12-16-2016

The Spike-and-Slab LASSO

Veronika Ročková

Edward I. George
University of Pennsylvania

Recommended Citation
Ročková, V., & George, E. I. (2016). The Spike-and-Slab LASSO. *Journal of the American Statistical Association, 113* (521), 431-444.
http://dx.doi.org/10.1080/01621459.2016.1260469

The Spike-and-Slab LASSO

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.

Keywords
high-dimensional regression, LASSO, penalized likelihood, posterior concentration, spike-and-slab, variable selection

Disciplines
Business | Business Analytics | Management Sciences and Quantitative Methods | Statistics and Probability

The Spike-and-Slab LASSO

Veronika Ročková and Edward I. George *

Submitted on 7^{th} August 2015

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 paper, 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 non-separable 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 Bayesian EM and 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 non-separable variants of the penalty, showing rate-optimality of the global mode as well as optimal posterior concentration when $p > n$. Thus, the modal estimates can be supplemented with meaningful uncertainty assessments.

Keywords: High-dimensional Regression; LASSO; Penalized Likelihood; Posterior Concentration; Spike-and-Slab; Variable Selection.

*Veronika Ročková is Postdoctoral Research Associate, Department of Statistics, University of Pennsylvania, Philadelphia, PA, 19104, vrockova@wharton.upenn.edu; Edward I. George is Professor of Statistics, Department of Statistics, University of Pennsylvania, Philadelphia, PA, 19104, edgeorge@wharton.upenn.edu.

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 with a known variance $\sigma^2 = 1$. 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 $1 \leq j \leq p$. 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

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

where $pen_\lambda(\boldsymbol{\beta})$ 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, i.e. $pen_\lambda(\boldsymbol{\beta}) = \sum_{j=1}^p \rho_\lambda(\beta_j)$. 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) (its closest concave[1] relative) uses $\rho_\lambda(\beta_j) = -\lambda|\beta_j|$. These two approaches stand at the two ends of a conceptual and a computational ideal for sparsity detection. Non-concave separable elaborations, intermediate between the two, have witnessed a surge of interest (e.g. MCP penalty of Zhang (2010), 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 $pen_\lambda(\boldsymbol{\beta}) = \log \pi(\boldsymbol{\beta} \mid \lambda)$. In particular, separable penalties stem from independent product priors.

[1]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.

Spike-and-slab approaches to Bayesian variable selection arise directly from probabilistic considerations. With a hierarchical prior over the parameter and model spaces, generic spike-and-slab priors are of the form

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

where $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_p)'$, $\gamma_i \in \{0, 1\}$, is an intermediate vector of binary variables, indexing the 2^p possible models. Here, $\psi_0(\beta)$ serves as a "spike distribution" for modeling irrelevant (zero) coefficients, and $\psi_1(\beta)$ serves as a "slab distribution" for modeling large effects. For the *Spike-and-Slab LASSO* we deploy the particular variant of (1.3) with $\psi_0(\beta) = \frac{\lambda_0}{2} \mathrm{e}^{-\lambda_0 |\beta|}$ and $\psi_1(\beta) = \frac{\lambda_1}{2} \mathrm{e}^{-\lambda_1 |\beta|}$. Proposed by Rockova (2015) in the context of sparse normal means, these two-point mixtures of Laplace distributions will be referred to as the Spike-and-Slab LASSO *(SSL)* priors. The scope of these priors is greatly enhanced by the flexibility of the model space prior $\pi(\boldsymbol{\gamma})$, which can be used to gear $\pi(\boldsymbol{\beta})$ towards preferred configurations $\boldsymbol{\gamma}$. For our development, we confine attention to exchangeable model space priors of the form

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

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

Conditionally on θ, the *SSL* prior (1.3) boils down to an independent product of mixtures

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

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 $\lambda_1 \to 0$), $\log \pi(\boldsymbol{\beta} \,|\, \theta)$ 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 form a non-concave continuum between these two ideals. Despite the wide adoption of spike-and-slab formulations for variable selection, their potential for modal estimation (1.2) through penalty creation has largely been overlooked (with the notable exceptions of George and Foster (1997) and Abramovich and Grinshtein (2010), who linked point-mass-Gaussian mixtures with ℓ_0 selection criteria, and Yuan and Lin (2006) who studied empirical Bayes calibration of point-mass-Laplace mixtures). Here, we unleash the potential of penalty functions arising from the *continuous SSL* priors in the context of high-dimensional regression, moving beyond the framework of independent product priors.

Whereas spike-and-slab priors with fixed θ are interesting constructs on their own, we shall be ultimately interested in the fully Bayes formulations, treating θ as unknown and random with $\theta \sim \pi(\theta)$. Such hierarchical mixture priors 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 et al., 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 *non-separable*. Here we explore the hidden potential of these new penalty constructions. These fully Bayes penalties are an essential building block of our approach.

The main thrust of this paper is to propose the *Spike-and-Slab LASSO* procedure for simultaneous variable selection and estimation in high-dimensional linear regression. Summarized in the points below, the paper makes contributions at three fronts: methodology, theory and implementation.

(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 non-separable fully Bayes penalties is introduced and developed, showing their potential for self-adaptivity and automatic hyper-parameter 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 the developments in non-concave optimization (Breheny and Huang, 2011; Mazumder et al., 2011). Drawing upon the similarities to the LASSO, we extend existing coordinate-wise optimization algorithms to the case of a non-separable *SSL* penalty. Such adaptations are feasible and natural due to the underlying coherent Bayesian formalism, which attributes a probabilistic meaning to adaptive selection thresholds. The *SSL* priors are also amenable to MCMC techniques, which can be made more efficient by leaving out the zero directions identified by the posterior modes.

(3) The *Spike-and-Slab LASSO* method for variable selection is introduced, entailing the deployment of a sequence of (non-separable) priors within a path-following scheme. Unlike the LASSO which 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 of Rockova and George (2014), who proposed an efficient EM algorithm for Bayesian model exploration with a Gaussian spike-and-slab mixture. Here, we revisit their approach for *SSL* priors as an alternative strategy to coordinate ascent. The EMVS implementation with a Laplace mixture has

many advantages: automatic variable selection through thresholding, diminished bias in estimation, and provably faster convergence.

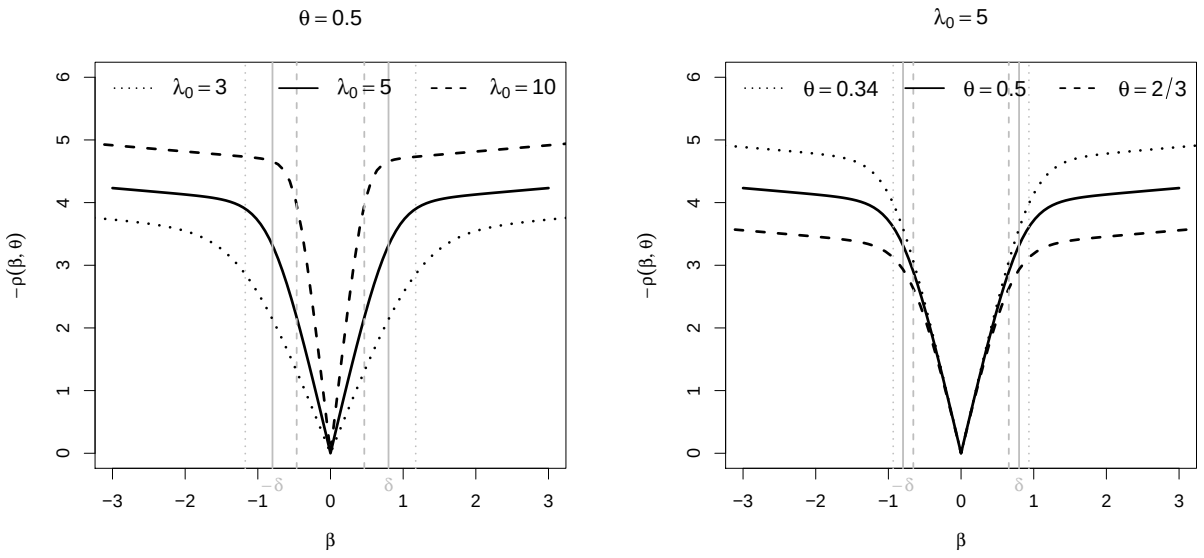
Path-following schemes are now routine for both concave/non-concave regularization. SCAD and MCP penalties have two hyper-parameters that require tuning, so that cross validation over a two-dimensional grid is often needed (Breheny and Huang, 2011; Mazumder et al., 2011). We also have two tuning parameters (λ_0, θ). However, by treating θ as random, the non-separable *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 of Rockova (2015), we show rate-optimality of the global mode under the separable penalty when $p > n$. This result is supplemented with an analogue involving the entire posterior measure. Building on the work of Castillo et al. (2015), we show that the *SSL* posterior keeps pace with the global mode by concentrating at the optimal rate when $p > n$. This result attributes meaning to the penalized likelihood surface, which can be used for meaningful uncertainty assessment, not only just as an objective function outputting a mode. Going further, we extend the analysis of the global mode to the case of non-separable *SSL* penalty functions, illuminating their potential for refining statistical rates.

The paper is structured as follows. Section 2 revisits the non-separable *SSL* penalty of Rockova (2015) in the context of high-dimensional linear regression. Section 3 introduces the framework of fully Bayes non-separable *SSL* penalties. Section 4 proposes a new coordinate ascent strategy for the non-separable *SSL* penalty and revisits the Bayesian EM strategy. Section 5 introduces the *Spike-and-Slab LASSO* approach and demonstrates its potential on a simulated example. 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 an independent product prior (1.5), assuming θ is fixed as if it were known. Paralleling the development of 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.



(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 $\rho(\beta\,|\,\theta)$ *with a minus sign* for different choices (λ_0, θ). The vertical lines correspond to the intersection point δ.

Definition 2.1. *Given $\theta \in (0,1)$, the separable Spike-and-Slab LASSO (SSL) penalty is defined as*

$$pen_S(\boldsymbol{\beta}\,|\,\theta) = \log\left[\frac{\pi(\boldsymbol{\beta}\,|\,\theta)}{\pi(\mathbf{0}_p\,|\,\theta)}\right] = \sum_{j=1}^{p} \log\left[\frac{\theta\,\psi_1(\beta_j) + (1-\theta)\psi_0(\beta_j)}{\theta\,\psi_1(0) + (1-\theta)\psi_0(0)}\right]. \tag{2.1}$$

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

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

which add up to yield

$$pen_S(\boldsymbol{\beta}\,|\,\theta) = \sum_{j=1}^{p} \rho(\beta_j\,|\,\theta) = -\lambda_1|\boldsymbol{\beta}| + \sum_{j=1}^{p} \log\left(\frac{p^\star_\theta(0)}{p^\star_\theta(\beta_j)}\right), \tag{2.3}$$

where

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

The alternative characterization (2.3) writes the separable *SSL* penalty as an adaptive sum of a LASSO penalty and a non-concave penalty, rendering it ultimately non-concave. The maximal non-concavity (Lv and Fan, 2009) 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 paper, we assume that λ_1 has been set to a small value (made precise by our theoretical study in Section 6) and thereby does not require tuning. 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 $\rho(\beta_j \,|\, \theta)$. As illustrated by Figure 1(a) (which plots $\rho(\beta_j \,|\, \theta)$ with a *minus* sign), the larger λ_0, the closer the approximation to ℓ_0. The plot also portrays $\rho(\beta_j \,|\, \theta)$ 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

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

The value δ represents a turning point, at which the slab has dominated the spike, and may be regarded as a threshold of practical significance (George and McCulloch, 1993). 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 θ. As seen from Figure 1(b), larger values θ represent a larger prior inclusion probability and thereby a smaller threshold. The particular choices $\theta = 2/3$ and $\theta = 0.34$ will be clarified in the next section, when linked to the non-separable *SSL* penalty. Figure 2 also shows that $\rho(\beta \,|\, \theta)$ shares many of the desirable properties required for separable regularizers (Zhang and Zhang, 2012): it is non-concave, non-increasing in $[0; \infty)$ and due to the convexity of $\log[p^\star_\theta(0)]/p^\star_\theta(\beta)]$ it is super-additive, i.e. $\rho(x + y \,|\, \theta) \geq \rho(x \,|\, \theta) + \rho(y \,|\, \theta)$ for all $x, y \geq 0$.

Before proceeding, it is worthwhile to examine more closely $p^\star_\theta(\beta_j)$ defined in (2.4), which is 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,

$$p^\star_\theta(\beta_j) = \mathsf{P}(\gamma_j = 1 \,|\, \beta_j, \theta) = \left[1 + \frac{\lambda_0}{\lambda_1} \frac{(1-\theta)}{\theta} \mathrm{e}^{-|\beta_j|(\lambda_0 - \lambda_1)}\right]^{-1}. \tag{2.6}$$

This quantity will keep reoccurring throughout the paper in many different contexts: implementation, statistical rates of the global mode, posterior concentration rates etc. We have already seen it in the definition of the intersection point (2.5). 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 2.1. *The derivative of the separable SSL penalty satisfies*

$$\frac{\partial pen_S(\boldsymbol{\beta} \,|\, \theta)}{\partial |\beta_j|} \equiv -\lambda^\star_\theta(\beta_j),$$

where

$$\lambda^\star_\theta(\beta_j) = \lambda_1 p^\star_\theta(\beta_j) + \lambda_0[1 - p^\star_\theta(\beta_j)]. \tag{2.7}$$

Proof. The result follows immediately from

$$\frac{\partial pen_S(\boldsymbol{\beta} \,|\, \theta)}{\partial |\beta_j|} = p^\star_\theta(\beta_j) \frac{\partial \log \psi_1(\beta_j)}{\partial |\beta_j|} + [1 - p^\star_\theta(\beta_j)] \frac{\partial \log \psi_0(\beta_j)}{\partial |\beta_j|}. \qquad \square$$

By exerting its influence through the Karush-Kuhn-Tucker (KKT) conditions (seen in (2.8) and (2.9) below), $\lambda^\star_\theta(\cdot)$ 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 $p^\star_\theta(\beta_j)$ (2.4), which gears $\lambda^\star_\theta(\beta_j)$ towards 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 *non-adaptive* sum of the ℓ_1 and a non-concave penalty (Liu and Wu, 2007; 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 $pen_S(\boldsymbol{\beta} \,|\, \theta)$. 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

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

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

which follow from the sub-differential calculus and Lemma 2.1. 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 et al., 2010) and thereby has instantaneous implications for the implementation (explored in Section 4). However, the LASSO penalty induces a constant shrinkage term λ, whereas the *SSL* penalty induces an adaptive term $\lambda^\star_\theta(\widehat{\beta}_j)$ that depends on the data through $\widehat{\beta}_j$ itself. As with the adaptive LASSO (Zou, 2006) and weighted ℓ_1 penalties (Candes et al., 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 ameliorates 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. 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 non-concavity

κ dominates the concavity of the log-likelihood when $(\lambda_0 - \lambda_1)^2 > 4\lambda_{min}$, where λ_{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 $pen_S(\boldsymbol{\beta} \,|\, \theta)$ 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 j^{th} direction, while keeping the other coordinates fixed at all but the j^{th} entry of $\widehat{\boldsymbol{\beta}}$. Thus, with z_j as before,

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

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

$$\Delta \equiv \inf_{t>0} \left[n\, t/2 - \rho(t \,|\, \theta)/t \right] \tag{2.12}$$

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

Theorem 2.1. *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 $pen_S(\boldsymbol{\beta} \,|\, \theta)$ satisfies*

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

where Δ is the selection threshold (2.12).

Theorem 2.1 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 $p_\theta^\star(0)$ will play an integral role in Δ.

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

$$g_\theta(x) = [\lambda_\theta^\star(x) - \lambda_1]^2 + 2\, n \log p_\theta^\star(x).$$

Denote by $c_+ = 0.5(1 + \sqrt{1 - 4n/(\lambda_0 - \lambda_1)^2}$ and

$$\delta_{c+} = \frac{1}{\lambda_0 - \lambda_1} \log \left[\frac{1-\theta}{\theta} \frac{\lambda_0}{\lambda_1} \frac{c_+}{1 - c_+} \right].$$

The value δ_{c+} is an inflection point of the univariate log-posterior in the j^{th} 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 $\lambda_\theta^\star(0) > \sqrt{2n \log[1/p_\theta^\star(0)]} + \lambda_1$, 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 2.2. *When $g_\theta(0) > 0$ and $(\lambda_0 - \lambda_1) > \sqrt{n}/2$, the threshold Δ in (2.12) is bounded by*

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

where

$$\Delta^L = \sqrt{2n \log[1/p_\theta^\star(0)] - d} + \lambda_1 \quad \textit{and} \quad \Delta^U = \sqrt{2n \log[1/p_\theta^\star(0)]} + \lambda_1. \tag{2.14}$$

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

As an aside of Theorem 2.2, we obtain that $\widehat{\boldsymbol{\beta}}$ has a zero gap, where the entries are either zero or above a certain threshold, i.e $|\widehat{\beta}_j| > \delta_{c+}$ when $\widehat{\beta}_j \neq 0$ (follows from Lemma 4.1 of Rockova (2015)).

Theorem 2.2 implies that for very non-concave 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 $\log[1/p_\theta^\star(0)]$, a fundamental quantity that affects statistical rates of the global mode (Section 6). Writing

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

we see that the parameters (λ_0, θ) have to work in concert to maintain the right balance. In order to achieve rate-minimaxity in sparse normal means under squared error loss, Rockova (2015) suggests setting $\lambda_0 \sim (1-\theta)/\theta \sim p/q$, when q is known. However, as will be seen in Section 5 we ultimately deploy *SSL* priors in a path following scheme, increasing λ_0, without knowledge of q. Assuming that somewhere along the solution path, λ_0 is actually approaching the right order p/q, we would like θ to adapt suitably. In the next section, we show that this can be achieved by treating θ as random.

3 The Wonder of a Non-separable *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 pre-specification 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

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

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

Recasting (3.2) as a penalty function, we obtain the following non-separable variant of the *SSL* penalty.

Definition 3.1. *The non-separable Spike-and-Slab LASSO (NSSL) penalty with $\theta \sim \pi(\theta)$ is defined as*

$$pen_{NS}(\boldsymbol{\beta}) = \log\left[\frac{\pi(\boldsymbol{\beta})}{\pi(\mathbf{0}_p)}\right] = -\lambda_1|\boldsymbol{\beta}| + \log\left[\frac{\int \frac{\theta^p}{\prod_{j=1}^{p} p_\theta^\star(\beta_j)} \mathrm{d}\,\pi(\theta)}{\int \frac{\theta^p}{\prod_{j=1}^{p} p_\theta^\star(0)} \mathrm{d}\,\pi(\theta)}\right]. \tag{3.3}$$

Again, we have centered the penalty so that $pen_{NS}(\mathbf{0}) = 0$. Contrasting (3.3) with (2.3), the *NSSL* penalty still writes as an additive composition of a (separable) LASSO part and a non-concave portion. But now, the non-concave part will be non-separable (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 non-separable analogue of Lemma 2.1. It will be convenient to let $\boldsymbol{\beta}_{\setminus j}$ denote the sub-vector of $\boldsymbol{\beta}$ containing all by the j^{th} entry.

Lemma 3.1. *The derivative of the non-separable Spike-and-Slab LASSO penalty* (3.3) *satisfies*

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

where

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

and

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

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

$$\begin{aligned}\frac{\partial \log \pi(\boldsymbol{\beta})}{\partial |\beta_j|} &= \frac{1}{\pi(\boldsymbol{\beta})} \int_0^1 \frac{\partial \pi(\boldsymbol{\beta} \,|\, \theta)}{\partial |\beta_j|} \pi(\theta) \mathrm{d}\, \theta = \int_0^1 \frac{\partial \log \pi(\boldsymbol{\beta} \,|\, \theta)}{\partial |\beta_j|} \pi(\theta \,|\, \boldsymbol{\beta}) \mathrm{d}\, \theta \\ &= -\lambda_1 \int_0^1 p_\theta^\star(\beta_j) \pi(\theta \,|\, \boldsymbol{\beta}) \mathrm{d}\, \theta - \lambda_0 \left[1 - \int_0^1 p_\theta^\star(\beta_j) \pi(\theta \,|\, \boldsymbol{\beta}) \mathrm{d}\, \theta \right]. \end{aligned}$$ □

We now pause a bit to appreciate the difference between (2.7) and (3.5). Instead of a "fixed-θ" mixing probability $p_\theta^\star(\beta_j)$, which appeared in the separable case, the non-separable penalty deploys an aggregated mixing probability $p^\star(\beta_j; \boldsymbol{\beta}_{\setminus j})$ obtained by averaging $p_\theta^\star(\cdot)$ over $\pi(\theta \,|\, \boldsymbol{\beta})$. 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 non-separable 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 $p_\theta^\star(\beta_j)$ is a non-linear 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 3.2. *Given $\boldsymbol{\beta} \in \mathbb{R}^p$ and prior $\pi(\theta)$ we can write*

$$p^\star(\beta_j; \boldsymbol{\beta}_{\setminus j}) = p_{\theta_j}^\star(\beta_j), \quad \textit{where} \quad \theta_j = \mathsf{E}\,[\theta \,|\, \boldsymbol{\beta}_{\setminus j}]. \tag{3.7}$$

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

$$\pi(\boldsymbol{\beta}) = \psi_1(\beta_j) \pi(\boldsymbol{\beta}_{\setminus j}) \int \frac{\theta}{p_\theta^\star(\beta_j)} \pi(\theta \,|\, \boldsymbol{\beta}_{\setminus j}) \mathrm{d}\, \theta. \tag{3.8}$$

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

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

Plugging in $p_\theta^\star(\beta_j)$ 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 non-separable case with the simple substitution $\theta = \mathsf{E}\,[\theta \,|\, \boldsymbol{\beta}_{\setminus j}]$. The identity (3.7) also illuminates how the fully Bayes formulation attributes a probabilistic meaning to the elements of the *NSSL* penalty.

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 3.2 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 $\rho(\beta_j \,|\, \theta) = -\lambda_1 |\beta_j| + \log[p_\theta^\star(0)/p_\theta^\star(\beta_j)]$ defined in (2.2). In a similar vein,

using (3.8) and (3.9), here we introduce conditional singletons in the j^{th} direction, while keeping $\boldsymbol{\beta}_{\setminus j}$ fixed:

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

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

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

where $\rho(\beta_j \,|\, \theta)$ is the singleton (2.2) of a separable penalty. In this way, the conditional singleton in the j^{th} 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 $\rho(\beta \,|\, \theta)$ 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 $\mathsf{E}\,\theta = 0.5$. The fixed choice θ leads to a singleton $\rho(\beta_1 \,|\, 0.5)$, which does not incorporate any information about β_2 (plotted in Figure 1(b) by a solid line). In contrast, assuming $\beta_2 = 0$, we obtain $\mathsf{E}\,[\theta \,|\, \beta_2 = 0] = 0.34$ which yields a conditional singleton $\rho(\beta_1 \,|\, 0.34)$ (plotted in Figure 1(b) with the dotted line). Relative to the fixed choice $\theta = 0.5$, $\mathsf{E}\,[\theta \,|\, \beta_2 = 0]$ 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 $\mathsf{E}\,[\theta \,|\, \beta_2 = 4] = 2/3$ (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. When more sparsity is detected in $\boldsymbol{\beta}_{\setminus j}$, the selection threshold for the j^{th} 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 non-separable case, we can readily obtain analogues 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 $pen_{NS}(\boldsymbol{\beta})$. A non-separable variant of the KKT necessary condition (2.10) writes as

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

where $z_j = \boldsymbol{X}_j'(\boldsymbol{Y} - \sum_{k \neq j} \boldsymbol{X}_k \widehat{\beta}_k)$ and $\theta_j = \mathsf{E}\,[\theta \,|\, \widehat{\boldsymbol{\beta}}_{\setminus j}]$. Contrasting (3.11) with (2.10), each coordinate now has a shrinkage term $\lambda^\star_{\theta_j}(\widehat{\beta}_j)$ which depends on all the coordinates, not just the j^{th}. This interconnection comes through θ_j.

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

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

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

$$\Delta_j \equiv \inf_{t>0} \left[n\, t/2 - \rho(t \,|\, \theta_j)/t \right]. \tag{3.13}$$

Combined with (3.11), this yields the following direct analogue of Theorem 2.1.

Theorem 3.1. *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 the non-separable penalty $pen_{NS}(\boldsymbol{\beta})$ satisfies*

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

where $\theta_j = \mathsf{E}\,[\theta \,|\, \widehat{\boldsymbol{\beta}}_{\setminus j}]$ 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 $\mathsf{E}\,[\theta \,|\, \widehat{\boldsymbol{\beta}}_{\setminus j}]$. This adaptation has an obvious empirical Bayes flavor. However, instead of estimating θ from the marginal likelihood (as in 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 3.2. *With $\theta_j = \mathsf{E}\,[\theta \,|\, \widehat{\boldsymbol{\beta}}_{\setminus j}]$ such that $g_{\theta_j}(0) > 0$ and with $(\lambda_0 - \lambda_1) > \sqrt{n}/2$, the adaptive threshold Δ_j defined in* (3.13) *satisfies*

$$\Delta_j^L < \Delta_j < \Delta_j^U,$$

where

$$\Delta_j^L = \sqrt{2n\, \log[1/p^{\star}_{\theta_j}(0)] - d_j} + \lambda_1 \quad \textit{and} \quad \Delta_j^U = \sqrt{2n\, \log[1/p^{\star}_{\theta_j}(0)]} + \lambda_1. \tag{3.14}$$

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

Proof. Follows from the proof of Lemma 4.1 of 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

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

which manifests the adaptability of the selection thresholds under the non-separable 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 non-entering the model $(1-\theta)/\theta$. In sharp contrast, (3.15) uses the "*posterior odds*" $[1 - \mathsf{E}\,(\theta \,|\, \widehat{\boldsymbol{\beta}}_{\setminus j})]/\mathsf{E}\,(\theta \,|\, \widehat{\boldsymbol{\beta}}_{\setminus j})$. 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 3.2, the global mode has a zero gap, where the nonzero estimates satisfy $|\widehat{\beta}_j| > \delta_j$ and δ_j is determined uniquely from $p^{\star}_{\theta_j}(\delta_j) = c_+$, where c_+ is defined in Section 2.

3.2 The Adaptive Weight

Because of the absolutely central role of $\mathsf{E}\,[\theta \,|\, \widehat{\boldsymbol{\beta}}_{\setminus j}]$ 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 obvious fact that the posterior expectations $\mathsf{E}\,[\theta \,|\, \widehat{\boldsymbol{\beta}}_{\setminus j}]$ will be very similar and close to $\mathsf{E}\,[\theta \,|\, \widehat{\boldsymbol{\beta}}]$, when p is sufficiently large. Thus, despite being coordinate-specific, Δ_j's may not be dramatically different. To continue, we examine the conditional distribution $\pi(\theta \,|\, \widehat{\boldsymbol{\beta}})$ 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

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

where $z = 1 - \frac{\lambda_1}{\lambda_0}$, $x_j = \left(1 - \frac{\lambda_1}{\lambda_0}\mathrm{e}^{|\widehat{\beta}_j|(\lambda_0 - \lambda_1)}\right)$. 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

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

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 $x \equiv \left(1 - \frac{\lambda_1}{\lambda_0}\mathrm{e}^{m(\lambda_0 - \lambda_1)}\right)$, where $m = \min\{\widehat{\beta}_j : \widehat{\beta}_j \neq 0\}$. In the

stylized scenario $x_j = x, 1 \leq j \leq p$, (3.18) is equal to a ratio of Appell F1 functions with shifted hyper-parameters, 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

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

is the Appell F1 function. Noting that the ratio (3.19) is monotone in x and z (Lemma 1 of Rockova and George (2015a)), suitable lower and upper bounds can be obtained for $\mathsf{E}[\theta \mid \widehat{\boldsymbol{\beta}}]$. 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 3.3. *Assume $\pi(\theta \mid \widehat{\boldsymbol{\beta}})$ is distributed according to* (3.17). *Let $\widehat{q} = ||\widehat{\boldsymbol{\beta}}||_0$. Then*

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

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 (2015a)

Lemma 3.3 has distinct implications in terms of the calibration 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 $\mathsf{E}[\theta \mid \widehat{\boldsymbol{\beta}}] \sim \widehat{q}/p$, 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.

Remark 3.1. *Lemma 3.3 provides a non-asymptotic 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 Aspects

A host of optimization algorithms have been proposed for non-concave *separable* penalties, including the local quadratic approximation LQA (Fan and Li, 2001), the local linear approximation LLA (Zou and Li, 2008; Candes et al., 2008), coordinate-wise optimization (Mazumder et al., 2011; Breheny and Huang, 2011), proximal gradient methods (Loh and Wainwright, 2014; Wang et al., 2014) or 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 (Fan et al., 2014; Wang et al., 2014). Here, we naturally extend two of these approaches, coordinate ascent and LLA, to the case of a non-separable *NSSL* penalty.

4.1 The Separable Case

We first revisit these two strategies in the context of the separable *SSL* penalty with fixed values of (λ_0, θ). By its striking resemblance to the LASSO regularization (made apparent by (2.10)), the *SSL* modal estimator naturally lends itself to coordinate-wise optimization. Such strategy makes use of the univariate soft-thresholding operator

$$S(z, \lambda) = \frac{1}{n}(|z| - \lambda)_+ \text{sign}(z).$$

Very much like for the LASSO (Friedman et al., 2010), stationary points satisfying (2.10) will be reached by cycling over one-site updates

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

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 j^{th} coordinate. Starting with an initial guess $\boldsymbol{\beta}^{(0)}$, parameters are cyclically updated according to (4.1) until convergence.

This computational strategy resembles the LLA algorithm (Zou and Li, 2008; Candes et al., 2008), which iterates over joint updates

$$\boldsymbol{\beta}^{(k+1)} = \arg\max_{\boldsymbol{\beta} \in \mathbb{R}^p} \left\{ -\frac{1}{2}||\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}||^2 - \sum_{j=1}^{p} \lambda_\theta^\star(\beta_j^{(k)})|\beta_j| \right\}. \tag{4.2}$$

From another point of view, (4.2) coincides with the M-step of a Bayesian EM algorithm for posterior mode detection under continuous spike-and-slab priors, which treats $\boldsymbol{\gamma}$ as missing data and keeps θ fixed. This connection is made apparent by the fact $\lambda_\theta^\star(\beta_j^{(k)}) = \lambda_1 p_\theta^\star(\beta_j^{(k)}) + \lambda_0[1 - p_\theta^\star(\beta_j^{(k)})]$, and by noting $p_\theta^\star(\beta_j) = \mathsf{E}\left(\gamma_j \mid \beta_j^{(k)}, \theta\right)$ (the E-step calculation). A similar strategy was implemented for a mixture of two Gaussian distributions in the EMVS procedure by Rockova and George (2014). Whereas their approach was based on iteratively solving adaptive ridge regressions, here it entails solving weighted LASSO regressions. The advantages of using EMVS with the LASSO updates are (a) automatic variable selection through thresholding, (b) faster speed of convergence (follows from considerations of Rockova and George (2015b)).

Both (4.1) and the EMVS implementation with (4.2) target all local maxima, including many peripheral modes. We can eliminate some of these suboptimal solutions with the aid of the refined characterization in Theorem 3.1. Following Mazumder et al. (2011), we define the generalized thresholding operator

$$\widetilde{S}(z, \lambda, \Delta) = \frac{1}{n}(|z| - \lambda)_+ \text{sign}(z)\mathbb{I}(|z| > \Delta). \tag{4.3}$$

With this operator, the refined coordinate-wise algorithm now cycles through

$$\beta_j^{(k+1)} = \widetilde{S}(z_j^{(k)}, \lambda_\theta^\star(\beta_j^{(k)}), \Delta), \tag{4.4}$$

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 $l \in \{1, \ldots, L\}$	
EMVS	NSCA
Set $k = 0$	Set $k = 0$
(a) Initialize: $\boldsymbol{\beta}_l^{(k)} = \boldsymbol{\beta}^\star$,$\theta^{(0)} = \theta^\star$	(a*) Initialize: $\boldsymbol{\beta}_l^{(k)} = \boldsymbol{\beta}^\star$,$\theta^{(0)} = \theta^\star$
(b) While $\mathsf{diff} > \varepsilon$	(b*) While $\mathsf{diff} > \varepsilon$
(i) Increment k	(i*) Increment k
(ii) Update $\boldsymbol{\beta}_l^{(k)}$ according to (4.2) with $\theta = \theta^{(k)}$	(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 (4.4) with $\theta = \theta^{(k)}$
	3. Update $\theta^{(k)} = \mathsf{E}\left[\theta \,|\, \boldsymbol{\beta}_l^{(k)}\right]$ using (3.19)
(iii) Update $\theta^{(k)}$ according to (4.5)	(iii*) $\mathsf{diff} = ||\boldsymbol{\beta}^{(k)} - \boldsymbol{\beta}^{(k-1)}||_2$
(iv) $\mathsf{diff} = ||\boldsymbol{\beta}^{(k)} - \boldsymbol{\beta}^{(k-1)}||_2$	(c*) Return $\boldsymbol{\beta}_l^{(k)}$
(c) Return $\boldsymbol{\beta}_l^{(k)}$	(d*) Assign $\boldsymbol{\beta}^\star = \boldsymbol{\beta}_l^{(k)}$
(d) Assign $\boldsymbol{\beta}^\star = \boldsymbol{\beta}_l^{(k)}$	

Table 1: Two variants of the Spike-and-Slab LASSO procedure

where $z_j^{(k)}$ is as in (4.1) and Δ is the selection threshold (2.12). The selection threshold Δ can be computed exactly using numerical optimization. This algorithm can be applied as a standalone or as a post-processing step after EMVS.

4.2 The Non-separable Case

We now turn to the optimization routines with λ_0 fixed and θ adaptive assuming $\pi(\theta) = \mathcal{B}(a, b)$. Extending LLA to the case of a non-separable penalty is achieved naturally within the Bayesian EM framework by treating θ as an additional model parameter. Instead of carrying forward the same fixed value, one now simply updates θ throughout the algorithm. The non-separable variant of the M-step thus uses $\theta = \theta^{(k)}$ to obtain $\boldsymbol{\beta}^{(k+1)}$ from (4.2). This step is followed by a new update $\theta^{(k+1)}$ according to

$$\theta^{(k+1)} = \frac{\sum_{j=1}^{p} p^\star_{\theta^{(k)}}(\beta_j^{(k+1)}) + a - 1}{a + b + p - 2}. \tag{4.5}$$

The calculation (4.5) follows directly from equation (3.12) of Rockova and George (2014). A variant of this strategy was implemented for sparse factor analysis by Rockova and George (2015b), where more details on this algorithm can be found.

Extending the coordinate ascent to the case of the non-separable *NSSL* penalty is made unapologetically simple by Lemma 3.2. Instead of using a fixed value θ, we can propagate it throughout the sweeps of coordinate ascent. Using Theorem 3.1, the k^{th} iteration of our proposed

Non-Separable Coordinate Ascent (NSCA) algorithm updates the j^{th} coordinate according to

$$\beta_j^{(k+1)} = \widetilde{S}\left(z_j^{(k)}, \lambda^{\star}_{\theta_j^{(k)}}(\beta_j^{(k)}), \Delta_j\right), \quad \text{where} \quad \theta_j^{(k)} = \mathsf{E}\left[\theta \,|\, \widetilde{\boldsymbol{\beta}}_{\setminus j}^{(k)}\right], \tag{4.6}$$

where Δ_j is the selection threshold (3.13) with $\theta = \theta_j^{(k)}$. 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 $\mathsf{E}\left[\theta \,|\, \widetilde{\boldsymbol{\beta}}_{\setminus j}\right]$ 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 3.3.

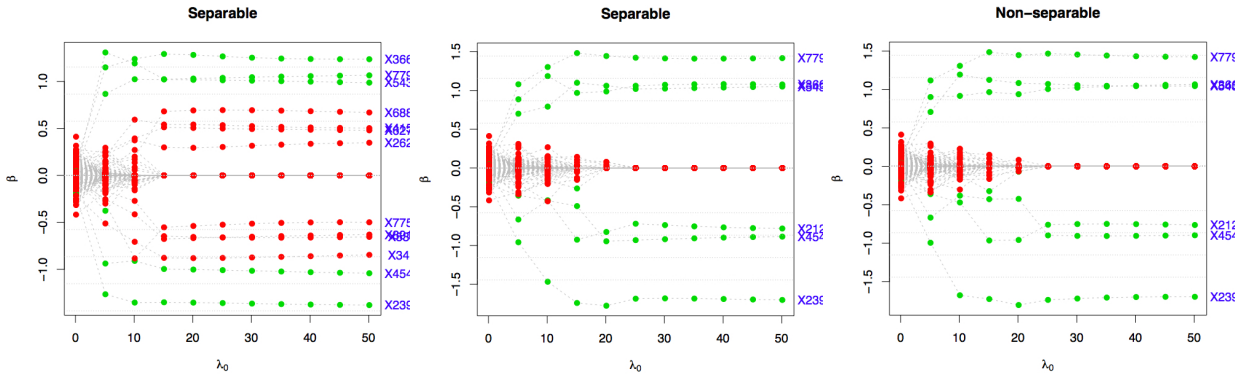
4.3 Bayesian Calculation

The *SSL* and *NSSL* priors are also amenable to posterior simulation. Direct Gibbs sampling is available through the exponential scale mixture representation of the Laplace distribution (Park and Casella, 2008), applying the SSVS strategy (George and McCulloch, 1993). Alternatively, one could deploy a variant of an orthant sampler developed for the Bayesian LASSO by Hans (2009). Whereas simulating from the full-dimensional posterior $\pi(\boldsymbol{\beta} \,|\, \boldsymbol{Y})$ will only be practical when p is not overwhelmingly big, initiating the sampler at a posterior mode can save burn-in time and provide a quick insight into uncertainty surrounding the mode. Alternatively, one could confine the simulation to a lower-dimensional subspace, sampling only from active coordinates identified by the mode hunting strategies. As will be shown in Section 6, the *SSL* posterior behaves optimally, providing an opportunity for meaningful uncertainty assessment.

5 The Spike-and-Slab LASSO

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

The path begins with λ_0^1 close to λ_1 so that the log-posterior is not too spiky, where a mode hunting algorithm finds the first solution $\widehat{\boldsymbol{\beta}}_1$. With $p < n$, choosing $\lambda_0^1 < 2\sqrt{\lambda_{min}} + \lambda_1$ ensures that $\widehat{\boldsymbol{\beta}}_1$ is the actual global mode. This output is then propagated with sequential reinitialization,



(a) $\theta = 0.5$ (b) $\theta = 8/1000$ (c) $\theta \sim \mathcal{B}(1, p)$

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

where $\widehat{\boldsymbol{\beta}}_{l-1}$ is used as a warm start for the next l^{th} calculation. At the end of the sequence, the method outputs the entire solution path $\{\widehat{\boldsymbol{\beta}}_1, \ldots, \widehat{\boldsymbol{\beta}}_L\}$ which inherently identifies a set of models through the inspection of the nonzero entries. Each individual $\widehat{\boldsymbol{\beta}}_l$ can be quickly obtained with either of the two non-concave optimization methods summarized in Table 1 and detailed in the previous section. Note that by performing model selection directly in the continuous $\boldsymbol{\beta}$ parameter space, the *Spike-and-Slab LASSO* is quite different from traditional spike-and-slab approaches which perform model selection on the basis of the discrete model space posterior $\pi(\boldsymbol{\gamma} \mid \boldsymbol{Y})$.

The sequential initialization is useful for the identification of a single good mode, which can be reported when further increases of λ_0 do not affect the solution. An example of such stabilization is seen in Figure 2 where, towards 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. For this purpose, it might be useful to rerun the *Spike-and-Slab LASSO* without sequential reinitialization (i.e. skipping steps (d) and (d*) in Table 1), in order to identify a more diverse set of models.

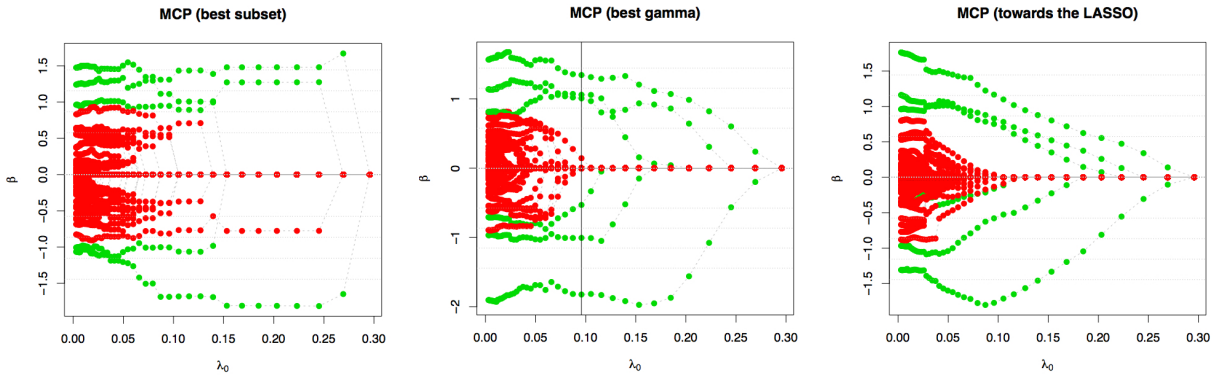
Compared with existing path-following methods with non-concave penalties (*SparseNet* of Mazumder et al. (2011), *ncvreg* of Breheny and Huang (2011)), the *Spike-and-Slab LASSO* permits the use of a self-adaptive *NSSL* penalty, avoiding the need for tuning its complexity parameter θ. We will illustrate this aspect in the next section. Non-adaptive variants of this strategy can be obtained by skipping the steps (ii) and (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 example. With $n = 100$ and $p = 1\,000$, we generate a data matrix $\boldsymbol{X}$ from a multivariate Gaussian distribution with mean $\mathbf{0}_p$ and $\Sigma = (\sigma_{ij})_{i,j=1}^{p}$, where $\sigma_{ij} = 0.6$ if $i \neq j$ and $\sigma_{ii} = 1$. The true vector $\boldsymbol{\beta}_0$ is constructed by assigning regression coefficients $\frac{1}{\sqrt{3}}\{-2.5, -2, -1.5, -1, 1, 1.5, 2, 2.5\}$ to $q = 8$ random directions and setting to zero all the remaining coefficients. 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 $\lambda_1 = 0.1$ 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, using the *NSCA* variant and starting at $\boldsymbol{\beta}^\star = \mathbf{0}_p$. We consider three settings: (a) a non-adaptive choice $\theta = 0.5$, clearly over-estimating the true nonzero fraction $8/1\,000$, (b) the non-adaptive oracle choice $\theta = 8/1\,000$ 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 green and false discoveries are depicted in red. The levels of true coefficients are marked by the horizontal dotted lines. Clearly, when θ is too large, there are many false positives (Figure 2(a)). When θ is set to the oracle choice $8/1000$, there are no false positives and only two false negatives. 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. This is also observed with the EMVS implementation, which outputs very similar solution paths. The purpose of this exercise has not been to compare the coordinate ascent with the EMVS (both of which are ultimately useful). The purpose has been to demonstrate that there are substantial gains when using the *NSSL* penalty. Fortunately, the practical implementation of the non-separable 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 et al., 2011). We applied the *SparseNet* algorithm of Mazumder et al. (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 solution surface are captured in Figure 3. The best subset regime (Figure 3(a)) outputs a solution path, whose middle part compares to *SSL* with $\theta = 0.5$ (Figure 2(a)). Again, we are finding quite a few false positives. As we increase γ, the solution begins to resemble a LASSO path, where all the coefficients are pulled towards zero with the same strength (Figure 3(c) with $\gamma = 8.562$). The best value of γ found by cross-validation was a compromise between the two (Figure 3(b) with $\gamma = 4.185$). The best solution (marked by a solid line) identifies correctly 6 coefficients (similarly as the oracle *SSL* solution), at the expense of one false positive and slightly



(a) $\gamma \approx 1$ (b) $\gamma = 4.185$ (c) $\gamma = 8.562$

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.

increased bias of the smaller nonzero coefficients. Overall, the performance of MCP after cross-validation is comparable to the performance of the adaptive *NSSL* penalty. It is 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 and requires cross-validation to identify the best-encountered solution.

6 Asymptotic Considerations

The purpose of this section is to provide affirmative statements about the suitability of the *SSL* and *NSSL* priors for sparse high-dimensional linear regression using 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 non-separable penalty can boost performance. In this section, we address all these points by studying statistical rates under squared error loss and assuming $||\boldsymbol{\beta}_0||_0 = q$.

Our analysis builds on existing theory developed for the LASSO (Bühlmann and van der Geer, 2011), non-concave separable regularizers (Zhang and Zhang, 2012) and posterior distributions under point-mass mixture priors (Castillo et al., 2015).

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

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

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 $||\boldsymbol{X}'\boldsymbol{\varepsilon}||_\infty < \eta\lambda$ (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 et al. (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 non-separable case (forthcoming Lemma 6.2). Moreover, Zhang and Zhang (2012) provide 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 non-separable *SSL* regularizers, $\boldsymbol{\Theta}$ lives inside a very specific set as follows from the following general lemma.

Lemma 6.1. *Assume that η-NC holds. Suppose $\widehat{\boldsymbol{\beta}} \in \mathbb{R}^p$ is the global mode* (1.2) *under a penalty $pen(\boldsymbol{\beta})$ and let $\boldsymbol{\Theta} = \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0$. Then $\boldsymbol{\Theta} \in C(\eta; \boldsymbol{\beta}_0)$, where*

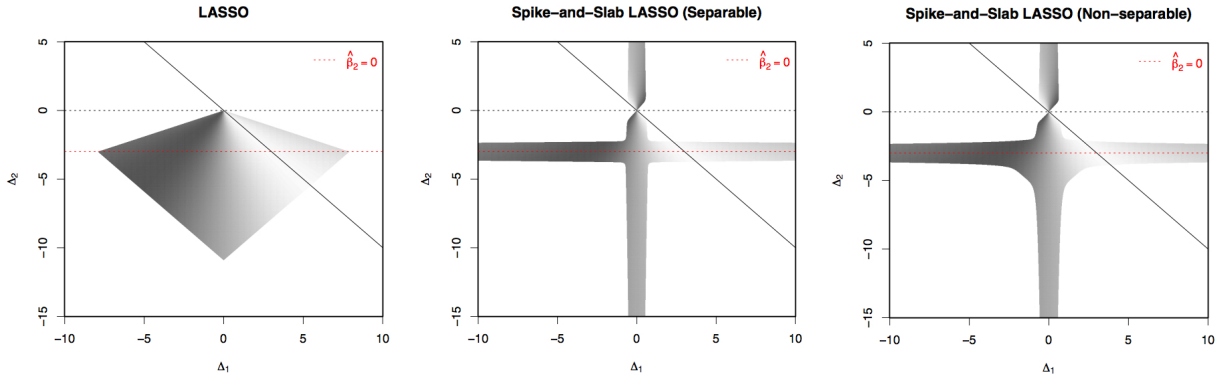
$$C(\eta; \boldsymbol{\beta}_0) = \left\{ \boldsymbol{\Theta} \in \mathbb{R}^p : pen(\boldsymbol{\Theta} + \boldsymbol{\beta}_0) \leq \frac{1}{\eta} \left[pen(\boldsymbol{\Theta} + \boldsymbol{\beta}_0) - pen(\boldsymbol{\Theta} - \boldsymbol{\beta}_0) \right] \right\}. \tag{6.1}$$

Proof. The resultfollows 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 6.2. *The minimal restricted eigenvalue is defined as*

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



(a) $\lambda_1 = \lambda_2$ (b) $\lambda_1 = 0.1, \lambda_0 = 10, \theta = 0.1$ (c) $\lambda_1 = 0.1, \lambda_0 = 10, a = 1, b = 9$

Figure 4: Plots of the feasible regions for $\boldsymbol{\Theta}$ under the LASSO penalty and the *SSL* penalty.

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 $C(\eta; \boldsymbol{\beta}_0)$. Figure 4 shows how this set depends on (λ_0, λ_1) and how it differs between the separable vs. non-separable *SSL* penalties. For the sake of illustration, we have assumed $\boldsymbol{\beta}_0 = (0, 3)'$ and $\eta = 0.45$. Under the LASSO penalty (where $\lambda_1 = \lambda_0$), the set (6.1) has a diamond shape (Figure 1(a)), embedded within a cone $\left\{\boldsymbol{\Theta} \in \mathbb{R}^p : |\boldsymbol{\Theta}_{S^c}| \leq \frac{1+\eta}{1-\eta} |\boldsymbol{\Theta}_S|\right\}$. On the other hand, for the other limiting case $\lambda_0 = \infty$ and $\lambda_1 = 0$, the *SSL* penalty corresponds to the ℓ_0 penalty. The set $C(\eta; \boldsymbol{\beta}_0)$ then consists of those values $\boldsymbol{\Theta} = \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0$ for which $||\widehat{\boldsymbol{\beta}}||_0 < 2$, as marked by the two solid lines corresponding to $\Theta_1 = 0$ and $\Theta_2 = -3$. The *SSL* penalty yields a compromise between these two extremes. With $1 < \lambda_0 < \infty$ and $0 < \lambda_1 < 1$, $C(\eta; \boldsymbol{\beta}_0)$ is a star-shaped wrap around the set $\{\boldsymbol{\Theta} : ||\widehat{\boldsymbol{\beta}}||_0 < 2\}$. With larger λ_1, the set begins to resemble a diamond. The non-separable penalty ties the coordinates together, making the set larger in the center.

Figures 4(a), 4(b) and 4(c) are actual heat-maps of the restricted eigenvalues $\frac{||\boldsymbol{X}\boldsymbol{\Theta}||}{||\boldsymbol{X}|| \, ||\boldsymbol{\Theta}||}$ inside $C(\eta; \boldsymbol{\beta}_0)$; the darker the shade of grey, the larger the value. Here, $\boldsymbol{X}$ contains $n = 100$ observations on 2 highly collinear variables (correlation $\rho = 0.96$). The diamond in Figure 4(a) is seen as a continuum of rays of equal eigenvalues, the minimum attained on the ray $\Theta_2 = -\Theta_1$ (marked by a solid line). This ray dissects all three sets in Figure 4. Under the ℓ_0 penalty, the intersection occurs at $\widehat{\boldsymbol{\beta}} = (3, 0)'$, the "opposite" of $\boldsymbol{\beta}_0 = (0, 3)$ where the correct variable was mistaken for its "knockoff". This unfavorable case is assigned a very small $c(\eta; \boldsymbol{\beta}_0)$ value, an indication that the true variable is not easily distinguishable. Interestingly, compared to the LASSO set, the *SSL* feasible regions indicate that at least one of the coordinates in $\widehat{\boldsymbol{\beta}}$ must be negligible, acknowledging the sparsity of the true $\boldsymbol{\beta}_0$. However, despite their different geometries, the minimal eigenvalue taken over these different sets is the same. Thus, regardless of the penalty, the same identifiability condition has to be imposed in this example.

To proceed with our analysis, we will need to borrow one more concept from the penalized likelihood literature, the compatibility.

Definition 6.3. *The compatibility number $\phi(C)$ of vectors in $C \subset \mathbb{R}^p$ is defined as*

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

For a nice description of this and related principles, we refer to Bühlmann and van der Geer (2011) and Castillo et al. (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 Appendix.

6.2 Asymptotic Properties

In this section, we build on the work of Rockova (2015), who showed that the global mode under the separable *SSL* penalty achieves rate-optimality when $\boldsymbol{X} = \mathrm{I}_n$ over sparse vectors under squared error loss. Going further, the entire posterior was shown to concentrate at the optimal rate in this setting. More precisely, Rockova (2015) shows that $\sup_{||\boldsymbol{\beta}_0||_0 \leq q} \mathsf{E}_{\boldsymbol{\beta}_0} ||\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0||^2 \sim q \log[1/p^\star_\theta(0)]$ and for some $M > 0$

$$\lim_{n\to\infty} \sup_{||\boldsymbol{\beta}_0||_0 \leq q} \mathsf{E}_{\boldsymbol{\beta}_0} \mathsf{P}(||\boldsymbol{\beta} - \boldsymbol{\beta}_0||^2 > M\, q \log[1/p^\star_\theta(0)] \,|\, \boldsymbol{y}, \theta) \to 0.$$

Both the risk and the posterior concentration rates are ultimately driven by the quantity $\log[1/p^\star_\theta(0)]$ and thereby can simultaneously achieve optimality with suitable hyper-parameter choices. In particular, with $C < \lambda_1 < \mathrm{e}^{-2}$ and with the oracle choice $\lambda_0 = (1-\theta)/\theta \sim n/q$ (assuming q is known) one achieves the minimax rate $q \log(n/q)$, whereas $\lambda_0 = (1-\theta)/\theta \sim n$ yields the near-minimax rate $q \log n$. Whereas the results in the orthogonal case do not imply analogues in the challenging high-dimensional regression case, in the forthcoming sections we show that this is indeed the case here.

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 6.1. *Let $\widehat{\boldsymbol{\beta}}$ be the Spike-and-Slab LASSO estimator under the separable penalty $pen_S(\boldsymbol{\beta}\,|\theta)$ and let $\widehat{q} = ||\widehat{\boldsymbol{\beta}}||_0$. Assume $(1-\theta)/\theta = p$, $\lambda_0 = p^d$, where $d \geq 2$, and $\sqrt{n}/p < \lambda_1 \leq \sqrt{2\,n \log p}$. Assume η-NC holds and let $c = c(\eta; \boldsymbol{\beta}_0)$ be the minimal restricted eigenvalue. Denote by $D = \left[\frac{\eta}{c} + \frac{\eta+1}{c\sqrt{d}}\right]^2$ and assume $D < 1 - \varepsilon$ for some $0 < \varepsilon < 1$. Then*

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

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

Proof. Appendix 8.1.1

Remark 6.1. *The smaller* $c(\eta; \boldsymbol{\beta}_0)$*, the harder it is to recover the true set* S_0*, which is manifested in Theorem 6.1 by a larger constant* K*.*

From existing theory about global posterior modes under separable non-concave regularizers (Theorem 1 by Zhang and Zhang (2012)), it turns out that the statistical rates (under η-NC) are ultimately guided by the selection threshold $\Delta \leq \min\{\lambda^\star_\theta(0), \sqrt{2n\log[1/p^\star_\theta(0)]}\}$. When λ_0 is not so large (i.e. $g_\theta(0) < 0$), Δ behaves similarly as $\lambda^\star(0)$, which in turn is very close to λ_0. In order to exert the influence of the spike-and-slab penalty, we need to increase λ_0 so that $g_\theta(0) > 0$. This condition is guaranteed when $\lambda_0 = p^d$ for some $d > 0$ and $(1-\theta)/\theta = p$, yielding $\Delta \sim \log[1/p^\star_\theta(0)] \sim \sqrt{2\,n\log p}$. This is the recognizable universal threshold (up to the a scaling factor[2] n), which produces the familiar near-minimax rates for the LASSO (van der Geer and Buhlmann). These considerations suggest that the global mode $\widehat{\boldsymbol{\beta}}$ will attain these rates with a careful tuning of Δ. We could apply Theorem 1 of Zhang and Zhang (2012) to obtain the statistical rates in terms of Δ. The following variant this theorem expresses the rates directly in terms of (q, n, p).

Theorem 6.2. *Let* $\widehat{\boldsymbol{\beta}}$ *be the Spike-and-Slab LASSO estimator under the separable penalty* $pen_S(\boldsymbol{\beta}\,|\theta)$*. Under the same conditions as in Theorem 6.1 we have*

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

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

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

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

Proof. Appendix 8.1.2

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. For this purpose, it is imperative that the entire posterior distribution concentrates at the right place and at the right rate (Castillo

[2]The scaling factor here emerges because we did not divide the likelihood portion of (1.2) by n.

and van der Vaart, 2012; Martin and Walker, 2014a). Adopting this perspective, we provide a comprehensive analysis of the entire posterior distribution, without restriction to data satisfying η-NC.

Our theoretical analysis follows closely Castillo et al. (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 and van der Vaart (2012); Castillo et al. (2015); Martin and Walker (2014a,b); van der Pas et al. (2014); Bhattacharya et al. (2015) and Rockova (2015).

To begin, we provide an analogue of Theorem 6.1 for the entire posterior measure. The *SSL* prior is inherently continuous, assigning zero mass to exactly sparse vectors. Following Rockova (2015) and Bhattacharya et al. (2015) we therefore use the following generalized notion of sparsity. Let δ be the intersection point between the spike and the slab densities in (2.5). 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, one can think of δ as a threshold of practical significance. With $\lambda_0 = (1-\theta)/\theta = p^d$, this threshold goes to zero rapidly, where $|\boldsymbol{\gamma}(\boldsymbol{\beta})|$ quickly approaches $||\boldsymbol{\beta}||_0$.

Throughout this section, we denote by $S_0 \subset \{1, \ldots, p\}$ the support $\boldsymbol{\beta}_0$, where $|S_0| = q$. As a natural continuation of Theorem 6.1, the following theorem shows that the expected posterior probability that the generalized dimensionality is a constant multiple larger than q, is asymptotically vanishing.

Theorem 6.3. *Assume $\lambda_0 = (1-\theta)/\theta = p^d$, where $d \geq 2$, and $\sqrt{n}/p < \lambda_1 \leq 4\sqrt{n \log p}$. Assume $p > n$ and $n, p \to \infty$. Then*

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

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

Proof. Appendix 8.2.1

The following theorem is a variant of Theorem 6.2, again involving the entire posterior not just its mode. Here, we obtain the same rates as in Theorem 6.2, with only slightly different multiplication constants (these are expressed in terms of modified compatibility numbers and restricted eigenvalues, defined in (8.5), (8.6) and (8.7) in the Appendix).

Theorem 6.4. *Assume* $\lambda_0 = (1-\theta)/\theta = p^d$, *where* $d \geq 2$, *and* $\sqrt{n}/p < \lambda_1 \leq 4\sqrt{n \log p}$. *Assume* $p > n$ *and* $n, p \to \infty$. *Then*

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

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

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

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

Proof. Appendix 8.2.2

To obtain a posterior mode and an entire posterior that converge at the same rate is not a property that is automatic. Indeed, Castillo et al. (2015) show that the the posterior under a single Laplace prior contracts at a far slower rate than its mode (the LASSO estimator). Thus, the Spike-and-Slab LASSO posterior is rate-optimal from both the penalized likelihood and full Bayes perspectives.

The Global Mode (Non-separable Case)

This section focuses on the non-separable *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. Intrigued by the possibility that such well-behaving posterior produces similarly rate-optimal modes, Rockova and George (2015a) studies the *NSSL* penalty in the context of normal means. Focusing on the high-dimensional regression, here we link existing results for separable regularizers to the *NSSL* penalty, soliciting evidence that hierarchical priors have the potential to refine statistical rates.

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 6.2. *Let* $\widehat{\boldsymbol{\beta}}$ *be the global mode under the NSSL penalty. Assume 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)

Proof. Denote by $\boldsymbol{e}_j$ the j^{th} canonical vector. The global optimality of $\widehat{\boldsymbol{\beta}}$ yields for any $t \in \mathbb{R}$

$$-||\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}||/2 + pen_{NS}(\widehat{\boldsymbol{\beta}}) \geq -||\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}} - t\boldsymbol{X}_j||/2 + pen_{NS}(\widehat{\boldsymbol{\beta}} + t\boldsymbol{e}_j),$$

Because $||\boldsymbol{X}_j||^2 = n$, this is equivalent to

$$t\boldsymbol{X}_j'(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}) \leq t^2 n/2 + \log\left[\frac{\pi(\widehat{\boldsymbol{\beta}})}{\pi(\widehat{\boldsymbol{\beta}} + t\boldsymbol{e}_j)}\right] < nt^2/2 - \widetilde{\rho}(t; \widehat{\boldsymbol{\beta}}_{\setminus j}),$$

where we used the definition of the conditional singleton (3.10) together with (3.8) and (3.9). The statement of the lemma follows from the definition of the selection threshold Δ_j. □

With this lemma, one immediately obtains Theorem 1 of 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 $(1-\theta)/\theta = p$, yielding

$$\Delta \sim \sqrt{2n \log\left(1 + \frac{\lambda_0}{\lambda_1} p\right)}.$$

For the non-separable case, applying Lemma (3.3) we immediately obtain

$$\bar{\Delta} \sim \sqrt{2n \log\left(1 + \frac{\lambda_0}{\lambda_1} \frac{p}{\widehat{q}}\right)}.$$

This comparison suggests that the non-separable case offers improvement. With $\lambda_0 = p^d$ and $\sqrt{n}/p < \lambda_1 < \sqrt{2n \log p}$, we obtain the same near-minimax rates (6.3), (6.4) and (6.5) also for the non-separable case.

The next natural step would be adapting also λ_0, either with $\pi(\theta)$ by linking λ_0 to θ, or with a separate prior distribution $\pi(\lambda_0)$. We anticipate that this type of strategy would ultimately provide a penalty, devoid of parameter tuning, which can attain shaper than the near-minimax statistical rate. Such remarkable property has been recently shown for the SLOPE penalty (Su and Candes, 2015) using different arguments.

7 Discussion

In this paper we have proposed a new class of self-adapting penalty functions arising from fully Bayes formulations. These *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. It is interesting to compare these two

shrinkage patterns with the adaptive LASSO (Zou, 2006), which penalizes the larger coefficients less to avoid bias, and the SLOPE estimator (Su and Candes, 2015) which penalizes the larger coefficients more to adjust for multiplicity. In contrast, the *NSSL* shrinkage meets both of these goals simultaneously and does so with self-adaptive data-driven penalization.

The non-separability of the *NSSL* penalty is a necessary consequence of its capacity to incorporate ensemble information. However, existing theory and implementations for separable regularizers are naturally extended. Its hierarchical form lends itself to fast implementation via EM and coordinate-wise algorithms within the path-following *Spike-and-Slab LASSO* strategy. As seen on a simulated example, the performance of the *NSSL* penalty mimics that of a oracle *SSL* penalty, providing a viable substitute for cross-validation. Asymptotic theory for separable and non-separable variants of the penalty establishes rate-optimality of the global mode as well as optimal posterior concentration.

Finally, 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).

8 Appendix

8.1 Proofs of Section 6.2.1

Throughout this section, we denote by $Q(\boldsymbol{\beta}) = -\frac{1}{2}||\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}||^2 + pen_S(\boldsymbol{\beta} \mid \theta)$ the log-posterior under the separable *SSL* penalty.

8.1.1 Proof of Theorem 6.1

Proof. Denote by $\boldsymbol{\Theta} = \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0$. Because $0 \geq Q(\boldsymbol{\beta}^0) - Q(\widehat{\boldsymbol{\beta}})$, we can write

$$0 \geq ||\boldsymbol{X}\boldsymbol{\Theta}||^2 - 2\boldsymbol{\varepsilon}'\boldsymbol{X}\boldsymbol{\Theta} + 2\log\left[\frac{\pi(\boldsymbol{\beta}_0 \mid \theta)}{\pi(\widehat{\boldsymbol{\beta}} \mid \theta)}\right]. \tag{8.1}$$

Using the fact $p^\star_\theta(\widehat{\beta}_j) > c_+$ when $\widehat{\beta}_j \neq 0$, we can write

$$\begin{aligned}\log\left[\frac{\pi(\boldsymbol{\beta}_0 \mid \theta)}{\pi(\widehat{\boldsymbol{\beta}} \mid \theta)}\right] &\geq -\lambda_1|\boldsymbol{\beta}_0 - \widehat{\boldsymbol{\beta}}_0| + \sum_{j=1}^{p}\log\left[\frac{p^\star_\theta(\widehat{\beta}_j)}{p^\star_\theta(0)}\right] + \sum_{j=1}^{p}\log\left[\frac{p^\star_\theta(0)}{p^\star_\theta(\beta_{0j})}\right] \\ &\geq -\lambda_1|\boldsymbol{\beta}_0 - \widehat{\boldsymbol{\beta}}_0| + \widehat{q}\,b + (\widehat{q} - q)\log[1/p^\star_\theta(0)],\end{aligned}$$

where $0 > b = \log c_+ > \log 0.5$ is a constant very close to 0. Because $||\boldsymbol{X}'\boldsymbol{\varepsilon}||_\infty \leq \eta\Delta$ under η-NC condition, we can use the Hölder inequality $|\boldsymbol{\alpha}'\boldsymbol{\beta}| \leq |\boldsymbol{\alpha}|_\infty|\boldsymbol{\beta}|$ to find that

$$0 \geq ||\boldsymbol{X}\boldsymbol{\Theta}||^2 - 2(\eta\Delta + \lambda_1)|\boldsymbol{\Theta}| + 2\,\widehat{q}\,b + 2(\widehat{q} - q)\log[1/p^\star_\theta(0)]. \tag{8.2}$$

From Lemma 1 we know that $\boldsymbol{\Theta}$ lives inside the cone $C(\eta;\boldsymbol{\beta}_0)$. Thus, we can use Definition 1 to find that $||\boldsymbol{X}\boldsymbol{\Theta}||^2 \geq c(\eta;\boldsymbol{\beta}_0)^2||\boldsymbol{\Theta}||^2\,||\boldsymbol{X}||^2$. Denote by $c = c(\eta;\boldsymbol{\beta}_0)$. Using the fact $|\boldsymbol{\Theta}| \leq ||\boldsymbol{\Theta}||\,||\boldsymbol{\Theta}||_0^{1/2}$, we have

$$0 \geq c^2||\boldsymbol{\Theta}||^2||\boldsymbol{X}||^2 - 2(\eta\Delta + \lambda_1)||\boldsymbol{\Theta}||\,||\boldsymbol{\Theta}||_0^{1/2} + 2\,\widehat{q}\,b + 2(\widehat{q} - q)\log[1/p^\star_\theta(0)],$$

which is equivalent to writing

$$\left[c||\boldsymbol{\Theta}||\,||\boldsymbol{X}|| - \frac{(\eta\Delta + \lambda_1)}{c||\boldsymbol{X}||}||\boldsymbol{\Theta}||_0^{1/2}\right]^2 - \frac{(\eta\Delta + \lambda_1)^2}{c^2||\boldsymbol{X}||^2}||\boldsymbol{\Theta}||_0 + 2\,\widehat{q} + 2(\widehat{q} - q)\log[1/p^\star_\theta(0)] \leq 0.$$

This yields

$$(\widehat{q} - q)\log[1/p^\star_\theta(0)] + \widehat{q}\,b \leq \frac{(\eta\Delta + \lambda_1)^2}{2c^2||\boldsymbol{X}||^2}||\boldsymbol{\Theta}||_0.$$

By noting $||\boldsymbol{\Theta}||_0 \leq \widehat{q_n} + q_n$ and $||\boldsymbol{X}||^2 = n$, we can write

$$\widehat{q} \leq q\left(1 + \frac{2A - b}{B + b - A}\right),$$

where $A = \frac{(\eta\Delta+\lambda_1)^2}{2c^2||\boldsymbol{X}||^2}$ and $B = \log[1/p^\star_\theta(0)]$. Using the fact $\lambda_1 < \sqrt{2\,n\log p} < \sqrt{2\,n/dB}$ we have $\frac{A}{B} < \left(\frac{\eta}{c} + \frac{\eta+1}{c\sqrt{d}}\right)^2 \equiv D$. We can then write $\widehat{q} \leq q\left(1 + M\frac{D}{1-D}\right)$. $\square$

8.1.2 Proof of Theorem 6.2

Proof. With $\boldsymbol{\Theta} = \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0$ and by noting $\log\left[\frac{\pi(\boldsymbol{\beta}_0\,|\,\theta)}{\pi(\boldsymbol{\beta}\,|\,\theta)}\right] > -\lambda_1|\boldsymbol{\Theta}| + q\log p^\star_\theta(0)$ we can write

$$0 \geq ||\boldsymbol{X}\boldsymbol{\Theta}||^2 - 2(\eta\Delta + \lambda_1)|\boldsymbol{\Theta}| + 2\,q\log[p^\star_\theta(0)], \tag{8.3}$$

where Δ is the selection threshold. From Lemma 6.1 we have $||\boldsymbol{\Theta}||_0 \leq (K+1)\,q$ under the η-NC condition. Denote by $\phi = \phi[C(\eta;\boldsymbol{\beta}_0)]$. From the definition of the compatibility number ϕ, and using $4uv \leq u^2 + 4v^2$, we find that

$$\begin{aligned} 2(\eta\Delta + \lambda_1)|\boldsymbol{\Theta}| &\leq 3(\eta\Delta + \lambda_1)\frac{||\boldsymbol{X}\boldsymbol{\Theta}||\,\sqrt{(K+1)\,q}}{||\boldsymbol{X}||\phi} - (\eta\Delta + \lambda_1)|\boldsymbol{\Theta}| \\ &\leq \frac{||\boldsymbol{X}\boldsymbol{\Theta}||^2}{2} + \frac{5(K+1)q(\eta\Delta + \lambda_1)^2}{||\boldsymbol{X}||^2\phi^2} - (\eta\Delta + \lambda_1)|\boldsymbol{\Theta}|. \end{aligned}$$

Thus it follows from (8.3) that

$$\frac{1}{2}||\boldsymbol{X}\boldsymbol{\Theta}||^2 + (\eta\Delta + \lambda_1)|\boldsymbol{\Theta}| \leq \frac{5(K+1)q(\eta\Delta + \lambda_1)^2}{||\boldsymbol{X}||^2\phi^2} + 2\,q\log[1/p^\star_\theta(0)]. \tag{8.4}$$

With $\lambda_0 = p^d$, $(1-\theta)/\theta = p$ and $\sqrt{n}/p < \lambda_1 < \sqrt{2\,n\log p}$ we have $(\eta\Delta + \lambda_1) < C_1\eta\sqrt{n\log p}$ and $\log[1/p^\star_\theta(0)] < C_2 \log p$. With $||\boldsymbol{X}||^2 = n$, the first two statements of the theorem follow directly from (8.4). Let $c = c(\eta, \boldsymbol{\beta}_0)$ be the minimal restricted eigenvalue. Then the last statement is obtained from $||\boldsymbol{X}\boldsymbol{\Theta}|| > c||\boldsymbol{X}||||\boldsymbol{\Theta}||$. □

8.2 Proofs of Section 6.2.2

The construction of the proof follows Castillo et al. (2015), where suitable modifications are required when using the notion of generalized dimensionality. Before proceeding, we need to introduce some more notation. Let

$$\Lambda_{n,\boldsymbol{\beta},\boldsymbol{\beta}_0} = \mathrm{e}^{-\frac{1}{2}||\boldsymbol{X}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)||^2 + (\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}_0)'\boldsymbol{X}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)}$$

and

$$\Pi(\boldsymbol{\beta}\,|\,\theta) = \prod_{i=1}^{p}[\theta\psi_1(\beta_i) + (1-\theta)\psi_0(\beta_i)].$$

Throughout this section we will denote by $\bar{\lambda} = 2\sqrt{n\log p}$ the universal threshold. The rates in this section will be expressed in terms of slightly different compatibility and minimal eigenvalue numbers. Following Castillo (2015), for $S \subset \{1,\ldots,p\}$, we define: the compatibility number $\widetilde{\phi}(S)$ of a model S by

$$\widetilde{\phi}(S) = \inf\left\{\frac{||\boldsymbol{X}\boldsymbol{\beta}||\,|S|^{1/2}}{||\boldsymbol{X}||\,|\boldsymbol{\beta}_S|} : |\boldsymbol{\beta}_{S^c}| \le 5|\boldsymbol{\beta}_S|, \boldsymbol{\beta}_S \neq 0\right\}, \tag{8.5}$$

the compatibility in vectors of generalized dimension s by

$$\bar{\phi}(s) = \inf\left\{\frac{||\boldsymbol{X}\boldsymbol{\beta}||\,s^{1/2}}{||\boldsymbol{X}||\,|\boldsymbol{\beta}|} : 0 < |\boldsymbol{\gamma}(\boldsymbol{\beta})| \le s\right\} \tag{8.6}$$

and the minimal eigenvalue restricted to vectors $\boldsymbol{\beta}$ of generalized dimensionality at most s by

$$\bar{c}(s) = \inf\left\{\frac{||\boldsymbol{X}\boldsymbol{\beta}||}{||\boldsymbol{X}||\,||\boldsymbol{\beta}||} : 0 < |\boldsymbol{\gamma}(\boldsymbol{\beta})| \le s\right\} \tag{8.7}$$

For $S \subset \{1,\ldots,p\}$, let $\boldsymbol{\beta}_S \in \mathbb{R}^p$ be a subset of $\boldsymbol{\beta}$ with coordinates in S. Denote by $\Pi_S(\boldsymbol{\beta}\,|\,\theta)$ the marginal prior confined to coordinates in S. Denote by δ the intersection point between *SSL* densities, by $\pi \equiv \mathsf{P}(|\beta_1| \le \delta)$ and by $\pi(s\,|\,\theta) = \binom{p}{s}\pi^s(1-\pi)^{p-s} = \mathsf{P}[|\boldsymbol{\gamma}(\boldsymbol{\beta})| = s\,|\,\theta]$ the prior distribution on the effective dimensionality. By assumption, we have $||\boldsymbol{X}||^2 = \max_{1\le i\le p}||\boldsymbol{X}_i||^2 = n$.

We will need the following analogue of Lemma 2 of Castillo et al. (2015) for the separable *SSL* prior.

Lemma 8.1. *Assume $\boldsymbol{\beta}_0 \in \mathbb{R}^p$ has a support $S_0 \subset \{1,\dots,p\}$, where $|S_0| = q$. Assume $\lambda_0 = (1-\theta)/\theta = Cp^a$, where $a \geq 2$ and $C > 0$, and $\sqrt{n}/p < \lambda_1 \leq 4\sqrt{n \log p}$. Assume $p > n$. Then*

$$\int \Lambda_{n,\boldsymbol{\beta},\boldsymbol{\beta}_0} \Pi(\boldsymbol{\beta} \,|\, \theta) \mathrm{d}\, \boldsymbol{\beta} \geq \frac{\pi(q \,|\, \theta)}{p^{2\,q}} \mathrm{e}^{-1-D-\lambda_1 |\boldsymbol{\beta}|_1},$$

where $D > 0$.

Proof. Denote by $g(\boldsymbol{\beta}) = \mathrm{e}^{-||\boldsymbol{X}\boldsymbol{\beta}||^2 + (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}_0)' \boldsymbol{X}\boldsymbol{\beta}}$. Using the fact $||\boldsymbol{X}\boldsymbol{\beta}||^2 \leq 2||\boldsymbol{X}\boldsymbol{\beta}_{S_0}||^2 + 2||\boldsymbol{X}\boldsymbol{\beta}_{S_0^c}||^2$, we can write

$$\Lambda_{n,\boldsymbol{\beta},\boldsymbol{\beta}_0} > g(\boldsymbol{\beta}_{S_0^c}) g(\boldsymbol{\beta}_{S_0} - \boldsymbol{\beta}_{0 S_0}).$$

By the Jensen's inequality we have

$$\int g(\boldsymbol{\beta}) \Pi(\boldsymbol{\beta} \,|\, \theta) \mathrm{d}\, \boldsymbol{\beta} \geq \int \mathrm{e}^{-||\boldsymbol{X}\boldsymbol{\beta}||^2} \Pi(\boldsymbol{\beta} \,|\, \theta) \mathrm{d}\, \boldsymbol{\beta}.$$

Conditionally on θ, the SSL prior is separable, implying $\Pi(\boldsymbol{\beta} \,|\, \theta) = \Pi_{S_0}(\boldsymbol{\beta} \,|\, \theta) \Pi_{S_0^c}(\boldsymbol{\beta} \,|\, \theta)$. Changing variables $\boldsymbol{b} \to (\boldsymbol{\beta} - \boldsymbol{\beta}_0)$ and noting $\Pi_{S_0}(\boldsymbol{\beta} \,|\, \theta) > \theta^q \left(\frac{\lambda_1}{2}\right)^q \mathrm{e}^{-\lambda_1 |\boldsymbol{\beta}_{S_0}|}$, we can write

$$\int \Lambda_{n,\boldsymbol{\beta},\boldsymbol{\beta}_0} \Pi(\boldsymbol{\beta} \,|\, \theta) \mathrm{d}\, \boldsymbol{\beta} > \int \mathrm{e}^{-||\boldsymbol{X}\boldsymbol{\beta}_{S_0^c}||^2} \Pi_{S_0^c}(\boldsymbol{\beta} \,|\, \theta) \mathrm{d}\, \boldsymbol{\beta}_{S_0^c} \tag{8.8}$$

$$\times\, \theta^q \mathrm{e}^{-\lambda_1 |\boldsymbol{\beta}_0|} \int \mathrm{e}^{-||\boldsymbol{X}\boldsymbol{b}_{S_0}||^2} \left(\frac{\lambda_1}{2}\right)^q \mathrm{e}^{-\lambda_1 |\boldsymbol{b}_{S_0}|} \mathrm{d}\, \boldsymbol{b}_{S_0}. \tag{8.9}$$

To simplify the integral in (8.9) we use arguments of Castillo (2015) in the proof of Lemma 2. Under the assumption $||\boldsymbol{X}||/p < \lambda_1 < 4||\boldsymbol{X}||\sqrt{\log p}$, we obtain

$$\int \mathrm{e}^{-||\boldsymbol{X}\boldsymbol{b}_{S_0}||^2} \left(\frac{\lambda_1}{2}\right)^q \mathrm{e}^{-\lambda_1 |\boldsymbol{b}_{S_0}|} \mathrm{d}\, \boldsymbol{b}_{S_0} > \mathrm{e}^{-1} \left(\frac{\lambda_1}{||\boldsymbol{X}||}\right)^q \frac{\mathrm{e}^{-\lambda_1/||\boldsymbol{X}||}}{q!} > \frac{\mathrm{e}^{-1}}{p^q q!} \tag{8.10}$$

To simplify the integral in (8.8), we use $||\boldsymbol{X}\boldsymbol{\beta}|| \leq ||\boldsymbol{X}||\,|\boldsymbol{\beta}|$ to find that

$$\int \mathrm{e}^{-||\boldsymbol{X}\boldsymbol{\beta}_{S_0^c}||^2} \Pi_{S_0^c}(\boldsymbol{\beta} \,|\, \theta) \mathrm{d}\, \boldsymbol{\beta}_{S_0^c} > \int_{|\beta_i| \leq \delta; i \notin S_0} \mathrm{e}^{-||\boldsymbol{X}||^2 |\boldsymbol{\beta}_{S_0^c}|^2} \Pi_{S_0^c}(\boldsymbol{\beta} \,|\, \theta) \mathrm{d}\, \boldsymbol{\beta}_{S_0^c} \tag{8.11}$$

$$\geq \mathrm{e}^{-(p-q)^2 ||\boldsymbol{X}||^2 \delta^2} \mathsf{P}(|\beta_1| \leq \delta)^{p-q}. \tag{8.12}$$

Combining (8.10) and (8.12) and noting $\theta > \pi \equiv \mathsf{P}(|\beta_1| > \delta \,|\, \theta)$, we obtain

$$\int \Lambda_{n,\boldsymbol{\beta},\boldsymbol{\beta}_0} \Pi(\boldsymbol{\beta} \,|\, \theta) \mathrm{d}\, \boldsymbol{\beta} > \mathrm{e}^{-(p-q)^2 ||\boldsymbol{X}||^2 \delta^2} \mathrm{e}^{-1-\lambda_1 |\boldsymbol{\beta}_0|} \pi^q (1-\pi)^{p-q} \frac{1}{p^q q!}.$$

Recall that $\pi(q \,|\, \theta) = \binom{p}{q} \pi^q (1-\pi)^{p-q}$ is the prior probability of the effective dimensionality q. Since $\binom{p}{q} q! \leq p^q$, we can write

$$\int \Lambda_{n,\boldsymbol{\beta},\boldsymbol{\beta}_0} \Pi(\boldsymbol{\beta} \,|\, \theta) \mathrm{d}\, \boldsymbol{\beta} > \mathrm{e}^{-(p-q)^2 ||\boldsymbol{X}||^2 \delta^2} \mathrm{e}^{-1-\lambda_1 |\boldsymbol{\beta}_0|} \frac{\pi(q \,|\, \theta)}{p^{2\,q}}.$$

Using the fact $\delta = \frac{1}{\lambda_0 - \lambda_1}\log[1/p^\star(0) - 1]$, we obtain

$$(p-q)^2||\boldsymbol{X}||^2\delta^2 = \frac{(p-q)^2||\boldsymbol{X}||^2}{(\lambda_0-\lambda_1)^2}\log^2[1/p^\star(0)-1] \tag{8.13}$$

Since $||\boldsymbol{X}|| = \sqrt{n} < \sqrt{p}$ we have $\lambda_1 \leq 4||\boldsymbol{X}||\sqrt{\log p} < 4p$. Because $\lambda_0 = p^d$ with $d \geq 2$, we have $\frac{p-q}{\lambda_0-\lambda_1} < \frac{1}{p^{d-1}-4}$. Because $(1-\theta)/\theta = p^d$, we have $\log[1/p^\star(0)-1] = 2\,d\log p$. Therefore, with $p > n$ and $d \geq 2$ we obtain

$$(p-q)^2||\boldsymbol{X}||^2\delta^2 < \frac{n\,4d^2\log^2 p}{(p^{d-1}-4)^2} < D. \quad \square$$

8.2.1 Dimensionality Result: Proof of Theorem 6.3

Proof. Denote by $\mathcal{B} = \{\boldsymbol{\beta} : |\gamma(\boldsymbol{\beta})| > R\}$. Then $\mathsf{E}_{\boldsymbol{\beta}_0}\mathsf{P}(\mathcal{B}\,|\,\boldsymbol{Y},\theta) \leq \mathsf{E}_{\boldsymbol{\beta}_0}\mathsf{P}(\mathcal{B}\,|\,\boldsymbol{Y},\theta)\mathbb{I}_{\tau_0} + \frac{2}{p}$, where $\tau_0 = \{||\boldsymbol{X}'(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}_0)||_\infty \leq \bar{\lambda}\}$ and $\bar{\lambda} = 2\sqrt{n\log p}$. Then

$$\mathsf{P}(\mathcal{B}\,|\,\boldsymbol{Y},\theta) = \frac{\int_{\mathcal{B}}\Lambda_{n,\boldsymbol{\beta},\boldsymbol{\beta}_0}\Pi(\boldsymbol{\beta}\,|\,\theta)\mathrm{d}\,\boldsymbol{\beta}}{\int\Lambda_{n,\boldsymbol{\beta},\boldsymbol{\beta}_0}\Pi(\boldsymbol{\beta}\,|\,\theta)\mathrm{d}\,\boldsymbol{\beta}} \leq A\mathrm{e}^{\lambda_1|\boldsymbol{\beta}_0|}\int_{\mathcal{B}}\mathrm{e}^{-\frac{1}{2}||\boldsymbol{X}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)||^2+(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}_0)'\boldsymbol{X}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)}\Pi(\boldsymbol{\beta}\,|\,\theta)\mathrm{d}\,\boldsymbol{\beta}, \tag{8.14}$$

where $A = \frac{p^{2\,q}}{\pi(q\,|\,\theta)}\mathrm{e}^{1+D}$. Similarly as in the proof of Theorem 12 of Castillo et al. (2015), we use Hölder's inequality to obtain on τ_0

$$(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}_0)'\boldsymbol{X}(\boldsymbol{\beta}-\boldsymbol{\beta}_0) \leq \bar{\lambda}|\boldsymbol{\beta}-\boldsymbol{\beta}_0|. \tag{8.15}$$

Therefore, the expectation under $\boldsymbol{\beta}_0$ of the integrand satisfies

$$\mathrm{e}^{-\frac{1}{2}||\boldsymbol{X}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)||^2}\mathsf{E}_{\boldsymbol{\beta}_0}\left[\mathrm{e}^{\left(1-\frac{\lambda_1}{2\bar{\lambda}}\right)(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta})'\boldsymbol{X}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)}\mathbb{I}_{\tau_0}\right]\mathrm{e}^{\frac{\lambda_1}{2}|\boldsymbol{\beta}-\boldsymbol{\beta}_0|} \tag{8.16}$$

$$\leq \mathrm{e}^{-\frac{1}{2}\left[1-\left(1-\frac{\lambda_1}{2\bar{\lambda}}\right)^2\right]||\boldsymbol{X}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)||^2}\mathrm{e}^{\frac{\lambda_1}{2}|\boldsymbol{\beta}-\boldsymbol{\beta}_0|} \tag{8.17}$$

$$\leq \mathrm{e}^{-\frac{\lambda_1}{4\bar{\lambda}}||\boldsymbol{X}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)||^2}\mathrm{e}^{\frac{\lambda_1}{2}|\boldsymbol{\beta}-\boldsymbol{\beta}_0|}, \tag{8.18}$$

where we used $\lambda_1 \leq 2\bar{\lambda}$ and invoked the expectation of a log-normally distributed r.v. Thus,

$$\mathsf{E}_{\boldsymbol{\beta}_0}\mathsf{P}(\mathcal{B}\,|\,\boldsymbol{Y},\theta)\,\mathbb{I}_{\tau_0} \leq A\,\mathrm{e}^{\lambda_1|\boldsymbol{\beta}_0|}\int_B\mathrm{e}^{-\frac{\lambda_1}{4\bar{\lambda}}||\boldsymbol{X}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)||^2}\mathrm{e}^{\frac{\lambda_1}{2}|\boldsymbol{\beta}-\boldsymbol{\beta}_0|}\mathrm{d}\,\Pi(\boldsymbol{\beta}\,|\,\theta). \tag{8.19}$$

Now, when $5|\boldsymbol{\beta}_{S_0}-\boldsymbol{\beta}_0| \leq |\boldsymbol{\beta}_{S_0^c}|$, then

$$|\boldsymbol{\beta}_0| + \frac{1}{2}|\boldsymbol{\beta}-\boldsymbol{\beta}_0| \leq |\boldsymbol{\beta}_{S_0}| + \frac{5}{4}|\boldsymbol{\beta}_S-\boldsymbol{\beta}_0| + \frac{3}{4}|\boldsymbol{\beta}_{S_0^c}| - \frac{1}{4}|\boldsymbol{\beta}-\boldsymbol{\beta}_0| < -\frac{1}{4}|\boldsymbol{\beta}-\boldsymbol{\beta}_0| + |\boldsymbol{\beta}| \tag{8.20}$$

$$< -\frac{1}{4}|\boldsymbol{\beta}-\boldsymbol{\beta}_0| + |\boldsymbol{\beta}| + \frac{1}{4\bar{\lambda}}||\boldsymbol{X}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)||^2 + 2\frac{\bar{\lambda}|S_0|}{||\boldsymbol{X}||^2\widetilde{\phi}(S_0)^2}. \tag{8.21}$$

When $5|\boldsymbol{\beta}_{S_0} - \boldsymbol{\beta}_0| > |\boldsymbol{\beta}_{S_0^c}|$, we use the definition of the compatibility number to find that

$$|\boldsymbol{\beta}_{S_0}| + \frac{5}{4}|\boldsymbol{\beta}_{S_0} - \boldsymbol{\beta}_0| + \frac{3}{4}|\boldsymbol{\beta}_{S_0^c}| - \frac{1}{4}|\boldsymbol{\beta} - \boldsymbol{\beta}_0| < -\frac{1}{4}|\boldsymbol{\beta} - \boldsymbol{\beta}_0| + |\boldsymbol{\beta}| + \frac{5}{4}\frac{||\boldsymbol{X}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)||\,|S_0|^{1/2}}{||\boldsymbol{X}||\widetilde{\phi}(S_0)}.$$

Invoking the inequality $4uv \le u^2 + 4v^2$, we can bound the last display from above by

$$-\frac{1}{4}|\boldsymbol{\beta} - \boldsymbol{\beta}_0| + |\boldsymbol{\beta}| + \frac{1}{4\bar{\lambda}}||\boldsymbol{X}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)||^2 + 2\frac{\bar{\lambda}\, q}{||\boldsymbol{X}||^2\widetilde{\phi}(S_0)^2}.$$

Thus (8.19) can be bounded by

$$A\,\mathrm{e}^{\frac{2\lambda_1\bar{\lambda}\,q}{||\boldsymbol{X}||^2\widetilde{\phi}(S_0)^2}} \int_{\mathcal{B}} \mathrm{e}^{\lambda_1|\boldsymbol{\beta}| - \frac{\lambda_1}{4}|\boldsymbol{\beta} - \boldsymbol{\beta}_0|}\Pi(\boldsymbol{\beta}\,|\,\theta)\mathrm{d}\,\boldsymbol{\beta}.$$

Note that $\pi(\beta\,|\,\theta) < \theta\lambda_1\mathrm{e}^{-\lambda_1|\beta|}$ when $|\beta| > \delta$. For $\mathcal{B} = \{\boldsymbol{\beta} : |\boldsymbol{\gamma}(\boldsymbol{\beta})| > R\}$, we can write

$$\begin{aligned}
\int_{\mathcal{B}} \mathrm{e}^{\lambda_1|\boldsymbol{\beta}| - \frac{\lambda_1}{4}|\boldsymbol{\beta} - \boldsymbol{\beta}_0|}\Pi(\boldsymbol{\beta}\,|\,\theta)\mathrm{d}\,\boldsymbol{\beta} &\le \sum_{S:|S|>R} \theta^{|S|}\lambda_1^{|S|} \int_{|\beta_i|>\delta; i\in S} \mathrm{e}^{-\frac{\lambda_1}{4}|\boldsymbol{\beta}_S - \boldsymbol{\beta}_{0S}|}\mathrm{d}\,\boldsymbol{\beta}_S && (8.22)\\
&\quad \times \int_{|\beta_i|\le\delta; i\in S^c} \mathrm{e}^{\lambda_1|\boldsymbol{\beta}_{S^c}| - \frac{\lambda_1}{4}|\boldsymbol{\beta}_{S^c} - \boldsymbol{\beta}_{0S^c}|}\Pi_{S^c}(\boldsymbol{\beta}\,|\,\theta)\mathrm{d}\,\boldsymbol{\beta}_{S^c} && (8.23)\\
&< \sum_{k=R+1}^{p} \binom{p}{k}(8\,\theta)^k \mathrm{e}^{\lambda_1\delta(p-k)}(1-\pi)^{p-k}. && (8.24)
\end{aligned}$$

Recall that $\pi \equiv \mathsf{P}(|\beta_1| > \delta\,|\,\theta) = \theta\,\mathrm{e}^{-\delta\lambda_1}\left(1 + \frac{\lambda_1}{\lambda_0}\right) < \theta$. Because $\theta < \pi\mathrm{e}^{\delta\lambda_1}$, we can bound the last display by

$$\mathrm{e}^{\lambda_1 p\,\delta} \sum_{k=R+1}^{p} 8^k \binom{p}{k}\pi^k(1-\pi)^{p-k} = \mathrm{e}^{\lambda_1 p\,\delta} \sum_{k=R+1}^{p} 8^k\pi(k\,|\,\theta).$$

Because $\pi/(1-\pi) < \theta/(1-\theta) \le 1/p^d$ for $d \ge 2$, we have

$$\pi(k\,|\,\theta) \le \left(\frac{1}{p}\right)^{d-1}\pi(k-1\,|\,\theta) \quad \text{for} \quad k \ge 1.$$

Thereby, we can write for $R > q$

$$\mathrm{e}^{\lambda_1 p\,\delta} \sum_{k=R+1}^{p} 8^k\pi(k\,|\,\theta) < \mathrm{e}^{\lambda_1 p\,\delta}\,8^q\,\pi(q\,|\,\theta)\left(\frac{8}{p^{d-1}}\right)^{R+1-q} \sum_{k=0}^{\infty}\left(\frac{8}{p^{d-1}}\right)^k.$$

With $(1-\theta)/\theta = \lambda_0 = p^d$ and $||\boldsymbol{X}||^2 = n$, we have $\lambda_1\delta p \le C_1$. Altogether

$$\begin{aligned}
\mathsf{P}(\mathcal{B}\,|\,\boldsymbol{Y}, \theta) &\preceq \mathrm{e}^{2q\log p + \lambda_1 p\,\delta + \frac{2\lambda_1\bar{\lambda}\,q}{n\widetilde{\phi}(S_0)^2}}\left(\frac{8}{p^{d-1}}\right)^{R+1-q} + \frac{2}{p}\\
&\preceq \mathrm{e}^{(R+1-q)\log 8 + 2q\log p[1 + 4\lambda_1/(\bar{\lambda}\widetilde{\phi}(S_0)^2)] - (R+1-q)(d-1)\log p} + \frac{2}{p}.
\end{aligned}$$

The right side of the above display goes to zero when $R > q\left[1 + \frac{M}{d-1}\left(1 + \frac{4\lambda_1}{\bar{\lambda}\widetilde{\phi}(S_0)^2}\right)\right]$ for some $M > 2$. □

8.2.2 Posterior Concentration Rate: Proof of Theorem 6.4

Proof. By Theorem 6.3, the posterior distribution is asymptotically supported on the event $E = \{\boldsymbol{\beta} : |\boldsymbol{\gamma}(\boldsymbol{\beta})| \leq q(1+K)|\}$, where $K = \frac{M}{d-1}\left(1 + \frac{4\lambda_1}{\bar{\lambda}\tilde{\phi}(S_0)^2}\right)$. Thus, we confine attention to $E^\star = E \cap \tau_0$, where τ_0 was defined in the proof of Theorem 6.3. From (8.14) and (8.15), we can see that

$$\Pi(\mathcal{B} \,|\, \boldsymbol{Y}, \theta)\mathbb{I}_{\tau_0} \leq \frac{p^{2q}\mathrm{e}^{1+D}}{\pi(q \,|\, \theta)} \int_{\mathcal{B}} \mathrm{e}^{-\frac{1}{2}||\boldsymbol{X}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)||^2 + 3\bar{\lambda}|\boldsymbol{\beta}-\boldsymbol{\beta}_0| + \lambda_1|\boldsymbol{\beta}|} \Pi(\boldsymbol{\beta} \,|\, \theta)\mathrm{d}\,\boldsymbol{\beta}. \tag{8.25}$$

We now use the definition of the compatibility number in vectors of generalized dimensionality (8.6). With the inequality $4uv \leq u^2 + 4v^2$, we can then write

$$\begin{aligned}(4-1)\bar{\lambda}|\boldsymbol{\beta}-\boldsymbol{\beta}_0| &\leq \frac{4\bar{\lambda}||\boldsymbol{X}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)||\,|\boldsymbol{\gamma}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)|^{1/2}}{\sqrt{n}\,\bar{\phi}(|\boldsymbol{\gamma}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)|)} - \bar{\lambda}|\boldsymbol{\beta}-\boldsymbol{\beta}_0| \\ &\leq \frac{1}{4}||\boldsymbol{X}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)||^2 + \frac{16\bar{\lambda}^2|\boldsymbol{\gamma}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)|}{n\,\bar{\phi}(|\boldsymbol{\gamma}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)|)^2} - \bar{\lambda}|\boldsymbol{\beta}-\boldsymbol{\beta}_0|\end{aligned}$$

Thus,

$$\Pi(\mathcal{B} \,|\, \boldsymbol{Y}, \theta)\mathbb{I}_{\tau_0} \leq \frac{p^{2q}\mathrm{e}^{1+D}}{\pi(q \,|\, \theta)}\mathrm{e}^{\frac{16\bar{\lambda}q(2+K)}{n[\bar{\phi}(2q+Kq)]^2}} \int_{\mathcal{B}} \mathrm{e}^{-\frac{1}{4}||\boldsymbol{X}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)||^2 - \bar{\lambda}|\boldsymbol{\beta}-\boldsymbol{\beta}_0| + \lambda_1|\boldsymbol{\beta}|} \Pi(\boldsymbol{\beta} \,|\, \theta)\mathrm{d}\,\boldsymbol{\beta}.$$

Denote now $\mathcal{B} = \{\boldsymbol{\beta} \in E^\star : ||\boldsymbol{X}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)|| > R\}$. Then

$$\begin{aligned}\Pi(\mathcal{B} \,|\, \boldsymbol{Y}, \theta)\mathbb{I}_{\tau_0} &\leq \frac{p^{2q}\mathrm{e}^{1+D}}{\pi(q \,|\, \theta)}\mathrm{e}^{\frac{16\bar{\lambda}q(2+K)}{n[\bar{\phi}(2q+Kq)]^2}}\mathrm{e}^{-\frac{R^2}{4}} \int_{\mathcal{B}} \mathrm{e}^{-\bar{\lambda}|\boldsymbol{\beta}-\boldsymbol{\beta}_0| + \lambda_1|\boldsymbol{\beta}|} \Pi(\boldsymbol{\beta} \,|\, \theta)\mathrm{d}\,\boldsymbol{\beta} \\ &\leq \frac{p^{2q}\mathrm{e}^{1+D}}{\pi(q \,|\, \theta)}\mathrm{e}^{\frac{16\bar{\lambda}q(1+K)}{n[\bar{\phi}(2q+Kq)]^2}}\mathrm{e}^{-\frac{R^2}{4}}\mathrm{e}^{\lambda_1\delta p}\sum_{k=0}^{p} 8^k \pi(k \,|\, \theta)\end{aligned}$$

Now, because $\pi > \theta\mathrm{e}^{-\delta\lambda_1}$ and $\theta = 1/(p^d+1)$ we can write

$$\pi(q \,|\, \theta) > \pi(q-1 \,|\, \theta)\frac{\mathrm{e}^{-\delta\lambda_1}}{p^d} > \pi(q-1 \,|\, \theta)\frac{C_2}{p^d},$$

where we used the fact $\mathrm{e}^{-\delta\lambda_1} > C_2$. Thus, $\pi(q \,|\, \theta) > C_2^q/p^{d\,q}\pi(0 \,|\, \theta)$. Thereby,

$$\Pi(\mathcal{B} \,|\, \boldsymbol{Y}, \theta)\mathbb{I}_{\tau_0} \leq C_2^{-q}p^{q(2+d)}\mathrm{e}^{1+D}\mathrm{e}^{\frac{16\bar{\lambda}^2 q(2+K)}{n[\bar{\phi}(2q+Kq)]^2}}\mathrm{e}^{-\frac{R^2}{4}}\mathrm{e}^{\lambda_1\delta p}\sum_{k=0}^{p}\left(\frac{8}{p^{d-1}}\right)^k.$$

This quantity will tend to zero for

$$R^2 \succeq 4q(2+d)\log p + \frac{16\bar{\lambda}^2 q(2+K)}{n[\bar{\phi}(2q+Kq)]^2} \succeq \frac{q(2+K)\log p}{[\bar{\phi}(2q+Kq)]^2}.$$

This proofs the first assertion. The second assertion follows from

$$\bar{\lambda}|\boldsymbol{\beta}-\boldsymbol{\beta}_0| \leq ||\boldsymbol{X}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)||^2 + \frac{\bar{\lambda}^2 q(2+K)}{2n\bar{\phi}(2q+Kq)^2}$$

and the last one from the definition of a minimal eigenvalue restricted to $\boldsymbol{\beta}$ of generalized dimensionality at most s in (8.7), which yields $||\boldsymbol{X}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)|| > \bar{c}(2q+Kq)||\boldsymbol{X}||\,||\boldsymbol{\beta}-\boldsymbol{\beta}_0||$.

□

References

Abramovich, F. and Grinshtein, V. (2010), "MAP model selection in Gaussian regression," *Electronic Journal of Statistics*, 4, 932–949.

Armero, C. and Bayarri, M. (1994), "Prior assessments in prediction in queues," *The Statistician*, 45, 139–153.

Bhattacharya, A., Pati, D., Pillai, N., and Dunson, D. (2015), "Dirichlet-laplace priors for optimal shrinkage," *Journal of the American Statistical Association (to appear)*.

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.

Bühlmann, P. and van der Geer, S. (2011), *Statistics for High-Dimensional Data*, Springer Series in Statistics.

Candes, E., Wakin, M., and Boyd, S. (2008), "Enhancing sparsity by reweighted l_1 minimization," *Journal of Fourier Analysis and Applications*, 14, 877–905.

Castillo, I., Schmidt-Hieber, J., and van der Vaart, A. (2015), "Bayesian linear regression with sparse priors," *The Annals of Statistics*, 43, 1986–2018.

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.

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.

Fan, Y. and Lv, J. (2014), "Asymptotic properties for combined L1 and concave regularization," *Biometrika*, 101, 57–70.

Fan, J., Zue, L., and Zou, H. (2014), "Strong oracle optimality of folded concave penalized estimation," *The Annals of Statistics*, 42, 819–849.

Friedman, J., Hastie, T., and Tibshirani, R. (2010), "Regularization paths for generalized linear models via coordinate descent," *Journal of Statistical Software*, 22, 1–22.

George, E. I. and Foster, D. (1997), "Calibration and empirical Bayes variable selection," *Biometrika*, 87, 731–747.

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

Gradshteyn, I. and Ryzhik, E. (2000), *Table of Integrals Series and Products*, Academic Press.

Hans, C. (2009), "Bayesian LASSO regression," *Biometrika*, 96, 835–845.

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.

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.

Liu, Y. and Wu, Y. (2007), "Variable selection via a combination of the L0 and L1 penalties," *Journal of Computational and Graphical Statistics*, 16, 782–798.

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.

Lv, J. and Fan, Y. (2009), "A unified approach to model selection and sparse recovery using regularized least squares," *The Annals of Statistics*, 37, 3498–3528.

Martin, R. and Walker, S. (2014a), "Asymptotically minimax empirical Bayes estimation of a sparse normal mean vector," *Electronic Journal of Statistics*, 8, 2188–2206.

Martin, R. and Walker, S. (2014b), "Empirical bayes posterior concentration in sparse high-dimensional linear models," *Submitted manuscript*.

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

Moreno, E., Girón, J., and Casella, G. (2015), "Posterior model consistency in variable selection as the model dimension grows," *Statistical Science*, 30, 228–241.

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

Rockova, V. (2015), "Bayesian estimation of sparse signals with a continuous spike-and-slab prior," *Submitted manuscript*.

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

Rockova, V. and George, G. (2015a), "Bayesian penalty mixing: The case of a non-separable penalty," *Manuscript*.

Rockova, V. and George, E. (2015b), "Fast Bayesian factor analysis via automatic rotations to sparsity," *Journal of the American Statistical Association (in revision)*.

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.

She, Y. (2009), "Thresholding-based iterative selection procedures for model selection and shrinkage," *Electronic Journal of Statistics*, 3, 384–415.

Su, W. and Candes, E. (2015), "Slope is adaptive to unknown sparsity and asymptotically minimax," *Submitted manuscript.*

Tibshirani, R. (1994), "Regression shrinkage and selection via the LASSO," *Journal of the Royal Statistical Society. Series B*, 58, 267–288.

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.

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.

Yuan, M. and Lin, Y. (2006), "Model selection and estimation in regression with grouped variables," *Journal of the Royal Statistical Society. Series B*, 68, 49–67.

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

Zhang, C. and Zhang, T. (2012), "A general theory of concave regularization for high-dimensional sparse estimation problems," *Statistical Science*, 27, 576–593.

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

Zou, H. and Li, R. (2008), "One-step sparse estimates in nonconcave penalized likelihood models," *The Annals of Statistics*, 36, 1509–1533.