Bayesian Penalty Mixing: The Case of a Non-separable Penalty

Veronika Ročková and Edward I. George

Abstract Separable penalties for sparse vector recovery are plentiful throughout statistical methodology and theory. Here, we confine attention to the problem of estimating sparse high-dimensional normal means. Separable penalized likelihood estimators are known to have a Bayesian interpretation as posterior modes under independent product priors. Such estimators can achieve rate-minimax performance when the correct level of sparsity is known. A fully Bayes approach, on the other hand, mixes the product priors over a shared complexity parameter. These constructions can yield a self-adaptive posterior that achieves rate-minimax performance when the sparsity level is unknown. Such optimality has also been established for posterior mean functionals. However, less is known about posterior modes in these setups. Ultimately, the mixing priors render the coordinates dependent through a penalty that is no longer separable. By tying the coordinates together, the hope is to gain adaptivity and achieve automatic hyperparameter tuning. Here, we study two examples of fully Bayes penalties: the fully Bayes LASSO and the fully Bayes Spike-and-Slab LASSO of Ročková and George (2015b). We discuss discrepancies and highlight the benefits of the two-group prior variant. We develop an Appell function apparatus for coping with adaptive selection thresholds. We show that the fully Bayes treatment of a complexity parameter is tantamount to oracle hyperparameter choice for sparse normal mean estimation.

Veronika Ročková
Statistics Department, The Wharton School of the University of Pennsylvania, Philadelphia, PA, 19104, USA, e-mail: vrockova@wharton.upenn.edu.

Edward I. George
Statistics Department, The Wharton School of the University of Pennsylvania, Philadelphia, PA, 19104, USA, e-mail: edgeorge@wharton.upenn.edu.

1 Introduction

1.1 Separable versus Non-separable

Separable penalty functions are limited by their inability to adapt to common features across model parameters because they treat these parameters independently. Non-separable penalties, on the other hand, can harvest structural knowledge such as groupings, networks or temporal orderings, to induce similarity across related components. The Fused LASSO (Tibshirani et al., 2005), the group LASSO (Meier et al., 2008), OSCAR (Bondell and Reich, 2008) are just a few prominent examples, which exploit structural similarities among parameters. Our interest in non-separable penalties here is fundamentally different. We want to explore the degree of adaptivity to unknown sparsity levels that can be achieved with exchangeable non-separable penalties when estimating sparse signals. To this end, we take a fully Bayesian perspective on penalized likelihood estimation of sparse normal means.

We consider the classic problem of estimating a mean vector from a single multivariate observation. With $y = (y_1, \ldots, y_n)'$ that arises from

$$y_i = \beta_{0i} + \varepsilon_i, \quad \text{where} \quad \varepsilon_i \overset{iid}{\sim} \mathcal{N}(0,1), \quad i = 1, \ldots, n, \tag{1}$$

independently, the goal is to estimate $\beta_0 = (\beta_{01}, \ldots, \beta_{0n})'$ under squared error loss, assuming that β_0 is possibly sparse. Throughout the paper, sparsity is understood as a requirement that $p_n = o(n)$ as $n \to \infty$, where $p_n = \|\beta_0\|_0$. The quality of recovery of β_0 is often assessed relative to the benchmark minimax risk $2\,p_n \log(n/p_n)(1+o(1))$ (Donoho et al., 1992) and the "near-minimax" risk $2\,p_n \log n(1+o(1))$, a perspective adopted here.

By regularizing towards sparsity, superbly behaving estimators can be obtained through maximization of the penalized likelihood

$$\widehat{\beta} = \arg\max_{\beta \in \mathbb{R}^n} \left\{ -\frac{1}{2}\sum_{i=1}^{n}(y_i - \beta_i)^2 + Pen_\lambda(\beta) \right\}, \tag{2}$$

where $Pen_\lambda(\beta)$ is a regularizing penalty and $\lambda > 0$ is a user-specified penalty parameter. The known equivalence between penalized likelihood estimators and Bayesian posterior modes is obtained by associating a penalty function with a (possibly improper) prior $\pi(\beta \mid \lambda)$ via $Pen_\lambda(\beta) = \log \pi(\beta \mid \lambda)$. Particularly appealing have been penalties that are separable in the sense: $Pen_\lambda(\beta) = \sum_{i=1}^n pen_\lambda(\beta_i)$. These correspond to independent product priors $\pi(\beta \mid \lambda) = \prod_{i=1}^n \pi(\beta_i \mid \lambda)$. Familiar examples of separable penalties include the ℓ_0 variable selection penalty $-\sum \lambda\, \delta_0(\beta_i)$, its closest concave relative the ℓ_1 LASSO penalty $-\sum \lambda |\beta_i|$ (Tibshirani, 1994) and the plentiful LASSO variants (Zou, 2006; Zou and Hastie, 2005). Recently there has been a surge of interest in separable penalties that are non-concave (SCAD of Fan and Li (2001), MC+ of Zhang (2010), log penalty of Friedman (2008)). Such penalties

eliminate modeling bias and can achieve more refined recovery rates (Wang et al., 2014).

Separable penalized likelihood estimators can achieve the minimax risk (up to a constant) with a suitable choice of λ that typically depends on p_n. Removing the need to assume that p_n is known, adaptive estimators can be obtained with empirical Bayes and fully Bayes approaches. The latter is of interest to us here.

To move beyond the independent product priors that give rise to separable penalties, more flexible penalty functions can be obtained from fully Bayes hierarchical prior constructions. Such priors are obtained by mixing over a shared parameter,

$$\pi(\beta) = \int \prod_{i=1}^{n} \pi(\beta_i \,|\, \eta)\pi(\eta)\mathrm{d}\,\eta, \tag{3}$$

where $\pi(\eta)$ is a prior distribution. The coefficients $\beta_1, \ldots, \beta_n$ are now conditionally independent given η. Appropriately formulated, such exchangeable priors can be used to create multivariate estimators that "borrow strength" across the components of β.

Such estimators have a long tradition. Perhaps the best known examples are the variety of posterior mean Stein estimators that can be motivated with hierarchical mixtures of normal priors. For example, the construction (3) with $\pi(\beta_i \,|\, \eta) = N(0,\eta)$ and uniform $\pi(\eta)$ over $(0,\infty)$, yields the harmonic prior $\pi_H(\beta) = \left(\sum_i \beta_i^2\right)^{-n-2}$ under which the components of β are no longer independent. With this prior, the formal Bayes Stein estimator $\widehat{\beta}^H$ is known to be minimax in the classical sense of having expected squared error loss less than than the minimax value n, for all β whenever $n \geq 3$ (Stein, 1974). It is also known to be admissible (Brown, 1971). To appreciate the adaptive shrinkage behavior of $\widehat{\beta}^H$, note that a large fixed value of η would cause the posterior mean to shrink very little, whereas a very small η would cause the posterior mean to shrink dramatically. By introducing a prior on η, $\widehat{\beta}^H$ adaptively shrinks more towards 0 when smaller values of η (and $\sum_i \beta_i^2$) are supported by the data through the posterior. Further elaborations of (3) can be used to obtain generalized Stein estimators that adaptively exploit ensemble structure among the components of β. For example, a location mixture of normals for each component of β, namely $\pi(\beta_i \,|\, \eta_1, \ldots, \eta_K) = \sum_{k=1}^{K} w_k N(u_{ik}, \eta_k)$ with independent uniform priors on the η_k, leads to a minimax multiple shrinkage estimator that adaptively shrinks towards the prior mean best supported by the data (George, 1968b,a). There are many possibilities for the construction of such estimators, which offer substantial risk reduction in the vicinity of their shrinkage targets.

The hierarchical normal priors for posterior mean Stein estimation could also be used to generate non-separable penalty functions $\log \pi(\beta)$ to obtain posterior mode estimators. Sometimes these modal estimators will be the same as the mean estimators, as occurs under the harmonic prior, but in general they will differ. In any case, the classical minimax guarantee offered by Stein estimators is less important in sparse settings where $p_n = o(n)$, as protection against all β is no longer needed. Instead, one can focus on obtaining risk reduction in the narrower region of interest

where p_n is small. By using posterior mode estimators that adaptively threshold irrelevant coefficients in such regions, one can aim for the asymptotic minimax value $2\,p_n \log(n/p_n)$, a vast improvement over n.

1.2 Bayesian Penalty Mixing

As opposed to the simple addition of penalty terms that typically occurs with separable penalties, prior constructions such as (3) exemplify another route for penalty construction. This general approach, which we call Bayesian penalty mixing, is a strategy that combines penalty functions via mixtures of their underlying prior distributions. More precisely, let $Pen(\beta \mid \eta) = \log \pi(\beta \mid \eta)$ be a set of penalties corresponding to a set of priors $\pi(\beta \mid \eta)$ indexed by η. Bayesian penalty mixing entails the creation of a penalty

$$Pen(\beta) = \log \pi(\beta) = \log \int \pi(\beta \mid \eta)\pi(\eta)\, d\eta$$

induced by mixing the underlying priors over $\pi(\eta)$. Note that this mixing occurs in the space of priors rather than the space of penalties.

As is well-known, the derivatives of $Pen(\beta)$ and $Pen(\beta \mid \eta)$ play a crucial role as the "bias terms" of the penalized likelihood solution $\widehat{\beta}$ to (2), (Fan and Li, 2001). The adaptive potential of Bayesian penalty mixing is reflected in the relationship between these derivatives given by the following fundamental identity

$$\frac{\partial Pen(\beta)}{\partial |\beta_i|} = \int \frac{\partial Pen(\beta \mid \eta)}{\partial |\beta_i|} \pi(\eta \mid \beta)\, \mathrm{d}\,\eta \tag{4}$$

which follows from

$$\frac{\partial \log \pi(\beta)}{\partial |\beta_i|} = \frac{1}{\pi(\beta)} \int \frac{\partial \pi(\beta \mid \eta)}{\partial |\beta_i|} \pi(\eta) \mathrm{d}\,\eta = \int \frac{\partial \log \pi(\beta \mid \eta)}{\partial |\beta_i|} \pi(\eta \mid \beta) \mathrm{d}\,\eta.$$

Thus the bias induced by $Pen(\beta)$ will be an adaptive convex combination of the biases induced by each of the combined penalties $Pen(\beta \mid \eta)$. More weight is put on shrinkage terms that are better supported by $\pi(\eta \mid \beta)$. In this way, $Pen(\beta)$ automatically emphasizes those shrinkage terms which are better suited for each β.

The principal thrust of this paper is to demonstrate the potential of Bayesian penalty mixing through three examples:

- A fully Bayes LASSO penalty obtained by mixing LASSO penalties *across* coordinates, (Section 2).
- A Spike-and-Slab LASSO penalty obtained by mixing pairs of LASSO penalties *within* coordinates, (Section 3).
- A fully Bayes Spike-and-Slab LASSO penalty obtained by mixing LASSO penalties both *within and across* coordinates, (Section 4).

These Spike-and-Slab LASSO examples complement the developments of Ročková (2015), who introduced Spike-and-Slab LASSO priors, and Ročková and George (2015a,b), who further developed and applied these priors for variable selection.

In each of these three examples, we pay particular attention to the ability of the modal estimator to adapt to the unknown sparsity level (a property known to hold for a posterior mean in similar spike-and-slab setups). Our exploration concludes with a very important finding. By mixing within coordinates and then adding a prior distribution over a complexity parameter, one achieves a level of adaptivity that is tantamount to an oracle hyperparameter choice. Our analysis constitutes an initial step towards the general development of more elaborate hierarchical penalty constructions for asymptotically minimax penalized likelihood estimation with unknown p_n.

In Section 2, we study the fully Bayes variant of the LASSO. Section 3 and 4 review the Spike-and-Slab LASSO and its fully Bayes variant. Section 5 then develops an Appell function apparatus to deal with expectations under a certain class of generalized beta distributions. Section 6 shows the adaptability of the fully Bayes Spike-and-Slab LASSO in sparse normal means. Section 7 concludes with a discussion of future directions.

The following notation will be used throughout. For sequences a_n and b_n, $a_n \sim b_n$ means $a_n/b_n \to c$ for some $c > 0$, $a_n \succeq b_n$ means $b_n = \mathcal{O}(a_n)$, $a_n \succ b_n$ means $b_n = o(a_n)$. We will denote by $|\cdot|$ the ℓ_1 norm.

2 Fully Bayes LASSO

The ℓ_1 penalty $Pen_\lambda(\beta) = -\lambda|\beta|$, arguably the most prominent separable regularizer, gives rise to the LASSO estimator $\widehat{\beta}^L$ in (2). With the universal threshold choice $\lambda \sim \sqrt{2\log n}$, the LASSO estimator enjoys oracle properties (Donoho and Johnstone 1994) and achieves a near-minimax risk rate when $p_n = o(n)$. In practice, however, the parameter λ is often chosen by performance based criteria such as cross-validation, generalized cross-validation or ideas based on Stein's unbiased estimate of risk (Tibshirani, 1994). Park and Casella (2008) offer Bayesian alternatives to choosing λ: an empirical Bayes strategy through marginal maximum likelihood and a fully Bayes solution with an appropriate prior distribution $\pi(\lambda)$. All of these approaches use the data to inform the penalty hyperparameter, thereby potentially boosting performance. Whereas Park and Casella (2008) used fully Bayes for posterior median estimation via MCMC, here we investigate the implications for direct penalized likelihood estimation.

For any given prior distribution $\pi(\lambda)$, a fully-Bayes-LASSO penalty is defined as

$$Pen_{FL}(\beta) = \log \int_0^\infty \left(\frac{\lambda}{2}\right)^n \mathrm{e}^{-\lambda|\beta|} \mathrm{d}\,\pi(\lambda). \tag{5}$$

Except for trivial point-mass priors $\pi(\lambda)$, (5) is a non-separable penalty. The well-known closed-form solution for the LASSO estimator, which treats λ as fixed,

writes $\widehat{\beta}^L = (\widehat{\beta}^L_1, \ldots, \widehat{\beta}^L_n)'$ as

$$\widehat{\beta}^L_i = (|y_i| - \lambda)_+ \text{sign}(y_i). \tag{6}$$

This follows from the KKT conditions, using the fact that $-\lambda$ is the derivative of the LASSO penalty. With the fully-Bayes-LASSO penalty, the derivative is instead

$$\frac{\partial Pen_{FL}(\beta)}{\partial |\beta_i|} = -\frac{\int_0^\infty \lambda \left(\frac{\lambda}{2}\right)^n e^{-\lambda|\beta|} d\pi(\lambda)}{\int_0^\infty \left(\frac{\lambda}{2}\right)^n e^{-\lambda|\beta|} d\pi(\lambda)} = -\mathsf{E}[\lambda \mid \beta],$$

a special case of (4). Thereby, the fully-Bayes-LASSO estimator $\widehat{\beta}^{FL} = (\widehat{\beta}^{FL}_1, \ldots, \widehat{\beta}^{FL}_n)'$ satisfies

$$\widehat{\beta}^{FL}_i = \left[|y_i| - \mathsf{E}\left(\lambda \mid \widehat{\beta}^{FL}\right)\right]_+ \text{sign}(y_i). \tag{7}$$

Here, the shrinkage term is adaptive, depending on the data through $\widehat{\beta}^{FL}$. Interestingly, the fully Bayes approach is seen to manifest itself through the plug-in choice $\mathsf{E}[\lambda \mid \widehat{\beta}^{FL}]$ in empirical Bayes-like fashion. Note, however, that $\widehat{\beta}^{FL}$ is contained on both sides of (7), in contrast with approaches which insert a single derived estimate for λ. With the fully Bayes plug-in, the amount of shrinkage reflects the size of $|\widehat{\beta}^{FL}|$, as will be seen below.

It can be shown[1] that the risk of the LASSO estimator $\widehat{\beta}^L$ satisfies

$$\mathsf{E}_{\beta_0} ||\widehat{\beta}^L - \beta_0||^2 \leq p_n(2 + 4\lambda^2) + (n - p_n) 4\lambda \phi(\lambda), \tag{8}$$

when $||\beta_0||_0 = p_n$. In order to obtain the near-minimax rate, one would need to select $\lambda \sim \sqrt{2 \log n}$ so that $\mathsf{E}_{\beta_0} ||\widehat{\beta}^L - \beta_0||^2 \preceq p_n \lambda^2 = 2\, p_n \log n$.

Analogously, it can be shown that the fully-Bayes-LASSO estimator satisfies

$$\mathsf{E}_{\beta_0} ||\widehat{\beta}^{FL} - \beta_0||^2 \leq p_n(2 + 4\mathsf{E}_{\beta_0}\widehat{\lambda}^2) + (n - p_n) 4\mathsf{E}_{\beta_0}\widehat{\lambda}\phi(\widehat{\lambda}),$$

where $\widehat{\lambda} = \mathsf{E}\left(\lambda \mid \widehat{\beta}^{FL}\right)$. One would hope that the fully Bayes variant will self-adapt, yielding ideally $\mathsf{E}_{\beta_0}\widehat{\lambda} \sim \sqrt{2 \log n}$, the desired LASSO tuning. To get a sense of this possibility, let's take a closer look at $\widehat{\lambda}$.

For the purpose of illustration, consider the conjugate class[2] of priors $\pi(\lambda) = \frac{b^a}{\Gamma(a)} \lambda^{a-1} e^{-b\lambda} \mathbb{I}(\lambda > 0)$, so that

[1] Following the proving technique of Ročková (2015), Remark 5.1. This upper bound is useful for illustration and can be sharpened.

[2] Casella and Park (2009) use a gamma prior on λ^2 for the ease of MCMC implementation.

$$\pi(\beta) = \frac{b^a}{2^n\Gamma(a)} \int_0^\infty \lambda^{n+a-1} \mathrm{e}^{-\lambda(|\beta|+b)} \mathrm{d}\lambda = \frac{b^a\Gamma(n+a)}{2^n\Gamma(a)(|\beta|+b)^{n+a}}.$$

This yields

$$\widehat{\lambda} = \mathsf{E}[\lambda \,|\, \widehat{\beta}^{FL}] = \frac{n+a}{|\widehat{\beta}^{FL}|+b}. \tag{9}$$

Thus $\widehat{\lambda}$ is an adaptive threshold that depends not only on the sparsity of $\widehat{\beta}^{FL}$, but also on the size of the nonzero coefficients. The dependence on $|\widehat{\beta}^{FL}|$, and hence on $|\beta_0|$, makes it impossible to tune a and b so that $\mathsf{E}_{\beta_0}\widehat{\lambda} \sim \sqrt{2\log n}$. Thus, there seems to be a disconnect between the universal LASSO tuning for near-minimax rates, and the fully Bayes LASSO approach.

Moreover, note that in (7), all coordinates are shrunk by the same amount. As opposed to the classical LASSO, the adaptive weight here depends on the size of the coefficients $\widehat{\beta}$. With just one tuning-parameter, despite its adaptivity, the penalty will fail to shrink globally and act locally (Polson and Scott, 2010). As a locally adaptive extension of the LASSO prior, Griffin and Brown (2012) propose individual penalties λ_i for each coordinate, where $\lambda_i \sim \pi(\lambda)$. Each λ_i is a realization from a mixing distribution. Marginally, such strategies induce sharply peaked separable penalties.

Although the fully Bayes treatment of the LASSO offers the possibility for self-tuning, the penalty $\mathsf{E}(\lambda \,|\, \widehat{\beta}^{FL})$ cannot both circumvent the bias issue and adapt to sparsity at the same time. For this, it is imperative that each coefficient is given a unique opportunity to escape the overall shrinkage effect. This is achieved with global/local shrinkage priors and two-group spike-and-slab priors. In the sequel, we explore the two group mixture prior formulations for penalty creation. Instead of mixing across coordinates immediately, we begin by mixing within coordinates.

3 Spike-and-Slab LASSO

Penalty mixing within coordinates provides an opportunity to enhance the performance of penalized likelihood estimators (2) by combining the benefits of single penalties. For instance, the elastic net (Zou and Hastie, 2005) blends the LASSO and the ridge, inducing a grouping effect in estimation while maintaining the ability to threshold. Other combinations of convex and concave penalties have been proposed to achieve good subset recovery and bias elimination (Fan and Lv, 2014). In these two cases, the mixing operates at the penalty level through the linear superposition of penalty terms.

In sharp contrast, Bayesian penalty mixing, which operates at the prior level, is inherently nonlinear and probabilistic. Such Bayesian penalty mixing arises naturally in the context of spike-and-slab mixture formulations for Bayesian variable selection. In particular, let us turn to the recently proposed Spike-and-Slab LASSO (SSL) prior of Ročková (2015). Viewed as a continuous relaxation of the point mass mixture prior (Castillo and van der Vaart (2012)), the SSL prior deploys a two-point

mixture of a Laplace spike distribution $\pi(\beta \mid \lambda_0) = \frac{\lambda_0}{2} e^{-|\beta|\lambda_0}$ and a Laplace slab $\pi(\beta \mid \lambda_1) = \frac{\lambda_1}{2} e^{-|\beta|\lambda_1}$, i.e.

$$\pi(\beta \mid \theta, \lambda_1, \lambda_0) = \prod_{i=1}^{n} [\theta \pi(\beta_i \mid \lambda_1) + (1-\theta)\pi(\beta_i \mid \lambda_0)],$$

where $\theta \in (0,1)$ is a mixing weight. It is assumed $\lambda_0 >> \lambda_1 > 0$. Note that the SSL prior reduces to a single Laplace (LASSO) prior when $\lambda_0 = \lambda_1$, and converges to a point mass spike-and-slab mixture as $\lambda_0 \to \infty$. Here we have three parameters $(\lambda_1, \lambda_0, \theta)$, where $0 < \lambda_1 < 1$ is a small fixed constant treated as known (made precise by the theoretical study of Ročková (2015)) and λ_0 and θ are the two parameters subject to tuning. In what follows, we suppress λ_1 from the conditioning.

Tailored for Bayesian variable selection (Ročková and George, 2014; Ročková, 2015; Ročková and George, 2015b), each β_i is thought of as arriving from the diffuse slab distribution $\pi(\beta_i \mid \lambda_1)$, with probability $\theta \in (0,1)$, or from the sharply peaked spike distribution $\pi(\beta_i \mid \lambda_0)$. In the context of Bayesian regression analysis, selection with a spike-and-slab prior has traditionally entailed a decision to include the ith regressor only when the "posterior probability that β_i came from $\pi(\beta_i \mid \lambda_1)$" is high. However, when approached from a penalized likelihood perspective, the selection strategy is very different.

In the context of penalized likelihood estimation, variable selection capitalizes on a penalty of the form

$$Pen_{SL}(\beta \mid \theta, \lambda_0) = \sum_{i=1}^{n} \log[\theta \pi(\beta_i \mid \lambda_1) + (1-\theta)\pi(\beta_i \mid \lambda_0)]. \tag{10}$$

The penalty (10) stands out from classical penalty functions as nonlinear in both $|\beta_i|$ and $(\lambda_1, \lambda_0, \theta)$. The benefits of using this rather complex nonlinear functional are revealed by its implicit shrinkage term, the derivative. As a special case of (4), we obtain

$$\frac{\partial Pen_{SL}(\beta \mid \theta, \lambda_0)}{\partial |\beta_i|} = -\lambda_1 p^\star_\theta(\beta_i) - \lambda_0[1 - p^\star_\theta(\beta_i)] \equiv -\lambda^\star_\theta(\beta_i), \tag{11}$$

where

$$p^\star_\theta(\beta_i) = \frac{\theta \pi(\beta_i \mid \lambda_1)}{\theta \pi(\beta_i \mid \lambda_1) + (1-\theta)\pi(\beta_i \mid \lambda_0)} = \frac{1}{1 + \frac{1-\theta}{\theta} \frac{\lambda_0}{\lambda_1} e^{-|\beta_i|(\lambda_0 - \lambda_1)}} \tag{12}$$

is the conditional probability that β_i came from $\pi(\beta_i \mid \lambda_1)$ having seen β_i. Thus, the spike-and-slab LASSO shrinkage term $\lambda^\star_\theta(\beta_i)$ mixes the two LASSO shrinkage terms (λ_1, λ_0) and does so adaptively. Ideally, one would like to use $\lambda_1 < 1$ when $|\beta_i|$ is large, and $\lambda_0 >> 1$ when $|\beta_i|$ is small. This is essentially effectuated with the adaptive exponential weight $0 < p^\star_\theta(\beta_i) < 1$, which gears $\lambda^\star_\theta(\beta_i)$ towards λ_0, when $|\beta_i|$ is small, and towards λ_1 when $|\beta_i|$ is large.

The first order KKT conditions immediately yield a necessary characterization of the global mode $\widehat{\beta}^{SL}$, where the coordinate estimators must satisfy

$$\widehat{\beta}_i^{SL} = \Big[|y_i| - \lambda_\theta^\star(\widehat{\beta}_i^{SL})\Big]_+ \operatorname{sign}(y_i),\ i = 1,\ldots,n. \tag{13}$$

The form (13) resembles the LASSO estimator (6). However, here the SSL penalty term $\lambda_\theta^\star(\widehat{\beta}_i^{SL})$ is not constant, but rather depends on the data through $\widehat{\beta}_i^{SL}$. As with the adaptive LASSO, (Zou, 2006), each coefficient has its own penalty, which serves as a basis for "differential" shrinkage. Adaptive in this way, the SSL estimator deploys a large penalty (close to λ_0) to threshold small estimates, and a small penalty (close to λ_1) to hold large estimates steady with only slight bias. This is very much in contrast to a single LASSO estimator which shrinks all estimates equally with a constant penalty.

To further eliminate local modal solutions, the necessary characterization (13) can be sharpened (Ročková, 2015), where $\widehat{\beta}^{SL}$ can be shown to be sandwiched between soft and hard thresholding operators. When λ_0 is large enough, the coordinate estimates must satisfy

$$\widehat{\beta}_i^{SL} = \Big[|y_i| - \lambda_\theta^\star(\widehat{\beta}_i^{SL})\Big] \operatorname{sign}(y_i)\, \mathrm{I}(|y_i| > \Delta),\ i = 1,\ldots,n, \tag{14}$$

where

$$\Delta \sim \sqrt{2\log[1/p_\theta^\star(0)]} + \lambda_1, \tag{15}$$

when $\lambda_1 \leq \mathrm{e}^{-2}$ and $\lambda_0 \geq 1/\theta + 3$, and $0 < \theta \leq 0.5$. Thus, letting λ_0 increase with the sample size, the asymptotic risk of $\widehat{\beta}^{SL}$ is ultimately governed by the threshold Δ through a functional

$$1/p_\theta^\star(0) = 1 + \frac{\lambda_0}{\lambda_1}\frac{(1-\theta)}{\theta}. \tag{16}$$

It follows from the proof of Theorem 5.1 of (Ročková, 2015) that

$$\mathsf{E}_{\beta_0}||\widehat{\beta}^{SL} - \beta_0||^2 \preceq p_n\Delta^2 + (n-p_n)\Delta\phi(\Delta). \tag{17}$$

This upper bound is similar to (8), where Δ now plays the role of selection threshold. The rate-minimax performance of the global posterior mode $\widehat{\beta}^{SL}$ will be then achieved when $1/p_\theta^\star(0) \sim (n/p_n)^\alpha$ for some $\alpha > 0$. This will occur, for instance, when $\lambda_1 < \mathrm{e}^{-2}$, $\theta \sim p_n/n$ and $\lambda_0 \sim 1/\theta$. Under this "oracle" choice, the threshold Δ is of the optimal order $\sqrt{2\log[n/p_n]}$, yielding (Corollary 5.1 of Ročková (2015))

$$\sup_{\beta_0 \in l_0[p_n;n]} \mathsf{E}_{\beta_0}||\widehat{\beta}^{SL} - \beta_0||^2 \sim p_n \log(n/p_n).$$

Moreover, with this parameter choice, the entire posterior distribution $\pi(\beta \,|\, y)$ concentrates around the true vector β_0 at the minimax rate (Corollary 6.1 of Ročková

(2015)). Namely,

$$\sup_{\beta_0 \in l_0[p_n;n]} \mathsf{E}_{\beta_0} \mathsf{P}(\beta \in \mathbb{R}^n : ||\beta - \beta_0||^2 > M\, p_n \log[n/p_n] \,|\, y^{(n)}) \to 0 \qquad (18)$$

for a suitably large constant $M > 0$. This property fails to hold for a single LASSO Laplace prior (Castillo et al. (2015)), despite the optimality of the LASSO mode.

Nevertheless, the oracle parameter setup will be unavailable when p_n is unknown, as will typically be the case in practice. To get around this problem, one could consider an empirical Bayes plug-in choice for θ and λ_0 (along the lines of Johnstone and Silverman (2004)). However, here we explore a fully Bayes approach which places a prior distribution on θ. The hope is that by treating θ as unknown in this way, we can mimic the performance obtained by the oracle choice.

It is important to note that the spike penalty λ_0 also requires proper tuning. Here, we focus exclusively on adapting θ, assuming that λ_0 satisfies $\lambda_0/\lambda_1 \sim n^{\alpha}$ for $\alpha > 0$). This scenario is instructive for understanding the basic mechanisms underlying the adaptivity of fully Bayes penalty mixing. It is also a fundamental first step towards fully adaptive procedures, obtained by inducing a prior on λ_0. We discuss some of these alternatives at the end of Section 6.

4 Fully Bayes Spike-and-Slab LASSO

In Section 2, we considered a fully Bayes treatment of the original LASSO, mixing its single penalty parameter over a prior distribution. This mixing produces a data-dependent penalty, which ties the coordinates together. In Section 3, we considered a very different approach, mixing penalties within coordinates over a spike-and-slab prior, keeping the value of the shared hyperparameter θ fixed. This mixing treated each coefficient uniquely, preventing the borrowing of strength across coordinates. We now proceed to combine these two perspectives, by mixing the spike-and-slab prior over a prior distribution on θ. This will allow the coordinates to share a global hyperparameter, while the spike-and-slab prior allows each coefficient to be treated locally.

Adaptive mixing of spike-and-slab penalties across the coordinates of β_i can be achieved with a suitable prior $\pi(\theta)$ (Ročková and George, 2015b). The prior mixing proportion θ plays the role of a complexity parameter, determining the sparsity of the solution. In the absence of knowledge about p_n, setting θ fixed to a constant may diminish performance by over/underestimating the dimensionality. The hope is that by introducing $\pi(\theta)$, the penalty can adapt to the sparsity level and help yield oracle-like performance automatically, avoiding the need for setting θ close to p_n/n.

With a prior on θ, the coordinates in β are now marginally dependent, yielding

$$\pi(\beta) = \int_0^1 \prod_{i=1}^{n} [\theta\, \pi(\beta_i \,|\, \lambda_1) + (1-\theta)\pi(\beta_i \,|\, \lambda_0)]\, \mathrm{d}\, \pi(\theta) \qquad (19)$$

which will be non-separable for all but trivial choices of $\pi(\theta)$. Generally, the integral above does not have a closed form solution, complicating the tractability of the penalty $Pen(\beta) = \log \pi(\beta)$. Fortunately, a revealing and simple form can still be found for its derivative.

With θ fixed, a simple necessary characterization of the global mode was obtained in (14). Following the development in Ročková and George (2015b), we can still obtain a similar simple characterization under the dependent prior $\pi(\beta)$. As another special case of (4), we have

$$\begin{aligned}\frac{\partial \log \pi(\beta)}{\partial |\beta_i|} &= \frac{1}{\pi(\beta)} \int_0^1 \frac{\partial \pi(\beta \,|\, \theta)}{\partial |\beta_i|} \pi(\theta) \mathrm{d}\,\theta = \int_0^1 \frac{\partial \log \pi(\beta \,|\, \theta)}{\partial |\beta_i|} \pi(\theta \,|\, \beta) \mathrm{d}\,\theta \\ &= -\lambda_1 \int_0^1 p^\star_\theta(\beta_i) \pi(\theta \,|\, \beta) \mathrm{d}\,\theta - \lambda_0 \left[1 - \int_0^1 p^\star_\theta(\beta_i) \pi(\theta \,|\, \beta) \mathrm{d}\,\theta \right]. \quad (20)\end{aligned}$$

The difference between (11) and (20) is illuminating. Instead of the "fixed-θ" mixing probability $p^\star_\theta(\beta_i)$ in (11), (20) employs an aggregated mixing probability $\int_0^1 p^\star_\theta(\beta_i) \pi(\theta \,|\, \beta) \mathrm{d}\,\theta$ that averages $p^\star_\theta(\beta_i)$ over $\pi(\theta \,|\, \beta)$. The implicit bias term (20) reveals the mechanism underlying the adaptivity. By averaging over the conditional distribution $\pi(\theta \,|\, \beta)$, the penalty is given an opportunity to learn about the level of sparsity of β.

Going further, substantial insight and simplification is provided by the following surprising identity from Ročková and George (2015b)

$$p^\star_{\theta_i}(\beta_i) \equiv \int_0^1 p^\star_\theta(\beta_i) \pi(\theta \,|\, \beta) \mathrm{d}\,\theta, \quad \theta_i \equiv E[\theta \,|\, \beta_{\setminus i}]. \quad (21)$$

Here $\beta_{\setminus i}$ denotes the sub-vector of β containing all but the ith entry. Even though $p^\star_\theta(\beta_i)$ is a nonlinear function of θ, its average value over $\pi(\theta \,|\, \beta)$ is obtained by simply substituting $\theta_i \equiv E[\theta \,|\, \beta_{\setminus i}]$ for θ to obtain $p^\star_{\theta_i}(\beta_i)$. With the representation (21), the implicit bias term (20) is of the simple form

$$\frac{\partial \log \pi(\beta)}{\partial |\beta_i|} = \lambda_1 \, p^\star_{\theta_i}(\beta_i) + \lambda_0 \left[1 - p^\star_{\theta_i}(\beta_i) \right] \equiv \lambda^\star_{\theta_i}(\beta_i), \quad (22)$$

which similarly comes from $\lambda^\star_\theta(\beta_i)$ in (11) by simple substitution of θ_i for θ.

A direct analogue of (14) is now readily available. The coordinates of the global modal estimator $\widehat{\beta}^{FSL}$ under the fully Bayes non-separable SSL penalty will jointly satisfy

$$\widehat{\beta}^{FSL}_i = \left[|y_i| - \lambda^\star_{\widehat{\theta}_i}(\widehat{\beta}^{FSL}_i) \right]_+ \mathrm{sign}(y_i) \, \mathrm{I}(|y_i| > \Delta_i), \quad (23)$$

where $\widehat{\theta}_i \equiv \mathsf{E}[\theta \,|\, \widehat{\beta}^{FSL}_{\setminus i}]$, and

$$\Delta_i \sim \sqrt{2 \log[1/p^\star_{\widehat{\theta}_i}(0)]} + \lambda_1. \quad (24)$$

Compared to equation (14), each coordinate in (23) now has a penalty $\lambda^{\star}_{\widehat{\theta}_i}(\widehat{\beta}_i^{FSL})$ which depends on all the coordinates through $\widehat{\beta}_i$ and $\widehat{\theta}_i$, not just the ith. This dependency originates from the mixing distribution $\pi(\theta \,|\, \widehat{\beta}^{FSL})$, used to obtain the "average" inclusion probability (21). The fewer zeros in $\widehat{\beta}^{FSL}$, the less concentrated this distribution will be around the origin, thereby leading to a larger value of $\widehat{\theta}_i$.

Another essential difference between (14) and (23), is the replacement of Δ by the adaptive thresholds Δ_i. In the separable case, we have a single threshold for which

$$(\Delta - \lambda_1)^2 \sim 2 \log \left[1 + \frac{\lambda_0}{\lambda_1} \frac{1-\theta}{\theta} \right]. \tag{25}$$

With θ fixed to a constant, (25) deploys *prior odds* of non-entering the model $(1-\theta)/\theta$. Substituting $\widehat{\theta}_i$ for θ, we have

$$(\Delta_i - \lambda_1)^2 \sim 2 \log \left[1 + \frac{\lambda_0}{\lambda_1} \frac{1 - \mathsf{E}(\theta \,|\, \widehat{\beta}^{FSL}_{\setminus i})}{\mathsf{E}(\theta \,|\, \widehat{\beta}^{FSL}_{\setminus i})} \right] \tag{26}$$

which shows how very different the non-separable formulation is from the separable one. By treating θ as random, (26) uses the "*posterior odds*" $\frac{[1-\mathsf{E}(\theta \,|\, \widehat{\beta}^{FSL}_{\setminus i})]}{\mathsf{E}(\theta \,|\, \widehat{\beta}^{FSL}_{\setminus i})}$. Through $\widehat{\beta}$, the data inform θ about the sparsity level through the proportion of its nonzero entries. This observation is fundamental to understanding the mechanism through which the fully Bayes formulation transmits information into the posterior mode.

5 Bounds and Rates for the Fully Bayes Selection Thresholds

The selection thresholds Δ_i are indispensable to describing the properties of the global mode estimator. For this purpose, Ročková and George (2015b) present bounds for the quantity $\mathsf{E}[\theta \,|\, \widehat{\beta}_{\setminus i}]$, and thereby also Δ_i when $\theta \sim \mathscr{B}(a,b)$. Note that the posterior expectations $\mathsf{E}(\theta \,|\, \widehat{\beta}_{\setminus i})$ will be very similar for each $i = 1, \ldots, n$, when n is sufficiently large. Thus, despite Δ_i's that are coordinate-specific, they will not be dramatically different. For a relatively uninformative choice of $\pi(\theta)$, one would expect $\mathsf{E}(\theta \,|\, \widehat{\beta}_{\setminus i})$ to be close to $\widehat{p}^{\,i}/n$, where $\widehat{p}^{\,i} = \|\widehat{\beta}_{\setminus i}\|_0$.

Let us assume that it is the first $\widehat{p}$ entries in $\widehat{\beta}$ that are nonzero. Under $\theta \sim \mathscr{B}(a,b)$, the density of the conditional distribution $\pi(\theta \,|\, \widehat{\beta})$ is given by

$$\pi(\theta \,|\, \widehat{\beta}) \propto \theta^{a-1}(1-\theta)^{b-1}(1-\theta z)^{n-\widehat{p}} \prod_{j=1}^{\widehat{p}} (1 - \theta x_j), \tag{27}$$

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

$$\mathsf{E}\left[\theta \,|\, \widehat{\beta}\right] = \frac{\int_0^1 \theta^a (1-\theta)^{b-1} (1-\theta z)^{n-\widehat{p}} \prod_{j=1}^{\widehat{p}} (1-\theta x_j)\,\mathrm{d}\,\theta}{\int_0^1 \theta^{a-1} (1-\theta)^{b-1} (1-\theta z)^{n-\widehat{p}} \prod_{j=1}^{\widehat{p}} (1-\theta x_j)\,\mathrm{d}\,\theta}. \tag{28}$$

Ročková and George (2015b) suggest approximating (28) by

$$\frac{\mathscr{B}(a+1,b)}{\mathscr{B}(a,b)} \frac{F_1(a+1,\widehat{p}-n,-\widehat{p},a+b+1;z,x)}{F_1(a,\widehat{p}-n,-\widehat{p},a+b;z,x)}, \tag{29}$$

for some suitable x, where

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

is the Appell F1 function. To obtain suitable lower and upper bounds for $\mathsf{E}\left[\theta \,|\, \widehat{\beta}\right]$, we begin with the following lemma which establishes that the ratio (29) is monotone in x and z.

Lemma 1. *(Monotonicity Ratio of Appell F1 Functions)*
Assume $\delta > 0$. Then the function

$$f_\delta(a',b',c',d';z,x) = \frac{F_1(a'+\delta,b',c',d'+\delta;z,-x)}{F_1(a',b',c',d';z,-x)} \tag{30}$$

is monotone increasing when $c' < 0$ and monotone decreasing when $c' > 0$ for $x > -1$.

Proof. Denote by $A_\delta = \frac{1}{\mathscr{B}(d'-a',a'+\delta)}$. Then

$$f_\delta(a',b',c',d';z,x) = \frac{A_\delta \int_0^1 \theta^{a'+\delta-1}(1+\theta x)^{-c'} g(a',d',b',z;\theta)\mathrm{d}\,\theta}{A_0 \int_0^1 \theta^{a'-1}(1+\theta x)^{-c'} g(a',d',b',z;\theta)\mathrm{d}\,\theta}, \tag{31}$$

where $g(a',d',b',z;\theta) = (1-\theta)^{d'-a'-1}(1-\theta z)^{-b'}$. By differentiating ratio (31) with respect to x, the function $f_\delta(a',b',c',d';z,x)$ is monotone increasing for $c' < 0$ (and monotone decreasing for $c' > 0$) if

$$\int_0^1 q(\theta)p(\theta)\mathrm{d}\,\theta \int_0^1 h(\theta)p(\theta)\mathrm{d}\,\theta < \int_0^1 q(\theta)h(\theta)p(\theta)\mathrm{d}\,\theta \int_0^1 p(\theta)\mathrm{d}\,\theta, \tag{32}$$

where

$$p(\theta) = \theta^{a'-1}(1+\theta x)^{-c'} g(a',d',b',z;\theta)$$
$$q(\theta) = \theta^\delta, \quad h(\theta) = \frac{\theta}{1+\theta x}$$

The inequality (32) follows from Chebyshev's integral inequality, because the function $p(\theta)$ is positive and both $q(\theta)$ and $h(\theta)$ are monotone increasing on $(0,\theta)$ for $x > -1$.

Remark 1. Lemma 1 is a generalization of Lemma 1.1 of Karp and Sitnik (2009), who showed the monotonicity of ratios of Gauss Hypergeometric functions (a special case of the Appell F1 functions) with shifted hyperparameters. Their result is obtained as a special case when either $b' = 0$ or $c' = 0$. Lemma 1 can also be formulated in terms of z, where the ratio will be monotone increasing in z when $b' < 0$ and decreasing when $b' > 0$.

The next lemma will be a stepping stone for deriving the upper bound on the selection threshold.

Lemma 2. *Assume $\delta > 0$ and let $f_\delta(a',b',c',d';z,x)$ be as in* (30)*. Assume $c' < 0$ and $0 < z < 1$. Then we have*

$$\lim_{x\to\infty} f_\delta(a',b',c',d';z,x) < \frac{\mathscr{B}(d'-a',a')}{\mathscr{B}(d'-a',a'+\delta)} \frac{\mathscr{B}(d'-a',a'+\delta-c')}{\mathscr{B}(d'-a',a'-c')}. \tag{33}$$

Proof. Let A_δ be as in the proof of Lemma 1. Repeatedly applying l'Hospital's rule (with respect to x), we obtain for $c' < 0$

$$\lim_{x\to\infty} f_\delta(a',b',c',d';z,x) = \frac{A_\delta}{A_0} \frac{\int_0^1 \theta^{a'+\delta-c'-1}(1-\theta)^{d'-a'-1}[1-\theta z]^{-b'}\,\mathrm{d}\,\theta}{\int_0^1 \theta^{a'-c'-1}(1-\theta)^{d'-a'-1}[1-\theta z]^{-b'}\,\mathrm{d}\,\theta} \tag{34}$$

$$= \frac{A_\delta}{A_0} \frac{\mathscr{B}(d'-a',a'+\delta-c')}{\mathscr{B}(d'-a',a'-c')} \frac{F_1(a'+\delta-c',b',0,d'+\delta;z,1)}{F_1(a'-c',b',0,d';z,1)} \tag{35}$$

Note that $F_1(a',b',0,d';z,1) = F_2^1(b',a',d';z)$, where F_2 is the Gauss hypergeometric function. We can apply Lemma 1.1 of Karp and Sitnik (2009) or Lemma 1 to conclude that the ratio of two Gauss functions with shifted arguments is monotone decreasing in z. Since $0 < z < 1$, we have $\frac{F_1(a'+\delta-c',b',0,d'+\delta;z,1)}{F_1(a'-c',b',0,d';z,1)} < \frac{F_1(a'+\delta-c',b',0,d'+\delta;0,1)}{F_1(a'-c',b',0,d';0,1)} = 1$.

Having developed the apparatus of Appell F1 functions, we are ready to state and prove the following key lemma.

Lemma 3. *Assume $\pi(\theta \mid \widehat{\beta})$ is distributed according to* (27)*. Let $\widehat{p} = \|\widehat{\beta}\|_0$. Then*

$$C\,\frac{\widehat{p}+a}{b+a+n} < \mathsf{E}\,[\theta \mid \widehat{\beta}] < \frac{\widehat{p}+a}{b+a+\widehat{p}},$$

where $0 < C < 1$. Moreover, when $a = 1, b = n$ and $(\lambda_0 - \lambda_1)^2\, n/\widehat{p}^2 \to \infty$, then $\lim_{n\to\infty} C = 1$.

Proof. Let $x_j = \left(1 - \frac{\lambda_1}{\lambda_0} \mathrm{e}^{|\widehat{\beta}_j|(\lambda_0 - \lambda_1)}\right)$ and assume (without loss of generality) that $x_j = x$ for $1 \leq j \leq \widehat{p}$. Note that $x < 1$ and $x \to -\infty$ as $\lambda_0 \to \infty$. Applying Lemma 1 and Lemma 2 (with $\delta = 1$) , we obtain

$$\mathsf{E}[\theta \,|\, \widehat{\beta}] = \frac{\mathscr{B}(a+1,b)}{\mathscr{B}(a,b)} \frac{F_1(a+1, \widehat{p}-n, -\widehat{p}, a+b+1; z, x)}{F_1(a, \widehat{p}-n, -\widehat{p}, a+b; z, x)} \tag{36}$$

$$< \frac{\mathscr{B}(\widehat{p}+a+1,b)}{\mathscr{B}(\widehat{p}+a,b)} = \frac{\widehat{p}+a}{\widehat{p}+a+b}. \tag{37}$$

Next, we can write

$$\mathsf{E}[\theta \,|\, \widehat{\beta}] > \frac{\int_0^1 \theta^{\widehat{p}+a}(1-\theta)^b(1-\theta z)^{n-\widehat{p}}\mathrm{d}\,\theta}{\int_0^1 \theta^{\widehat{p}+a-1}(1-\theta)^b(1-\theta z)^{n-\widehat{p}}\mathrm{d}\,\theta} \times \frac{\int_0^1 \theta^{\widehat{p}+a-1}(1-\theta)^b(1-\theta z)^{n-\widehat{p}}\mathrm{d}\,\theta}{\int_0^1 \theta^{\widehat{p}+a-1}(1-\theta)^b(1-\theta z)^{n-\widehat{p}}(1+\frac{1-\theta}{\theta}\frac{\lambda_0}{\lambda_1}\mathrm{e}^{-|\beta_i|(\lambda_0-\lambda_1)})^{\widehat{p}}\mathrm{d}\,\theta},$$

where we used the fact that $1/p^\star_\theta(\beta_i) = 1 + \frac{1-\theta}{\theta}\frac{\lambda_0}{\lambda_1}\mathrm{e}^{-|\beta_i|(\lambda_0-\lambda_1)]} > 1$. Denote by $R(z)$ the first term in the product above and by C the second term. Then

$$R(z) = \frac{\mathscr{B}(\widehat{p}+a+1,b)}{\mathscr{B}(\widehat{p}+a,b)} \frac{F_1(\widehat{p}+a+1, 0, \widehat{p}-n, \widehat{p}+a+b+1; 0, z)}{F_1(\widehat{p}+a, 0, \widehat{p}-n, \widehat{p}+a+b; 0, z)} > R(1) = \frac{\widehat{p}+a}{b+a+n}, \tag{38}$$

where we applied Lemma 1.

Next, we show that $C \to 1$ as $\lambda_0 \to \infty$. First, we denote by $\widetilde{\pi}(\theta) \propto \theta^{a+\widehat{p}-1}(1-\theta)^{b-1}(1-\theta z)^{n-\widehat{p}}(1+\frac{1-\theta}{\theta}\frac{\lambda_0}{\lambda_1}\mathrm{e}^{-|\beta_i|(\lambda_0-\lambda_1)})^{\widehat{p}}$, the density of a generalized Gauss hypergeometric distribution, and by $\mathsf{E}_{\widetilde{\pi}}[\cdot]$ its expectation operator. Then

$$C = \mathsf{E}_{\widetilde{\pi}}\left\{\left[1 + \frac{1-\theta}{\theta}\frac{\lambda_0}{\lambda_1}\mathrm{e}^{-|\widehat{\beta}_i|(\lambda_0-\lambda_1)}\right]^{-\widehat{p}}\right\}.$$

We now use the fact that nonzero values $\widehat{\beta}_i$ are larger than a certain threshold, $p^\star_{\theta_i}(\widehat{\beta}_i) > 0.5(1+\sqrt{1-4/(\lambda_0-\lambda_1)^2}$, (Ročková, 2015). This can be equivalently written as

$$\frac{\lambda_0}{\lambda_1}\mathrm{e}^{-|\widehat{\beta}_i|(\lambda_0-\lambda_1)} < \frac{\mathsf{E}[\theta \,|\, \widehat{\beta}_{\setminus i}]}{1-\mathsf{E}[\theta \,|\, \widehat{\beta}_{\setminus i}]} \frac{1}{(\lambda_0-\lambda_1)^2/2-1} < \frac{\widehat{p}+a}{b-1}\frac{1}{(\lambda_0-\lambda_1)^2/2-1}.$$

In the second inequality above, we applied the upper bound (36). Next,

$$1 > \left[1 + \frac{1-\theta}{\theta}\frac{\lambda_0}{\lambda_1}\mathrm{e}^{-|\widehat{\beta}_i|(\lambda_0-\lambda_1)}\right]^{-\widehat{p}} > \mathrm{e}^{-\frac{1-\theta}{\theta}\frac{\widehat{p}+a}{b-1}\frac{\widehat{p}}{(\lambda_0-\lambda_1)^2/2-1}} \equiv g(\theta)$$

Assuming $(\lambda_0-\lambda_1)^2 n/\widehat{p}^2 \to \infty$ as $n \to \infty$ and $b=n, a=1$, we have $\lim_{n\to\infty} g(\theta) = 1 \forall \theta \in (0,1)$ and $\lim_{n\to\infty} C = 1$, by the bounded convergence theorem.

The arguments apply also when $x_i \neq x_j, 1 \leq i,j \leq \widehat{p}_n$, yielding ultimately the same upper/lower bounds.

Lemma 3 has important implications for the tuning of a and b. With $a=1$ and $b=n$, we obtain $\mathsf{E}[\theta \,|\, \widehat{\beta}] \sim \frac{\widehat{p}}{n}$, which is the actual proportion of nonzero coefficients in $\widehat{\beta}$. Using Lemma 3, we obtain that with $a=1$ and $b=n$, the posterior odds satisfy

$$\frac{n+\widehat{p}+1}{\widehat{p}+1} - 1 < \frac{1-\mathsf{E}[\theta \,|\, \widehat{\beta}]}{\mathsf{E}[\theta \,|\, \widehat{\beta}]} < \frac{2n+1}{\widehat{p}+1} - 1. \tag{39}$$

These posterior odds play a key role in determining the selection thresholds Δ_i, which in turn drive the risk of the global mode estimator.

6 Risk Properties of the Global Mode

Now consider the fully-Bayes SSL estimator $\widehat{\beta}^{FSL}$, and again assume that it is the first p entries in β_0 that are nonzero. Adapting the proof of Theorem 5.1 of Ročková (2015), we obtain

$$\mathsf{E}_{\beta_0} ||\widehat{\beta}^{FSL} - \beta_0||^2 \preceq \sum_{i=1}^{p_n} \mathsf{E}_{\beta_0} {\Delta_i}^2 + \sum_{i=p_n+1}^{n} \mathsf{E}_{\beta_0} \Delta_i \phi(\Delta_i). \tag{40}$$

With $a=1$ and $b=n$, using (39) we obtain

$$\mathsf{E}_{\beta_0} ||\widehat{\beta}^{FSL} - \beta_0||^2 \preceq p_n \mathsf{E}_{\beta_0} \log\left[1 + \frac{\lambda_0}{\lambda_1}\frac{n}{\widehat{p}+1}\right]. \tag{41}$$

It is worthwhile to compare (41) with the upper risk bound (17) obtained for the non-adaptive estimator. Here, we have a different selection threshold for each coordinate and deploy an expected value of these thresholds under $\pi(Y \,|\, \beta_0)$. In the absence of knowledge of p_n, the automatic choice $\lambda_0/\lambda_1 = n^\alpha$, for $\alpha > 0$, and $\theta = 1/n$ yielded $\Delta \sim \sqrt{2\log(1+n^{\alpha+1})}$ in the non-adaptive case. Here, by adapting the parameter θ, we obtain an improvement, where $\Delta_i \sim \sqrt{2\log(1+n^{\alpha+1}/\widehat{p})}$. In either case, with $\lambda_0/\lambda_1 = n^\alpha$ one achieves the near-minimax risk rate $\sqrt{2\log n}$. However, in the adaptive case we have obtained a sharper upper bound.

Whereas the fully Bayes LASSO selection threshold $\mathsf{E}[\lambda \,|\, \widehat{\beta}^L]$ could not be scaled as $\sqrt{2\log n}$, here the adaptive thresholds Δ_i are themselves logarithms and scale as $\sqrt{2\log n}$ under a suitable beta prior $\mathscr{B}(1,n)$. Thus, with the spike-and-slab LASSO, there is no longer a disconnect between the fully Bayes and universal hyperparameter tuning.

6.1 Adapting to the Dimensionality

The purpose of this section is to demonstrate the ability of the "posterior odds" $[1-\mathsf{E}(\theta\,|\,\widehat{\beta}_{\setminus i})]/\mathsf{E}(\theta\,|\,\widehat{\beta}_{\setminus i})$ to adapt to the true unknown sparsity level p_n. For $\theta\sim\mathscr{B}(1,n)$, the asymptotic rate of the odds ratio (39) is governed by $n/\widehat{p}$. Our goal in this section is to show that these odds are of the optimal order n/p_n with large probability. The following lemma will be instrumental in the result to follow.

Lemma 4. *We have*

$$\log\left[\frac{\pi(\beta_0)}{\pi(\widehat{\beta})}\right] > -\lambda_1|\widehat{\beta}-\beta_0| + (\widehat{p}-p_n)\log\left[\frac{\lambda_0}{\lambda_1}\frac{b+k+a}{\widehat{p}+a}\right] + \log C,$$

where $k=\widehat{p}\,\mathrm{I}(\widehat{p}<p_n)+n\,\mathrm{I}(\widehat{p}>p_n)$ *and* C *was defined in the proof of Lemma 3.*

Proof. We can write

$$\log\left[\frac{\pi(\beta_0)}{\pi(\widehat{\beta})}\right] > -\lambda_1|\widehat{\beta}-\beta_0| + (\widehat{p}-p_n)\log\left(\frac{\lambda_0}{\lambda_1}\right) + \log\left[\frac{N(z)}{D(z)}\right] + \log C \tag{42}$$

where $z=(1-\frac{\lambda_1}{\lambda_0})$ and

$$N(z)\equiv\int_0^1\theta^{p_n+a-1}(1-\theta)^{b-1}(1-\theta z)^{n-p_n}\mathrm{d}\,\theta, \tag{43}$$

$$D(z)\equiv\int_0^1\theta^{\widehat{p}+a-1}(1-\theta)^{b-1}(1-\theta z)^{n-\widehat{p}}\mathrm{d}\,\theta. \tag{44}$$

Denote by $R(z)=\frac{\mathscr{B}(\widehat{p}_n+a,b)}{\mathscr{B}(p_n+a,b)}\frac{N(z)}{D(z)}$. First, assume $p_n=\widehat{p}+\delta$ for some $\delta>0$. We can write

$$R(z)=\frac{F_1(\widehat{p}+a+\delta,0,\widehat{p}-n+\delta,b+\widehat{p}+a+\delta;0,z)}{F_1(\widehat{p}+a,0,\widehat{p}-n,b+\widehat{p}+a;0,z)}. \tag{45}$$

As in Lemma 1, we can show (using similar arguments) that $R(z)$ is monotone decreasing in z and thus can be lower-bounded by $R(1)$. Therefore

$$\frac{N(z)}{D(z)}>\frac{\mathscr{B}(p_n+a,b+n-p_n)}{\mathscr{B}(\widehat{p}+a,b+n-\widehat{p})}>\left(\frac{b+n+a}{\widehat{p}+a}\right)^{\widehat{p}-p_n}. \tag{46}$$

Now assume $\widehat{p}=p_n+\delta$ for some $\delta>0$. Using again the monotonicity argument, we find that $1/R(z)$ can be upper-bounded by $1/R(0)$. This yields

$$\frac{N(z)}{D(z)}>\frac{\mathscr{B}(p_n+a,b)}{\mathscr{B}(\widehat{p}+a,b)}>\left(\frac{b+\widehat{p}+a}{\widehat{p}+a}\right)^{\widehat{p}-p_n}.$$

In the following theorem we show that, with high probability, $\widehat{p}$ has the same order as p_n, assuming that the signal is strong enough. Namely, we provide a non-

asymptotic upper and lower bound for $\widehat{p}$, focusing on a set $\tau_0 = \left\{Y : ||Y-\beta_0||_\infty \leq \bar{\lambda}\right\}$, where $\bar{\lambda} = 2\sqrt{\log n}$. The complement of this set has a small probability, i.e. $\mathsf{P}(\tau_0^c) \leq \frac{2}{n}$ (Castillo et al. (2014), Lemma 2).

Theorem 1. *Assume $\lambda_0/\lambda_1 = n^\alpha$ where $\alpha > 0$ and $\lambda_1 < \bar{\lambda}$. Assume $|\beta_{0i}| \geq b_0$, when $\beta_{0i} \neq 0$, where $b_0 > C_1\sqrt{p_n \log n}$ for some $C_1 > 0$. Then with probability at least $1-\frac{2}{n}$, we have*

$$p_n \leq \widehat{p} \leq 2\,p_n + 1. \tag{47}$$

Proof. Denote by $Q(\beta) = -\frac{1}{2}||Y-\beta||^2 + \log\pi(\beta)$, where $\pi(\beta)$ is the non-separable prior (19). Using the global optimality $0 \geq Q(\beta^0) - Q(\widehat{\beta})$, we can write

$$0 \geq ||\widehat{\beta}-\beta^0||^2 - 2\varepsilon'(\widehat{\beta}-\beta^0) + 2\log\left(\frac{\pi(\beta_0)}{\pi(\widehat{\beta})}\right) \tag{48}$$

Now, we condition on the set τ_0 and use the Hölder inequality $|\alpha'\beta| \leq |\alpha|_\infty|\beta|$ to find that

$$0 \geq ||\widehat{\beta}-\beta^0||^2 - 2\,\bar{\lambda}\,|\widehat{\beta}-\beta_0| + 2\log\left(\frac{\pi(\beta_0)}{\pi(\widehat{\beta})}\right). \tag{49}$$

Denote by $\delta = \widehat{\beta}-\beta_0$. Using the fact $|\delta| \leq ||\delta||\,||\delta||_0^{1/2}$, we have

$$0 \geq ||\delta||^2 - 2\,\bar{\lambda}\,||\delta||\,||\delta||_0^{1/2} + 2\log\left(\frac{\pi(\beta_0)}{\pi(\widehat{\beta})}\right). \tag{50}$$

We will first show the upper bound $\widehat{p} \leq C_2\,p_n$ for some $C_2 > 1$. To this end, we assume $\widehat{p} > p_n$. Using the lower-bound of the log-prior ratio in Lemma 4 we have

$$\log\left[\frac{\pi(\beta_0)}{\pi(\widehat{\beta})}\right] \geq -\lambda_1|\beta_0-\widehat{\beta}_0| + (\widehat{p}-p_n)\log\left(\frac{\lambda_0}{\lambda_1}\frac{b+n+a}{\widehat{p}+a}\right) + 2\log C.$$

To continue with (50), we can write

$$\left[||\delta|| - (\bar{\lambda}+\lambda_1)||\delta||_0^{1/2}\right]^2 + 2\,(\widehat{p}-p_n)\log\left(\frac{\lambda_0}{\lambda_1}\frac{b+n+a}{\widehat{p}+a}\right) \leq (\bar{\lambda}+\lambda_1)^2||\delta||_0 + 2\log 1/C. \tag{51}$$

With (51) and using the fact $||\delta||_0 \leq \widehat{p}+p_n$, we obtain

$$2\,(\widehat{p}-p_n)\log\left(\frac{\lambda_0}{\lambda_1}\frac{b+n+a}{\widehat{p}+a}\right) \leq (\bar{\lambda}+\lambda_1)^2(\widehat{p}+p_n) + 2\log 1/C$$

which is equivalent to writing

$$\widehat{p} \leq p_n \left(1 + \frac{2(\bar{\lambda}+\lambda_1)^2}{2\log\left(\frac{\lambda_0}{\lambda_1}\frac{b+n+a}{\widehat{p}_n+a}\right) - (\bar{\lambda}+\lambda_1)^2} \right) + 2\log 1/C.$$

With $\lambda_1 < \bar{\lambda}$ and $\lambda_0/\lambda_1 = n^\alpha$ for sufficiently large $\alpha > 0$, we have $\log\left(\frac{\lambda_0}{\lambda_1}\frac{b+n+a}{\widehat{p}+a}\right) > (\bar{\lambda}+\lambda_1)^2$. Because $C \to 1$ under given assumptions, $\log 1/C < 1/2$ for n large enough. We obtain the upper bound in (47) with $C_2 = 2$.

What remains to be shown is $p_n \leq \widehat{p}$. We prove this statement by contradiction. Assume $\widehat{p} < p_n$ and let $0 < q = p_n - \widehat{p}$. To continue with (50), we use Lemma 4 to obtain

$$0 \geq ||\delta|| \left[||\delta|| - 2(\bar{\lambda}+\lambda_1)||\delta||_0^{1/2} \right] + 2(\widehat{p}-p_n)\log\left(\frac{\lambda_0}{\lambda_1}\frac{b+\widehat{p}+a}{\widehat{p}+a}\right) + 2\log C.$$

Because $||\delta|| < p_n + \widehat{p} < (1+C_2)p_n + 1$, this writes as

$$0 \geq ||\delta|| \left[||\delta|| - 2(\bar{\lambda}+\lambda_1)\sqrt{p_n(1+C_2)+1} \right] + 2(\widehat{p}-p_n)\log\left(\frac{\lambda_0}{\lambda_1}\frac{b+\widehat{p}+a}{\widehat{p}+a}\right) + 2\log C. \tag{52}$$

Assuming the minimal-strength condition $|\beta_{0i}| > b_0$ when $\beta_{0i} \neq 0$, we can write

$$\frac{1}{2}||\delta|| > \frac{1}{2}\sqrt{q}b_0 > C_1\sqrt{p_n \log n} > 2(\bar{\lambda}+\lambda_1)\sqrt{p_n(1+C_1)+1}$$

for suitably large C_1. Assuming $\lambda_0/\lambda_1 = n^\alpha$, (52) yields

$$\begin{aligned} 0 &\geq \frac{1}{2}||\delta||^2 - 2q\log\left(\frac{\lambda_0}{\lambda_1}\frac{b+\widehat{p}+a}{\widehat{p}+a}\right) + 2\log C \\ &> 2C_1^2\sqrt{p_n\log n} - 2q\log\left(\frac{\lambda_0}{\lambda_1}\frac{b+\widehat{p}+a}{\widehat{p}+a}\right) + 2\log C > 0 \end{aligned}$$

for C_1 sufficiently large.

Remark 2. The minimal strength condition in Lemma 1 is a bit stronger than typical beta-min conditions in the LASSO literature. This stronger condition was used previously by Fan and Lv (2014) and Zheng et al. (2014) to show sign consistency of non-concave regularizers.

We conclude the paper with the following result which follows directly from Lemma 3 and Theorem 1.

Corollary 1. *Assume $\theta \sim \mathscr{B}(1,n)$. Under the same conditions as in Theorem 1, we obtain*

$$\frac{1 - \mathsf{E}[\theta \,|\, \widehat{\beta}_{\setminus i}]}{\mathsf{E}[\theta \,|\, \widehat{\beta}_{\setminus i}]} \sim n/p_n$$

and thus $(\Delta_i - \lambda_1)^2 \sim 2\log\left(\frac{\lambda_0}{\lambda_1}\frac{n}{p_n}\right)$, with probability at least $1 - \frac{2}{n}$.

Corollary 1 conveys the very important conclusion that the portion of the selection threshold Δ_i involving θ is *self-adaptive*. In other words, the non-separable penalty obtained with the prior $\pi(\theta) = \mathscr{B}(1,n)$ removes the need for setting θ equal to the true proportion of true coefficients p_n/n, because it can adapt to the ambient dimensionality of the data. Thus, the fully Bayes treatment of θ here mimics oracle performance. This behavior was confirmed by simulations, where self-tuning θ with the fully Bayes formulation was tantamount to selecting θ by cross-validation.

Adapting θ is only halfway towards a fully automatic procedure that would adapt λ_0 and θ simultaneously, removing the need for assuming p_n is known. Observing that $(1-\theta)/\theta$ and λ_0 should be of the same order n/p_n, Ročková (2015) proposed tying λ_0 and θ through $\lambda_0 \sim 1/\theta$ to borrow strength. This amounts to inducing a beta prime distribution on the spike penalty. Another potentially useful approach would be to treat θ and λ_0 independently by assigning two prior distributions. In any case, adapting λ_0 simultaneously with θ requires several nontrivial modifications of our approach and will be reported elsewhere.

7 Discussion

In this paper we demonstrated the potential of the fully Bayes approach for adaptive penalty creation. We compared two deployments of this strategy in terms of their ability to adapt to unknown sparsity: the fully Bayes LASSO and the fully Bayes Spike-and-Slab LASSO. In the first example, the fully Bayes adaptation could not overcome the restrictive form of the penalty. In the second example, however, the fully Bayes adaptation automatically performed universal hyperparameter tuning. For the Spike-and-Slab LASSO, treating a penalty hyperparameter as random with a prior was shown to be tantamount to an oracle choice of the hyperparameter. Thus, penalty functions arising from such fully Bayes prior constructions exert self-adapting ability. This adaptability is reminiscent of an empirical Bayes strategy, and constitutes an alternative to cross-validation and other calibration approaches.

Acknowledgements This work was supported by NSF grant DMS-1406563 and AHRQ grant R21-HS021854.

References

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

Bondell, H. and Reich, B. (2008), "Simultaneous regression shrinkage, variable selection, and supervised clustering of predictors with OSCAR," *Biometrics*, 64, 115–123.

Brown, L. (1971), "Admissible estimators, recurrent diffusions, and insoluble boundary value problems," *Annals of Mathematical Statistics*, 42, 855–903.

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.

Donoho, D., Johnstone, I. M., Hoch, J. C., and Stern, A. S. (1992), "Maximum entropy and the nearly black object," *Journal of the Royal Statistical Society. Series B*, 54, 41–81.

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, 67–70.

Friedman, J. (2008), *Fast sparse regression and classification*, Technical report, Department of Statistics, Stanford University.

George, E. I. (1968a), "Combining minimax shrinkage estimators," *Journal of American Statistical Association*, 81, 437–445.

George, E. I. (1968b), "Minimax multiple shrinkage estimation," *Annals of Statistics*, 14, 188–205.

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

Griffin, J. E. and Brown, P. J. (2012), "Bayesian hyper-LASSOS with non-convex penalization," *Australian & New Zealand Journal of Statistics*, 53, 423–442.

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.

Karp, D. and Sitnik, S. M. (2009), "Inequalities and monotonicity of ratios for generalized hypergeometric function," *Journal of Approximation Theory*, 161, 337–352.

Meier, L., Van de Geer, S., and Bühlmann, P. (2008), "The group LASSO for logistic regression," *Journal of the Royal Statistical Society. Series B*, 70, 53–71.

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

Polson, N. and Scott, J. (2010), "Shrink globally, act locally: Sparse Bayesian regularization and prediction," *Bayesian Statistics*, 9, 501–539.

Ročková, V. (2015), "Bayesian estimation of sparse signals with a continuous spike-and-slab prior," *in revision*.

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

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

Ročková, V. and George, E. (2015b), "The Spike-and-Slab LASSO," *Submitted*.

Stein, C. (1974), "Estimation of the mean of a multivariate normal distribution," in "Prague Symposium on Asymptotic Statistics, Ed. J. Hajek: Univerzita Karlova," .

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

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., and Knight, K. (2005), "Sparsity and smoothness via the fused LASSO," *Journal of the Royal Statistical Society. Series B*, 67, 91–108.

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.

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

Zheng, Z., Fan, Y., and Lv, J. (2014), "High dimensional thresholded regression and shrinkage effect," *Journal of the Royal Statistical Society. Series B*, 76, 627–649.

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

Zou, H. and Hastie, T. (2005), "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society. Series B*, 67, 301–320.