Electronic Journal of Statistics
Vol. 7 (2013) 973–990
ISSN: 1935-7524
DOI: 10.1214/13-EJS795

Hierarchical Bayes, maximum *a posteriori* estimators, and minimax concave penalized likelihood estimation

Robert L. Strawderman

Department of Biostatistics and Computational Biology
University of Rochester, Rochester NY 14642 USA
e-mail: robert_strawderman@urmc.rochester.edu

Martin T. Wells

Department of Statistical Science
Cornell University, Ithaca NY 14853 USA
e-mail: mtw1@cornell.edu

and

Elizabeth D. Schifano

Department of Statistics
University of Connecticut, Storrs CT 06269 USA
e-mail: elizabeth.schifano@uconn.edu

Abstract: Priors constructed from scale mixtures of normal distributions have long played an important role in decision theory and shrinkage estimation. This paper demonstrates equivalence between the maximum aposteriori estimator constructed under one such prior and Zhang's minimax concave penalization estimator. This equivalence and related multivariate generalizations stem directly from an intriguing representation of the minimax concave penalty function as the Moreau envelope of a simple convex function. Maximum aposteriori estimation under the corresponding marginal prior distribution, a generalization of the quasi-Cauchy distribution proposed by Johnstone and Silverman, leads to thresholding estimators having excellent frequentist risk properties.

AMS 2000 subject classifications: Primary 62C60, 62J07.
Keywords and phrases: Convex optimization, Lasso penalty, Moreau regularization, minimax concave penalty, sparsity, smoothly clipped absolute deviation penalty, thresholding.

Received August 2012.

Contents

1. Introduction

Most penalized likelihood estimators have a formal Bayesian interpretation. In particular, the estimates obtained from maximizing a penalized likelihood function may be viewed as the posterior mode, or maximum aposteriori (MAP) estimator, under a (possibly improper) prior distribution implied by the choice of penalty. For example, in the case of the LASSO with design matrix given by the identity, the minimization with respect to $\boldsymbol{\theta}$ of $\frac{1}{2}\|\mathbf{Z}-\boldsymbol{\theta}\|^2+\lambda||\boldsymbol{\theta}||_1$, for $||\boldsymbol{\theta}||_1=\sum_{i=1}^p|\theta_i|$ and $||\boldsymbol{\theta}||=(\boldsymbol{\theta}'\boldsymbol{\theta})^{1/2}$, is observed to be equivalent to computing the MAP estimator of $\boldsymbol{\theta}$ under the model specification $\mathbf{Z}\sim \mathrm{N}_p(\boldsymbol{\theta},\boldsymbol{I}_p)$, where $\boldsymbol{\theta}$ has a prior distribution satisfying $\theta_i \overset{iid}{\sim}$ DoubExp(λ), $i=1\ldots p$ for a constant $\lambda>0$. The solution to this optimization problem is known to be $\widehat{\theta}_i(\mathbf{Z})=\mathrm{sign}(Z_i)(|Z_i|-\lambda)_+$, $i=1\ldots p$ [36]. The hyperparameter λ, held fixed for the purposes of estimating $\boldsymbol{\theta}$, is usually estimated in some adhoc manner (e.g., cross validation), resulting in an estimator with an empirical Bayes flavor.

The double exponential prior implicit in the LASSO penalization has broader connections to estimation under hierarchical prior specifications involving scale mixtures of normal distributions. For example, the double exponential distribution is an obvious special case of $\pi_p(\boldsymbol{\theta}|\lambda)\propto\lambda^p\exp\{-\lambda\|\boldsymbol{\theta}\|\}$, $p\geq 1$, a class of densities that is itself a special case of a very general class of normal scale mixtures known as the multivariate exponential power distributions [cf. 19, Thm. 2.1]. Treating λ as a fixed hyperparameter and considering the corresponding generalization of the LASSO penalization, computation of the resulting MAP estimator under the likelihood specification $\mathbf{Z}\sim \mathrm{N}_p(\boldsymbol{\theta},\boldsymbol{I}_p)$ reduces to determining the value of $\boldsymbol{\theta}$ that minimizes $\frac{1}{2}\|\mathbf{Z}-\boldsymbol{\theta}\|^2+\lambda\|\boldsymbol{\theta}\|$. The resulting estimator is easily shown to be $\hat{\boldsymbol{\theta}}_{GL}(\mathbf{Z})=\left(1-\lambda\|\mathbf{Z}\|^{-1}\right)_+\mathbf{Z}$, an estimator that (i) coincides with the solution to the canonical version of the grouped LASSO problem involving a single group of parameters [38]; and, (ii) for $p=1$, reduces to the soft-thresholding operator $\widehat{\theta}(Z)=\mathrm{sign}(Z)(|Z|-\lambda)_+$.

In a very interesting paper, Takada [35] proves that the positive part James-Stein estimator $\hat{\boldsymbol{\theta}}_{JS+}(\mathbf{Z})=\left(1-(p-2)\|\mathbf{Z}\|^{-2}\right)_+\mathbf{Z}$ is the MAP estimator of $\boldsymbol{\theta}$ under the hierarchical model specification $\mathbf{Z}\sim N_p(\boldsymbol{\theta},\mathbf{I}_p)$ and $\pi(\boldsymbol{\theta},\kappa)=\pi(\boldsymbol{\theta}|\kappa)\pi_T(\kappa)$, where $\boldsymbol{\theta}|\kappa\sim \mathrm{N}_p(\mathbf{0},(\kappa^{-1}-1)\boldsymbol{I}_p)$, and $\pi_T(\kappa)\propto(1-\kappa)^{p/2}\kappa^{-1}1_{[0<\kappa<1]}$.

The marginal prior on $\boldsymbol{\theta}$ is again observed to correspond to a scale mixture of normal distributions. However, in contrast to the LASSO and grouped LASSO estimators above, the minimax estimator $\hat{\boldsymbol{\theta}}_{JS+}(\mathbf{Z})$ is obtained as a MAP estimator when the posterior distribution under the Takada prior is maximized *jointly* in both $\boldsymbol{\theta}$ and κ, not in $\boldsymbol{\theta}$ alone for a fixed value of κ. It is interesting to note that $\hat{\boldsymbol{\theta}}_{GL}(\mathbf{Z})$ provides an empirical Bayes interpretation for $\hat{\boldsymbol{\theta}}_{JS+}(\mathbf{Z})$ upon replacing λ in $\hat{\boldsymbol{\theta}}_{GL}(\mathbf{Z})$ with $\hat{\lambda} = (p-2)/\|\mathbf{Z}\|$; see [34] for further discussion.

Prior distributions constructed from scale mixtures of normal distributions have a rich history in the theory of shrinkage estimation and Bayesian decision theory. Recently, there has been a resurgence of interest in such priors in connection with both frequentist and Bayesian treatments of sparse estimation problems; see, for example, [37, 23, 26, 22, 21, 12, 3, 34], and [27]. Let $\phi(\cdot)$ and $\Phi(\cdot)$ respectively denote the standard normal density and cumulative distribution functions. Let $\mathrm{TN}(a, b)$ denote a normal random variable W having mean a and variance b, but truncated below at zero. Consider the hierarchical class of proper priors

$$\boldsymbol{\theta}|\psi \sim \mathrm{N}_p(\mathbf{0}, \psi \boldsymbol{I}_p),\ \psi|\gamma \sim \mathrm{Gamma}\left(\frac{p+1}{2}, \frac{\gamma^2}{2}\right),\ \gamma|\alpha, \lambda \sim \mathrm{TN}(\lambda, \{2\alpha\}^{-1}) \quad (1.1)$$

where $p \geq 1$, $\alpha > 0$, and $\lambda \geq 0$, the prior density function on γ being given by

$$\pi(\gamma|\alpha, \lambda) = \frac{e^{-\alpha(\gamma-\lambda)^2}\alpha^{1/2}}{\pi^{1/2}\Phi\left\{\lambda(2\alpha)^{1/2}\right\}} 1_{[\gamma \geq 0]}.$$

When $\alpha = 1$, integrating γ out of (1.1) yields a proper version of the prior for $\psi = \kappa^{-1} - 1$ in the Takada prior [34, Sec. 2.2]. When $p = 1$ (i.e., whether or not $\alpha = 1$), the marginal prior for $\theta \in \mathbb{R}$ under (1.1) can be shown to reduce to

$$\pi(\theta|\alpha, \lambda) \quad = \quad \frac{1 - r_{\alpha,\lambda}(|\theta|)\, M\{r_{\alpha,\lambda}(|\theta|)\}}{2^{3/2}\alpha^{1/2}\, M\{r_{\alpha,\lambda}(0)\}} \quad (1.2)$$

where $r_{\alpha,\lambda}(|\theta|) = |\theta|/(2\alpha)^{1/2} - \lambda(2\alpha)^{1/2}$ and $M(s) = (1 - \Phi(s))/\phi(s)$ denotes Mills Ratio. For $\lambda \geq 0$, the prior (1.2) is symmetric about $\theta = 0$, as well as bounded and non-differentiable there. For $\lambda = 0$, (1.2) reduces to the quasi-Cauchy prior of [23] [e.g., see 34, Sec. 2.1] with scale parameter $(2\alpha)^{-1/2}$. Strongly related classes of normal mixture priors include those considered in [33, 20, 12, 3, 21] and [27].

Remarkably, upon setting $p = 1$, the marginal prior (1.2) can also be derived directly from the prior class

$$\pi(\boldsymbol{\theta}|\gamma, \alpha) \propto \gamma^p \exp\{-\gamma\|\boldsymbol{\theta}\|\}, \quad \pi(\gamma|\alpha, \lambda) \propto \alpha^{1/2} \exp\{-\alpha(\gamma - \lambda)^2\}\, 1_{[\gamma \geq 0]} \quad (1.3)$$

a result easily proved by integrating out ψ (i.e., instead of γ) in (1.1). The conditional prior $\pi(\boldsymbol{\theta}|\gamma, \alpha)$ takes the same form as $\pi_p(\boldsymbol{\theta}|\lambda)$; in view of the fact that $\alpha = \infty$ corresponds to placing a point mass at $\gamma = \lambda$, the prior (1.3) contains $\pi_p(\boldsymbol{\theta}|\lambda)$, hence the double exponential density with scale parameter λ,

as special cases. The results of [35], combined with this observation, motivate [34] to propose jointly maximizing the posterior distribution of $(\boldsymbol{\theta}, \gamma)|\mathbf{Z}$ induced by the likelihood specification $\mathbf{Z} \sim N_p(\boldsymbol{\theta}, \mathbf{I}_p)$ and modified prior distribution

$$\pi(\boldsymbol{\theta}|\gamma, \alpha, \lambda) \propto \gamma^p \exp\{-\gamma\|\boldsymbol{\theta}\|\}, \quad \tilde{\pi}(\gamma|\alpha, \lambda) \propto \alpha^{1/2}\gamma^{-p} \exp\{-\alpha(\gamma - \lambda)^2\}\, 1_{[\gamma \geq 0]}, \tag{1.4}$$

where $\alpha, \lambda > 0$ are fixed hyperparameters. Comparing (1.3) and (1.4), the difference lies in replacing the proper prior $\pi(\gamma|\alpha, \lambda)$ with the improper prior $\tilde{\pi}(\gamma|\alpha, \lambda)$. The nature of the modification that leads from (1.3) to (1.4) is analogous to the use of the factor $(1 - \kappa)^{p/2}$ in the Takada prior $\pi_T(\kappa)$. The prior distribution (1.4) is evidently improper; however, the corresponding posterior distribution for $(\boldsymbol{\theta}, \gamma)|\mathbf{Z}$ remains proper and can thus be used as a statistical model for the purposes of estimation and inference. In particular, the MAP estimator for $(\boldsymbol{\theta}, \gamma)$ under (1.4) is obtained by jointly minimizing

$$L(\boldsymbol{\theta}, \gamma) = \frac{1}{2}\|\mathbf{Z} - \boldsymbol{\theta}\|^2 + \gamma\|\boldsymbol{\theta}\| + \alpha(\gamma - \lambda)^2 \tag{1.5}$$

for $\boldsymbol{\theta} \in \mathbb{R}^p$ and $\gamma > 0$. The solution obtained evidently corresponds to solving the LASSO ($p = 1$) or grouped LASSO ($p > 1$) problem with an additional penalty on LASSO penalty parameter γ.

In this article, we focus on the implications of using the priors (1.3) and (1.4) in two MAP estimation problems when $p = 1$: assuming $Z \sim N(\theta, 1)$, derive (i) the MAP estimator for (θ, γ) under the improper prior (1.4) by minimizing (1.5) for $p = 1$ in both θ and γ; and, (ii) the MAP estimator for θ under the proper marginal prior (1.3), or equivalently, (1.2). We show that the optimization problem in (i) results in an estimator for θ that coincides with the univariate version of the MCP threshold estimator [39]. We further show that the MCP penalty function can be characterized as the Moreau envelope of a simple convex function and subsequently generalize these observations to a much broader class of estimation problems. Finally, considering (ii), we establish a new class of thresholding estimators with desirable frequentist properties, as developed in [2] and [16].

2. Main results

2.1. MAP estimation and minimax concave penalization

Given $\theta \in \mathbb{R}$ and $\gamma \geq 0$, suppose Z satisfies the canonical normal model $Z = \theta + \epsilon$, where $\epsilon \sim N(0, 1)$. Let $\alpha = a/2$; then, for $a > 1$ (i.e., $\alpha > 1/2$) and $\lambda > 0$, suppose (θ, γ) has the improper prior distribution $\pi(\theta, \gamma|a/2, \lambda)$ given in (1.4) (i.e., for $p = 1$). Let $z \in \mathbb{R}$ denote the observed value of Z. Then, following (1.5), the MAP estimator for (θ, γ) is formally obtained by minimizing

$$G(\theta, \gamma; z) = \frac{1}{2}(\theta - z)^2 + \gamma|\theta| + \frac{a}{2}(\gamma - \lambda)^2 \tag{2.1}$$

jointly in $\theta \in \mathbb{R}$ and $\gamma \geq 0$, where $a > 1$ and $\lambda \geq 0$ are constants.

As noted above, (2.1) is formally motivated as the posterior distribution of (θ, γ) under the hierarchically specified improper prior (1.4). It is well known that improper priors must be used with care in Bayesian estimation and inference problems, as these can lead to marginalization paradoxes and other problems; for example, see Kass and Wasserman [24] and Robert [28, p. 29]. Such concerns are less relevant in settings where the focus lies on the corresponding posterior distribution, provided this posterior distribution is well defined and generates a useful statistical model [e.g. 28, p. 128]. Indeed, the literature on Bayesian inference is rich with examples of improper priors that lead to useful estimators with good properties [e.g., 6, 7, 17]. The following result shows that the objective function (2.1) is derived from a valid posterior distribution; hence, the MAP for (θ, γ) is well-defined from a generalized Bayes perspective. The proof of this result may be found in Appendix A; the argument used there can be extended in a straightforward manner to similarly justify (1.5) for general p and vectors $\mathbf{Z}$ and $\boldsymbol{\theta}$.

Theorem 2.1. *Let*

$$\pi(\theta, \gamma) \propto (a/2)^{1/2} \exp\{-\gamma|\theta|\} \exp\{-(a/2)(\gamma - \lambda)^2\}$$

denote the improper prior (1.4) *with* $p = 1$ *and* $\alpha = a/2$. *Define*

$$m(z) \quad = \quad \int_{\mathbb{R}} \int_0^{\infty} \phi(z - \theta)\pi(\theta, \gamma)\, d\gamma\, d\theta. \tag{2.2}$$

Then, $m(z) < \infty$ *for each* $z \in \mathbb{R}$, *implying the existence of the posterior distribution*

$$\pi(\theta, \gamma|z) = \frac{\phi(z - \theta)\pi(\theta, \gamma)}{m(z)}.$$

Returning to the problem of minimizing (2.1), results in Strawderman and Wells [34, Thm. 4.2] imply (2.1) is strictly convex for $(\theta, \gamma) \in \mathbb{R} \times \mathbb{R}_+$, with unique solution for θ given by

$$\eta_M(z; \lambda, a) \quad = \quad \begin{cases} \frac{a}{a-1}\text{sign}(z)\,(|z| - \lambda)_+ & \text{if } |z| < a\lambda \\ z & \text{if } |z| \geq a\lambda \end{cases}. \tag{2.3}$$

The estimator (2.3) is equivalent to the univariate MCP threshold estimator, derived as the minimizer of

$$H(\theta; z) = \frac{1}{2}(\theta - z)^2 + \lambda \int_0^{|\theta|} \left(1 - \frac{x}{a\lambda}\right)_+ dx \tag{2.4}$$

in θ only for fixed $\lambda > 0$ and $a > 1$; see Zhang [39, §2.1, §7.3]. As discussed in [39], (2.3) is also equivalent to the firm threshold estimator of [11], reducing to the hard- and soft-thresholding operators as a respectively approaches one and infinity. The thresholding estimator (2.3) exhibits the sparsity, continuity and unbiasedness properties recommended in [16] and, following Antoniadis and Fan [2, §3.3], Zhang [39] and Gao and Bruce [18, §3.1], can be shown to possess various oracle properties under suitable regularity conditions.

Remark 2.2. It can be shown directly that (2.4) corresponds to a profiled version of (2.1) with γ set equal to its minimizer $\hat{\gamma} = (\lambda - \frac{|\theta|}{a})_+$ in Theorem 2.3 and $t = |\theta| \geq 0$; see Schifano [31, Ch. 6]. As a result, in the given form, (2.4) does not correspond to the estimator of θ that would be obtained using the marginal prior "density" of θ derived from (1.4). This fact can also be seen directly from the proof of Theorem 2.1, where an expression for the marginal prior "density" of θ is obtained in (A.2).

The appearance of (2.3) as the solution for θ when jointly minimizing (2.1) is very surprising and suggests a direct connection between the respective optimization problems associated with (2.1) and (2.4). Theorem 2.3 establishes the exact nature of this connection.

Theorem 2.3. *Let $\lambda > 0$ and $a > 0$. Then, for any $t \geq 0$,*

$$\lambda \int_0^t \left(1 - \frac{x}{a\lambda}\right)_+ dx \quad = \quad \min_{\gamma \geq 0} \left\{\gamma t + \frac{a}{2}(\gamma - \lambda)^2\right\}, \tag{2.5}$$

and $\hat{\gamma}_\lambda = \lambda(1 - t/(a\lambda))_+$ is the unique solution to the right hand side of (2.5).

A proof of this result is provided in Appendix B. Remarkably, the equivalence result in Theorem 2.3 can also be established using results from convex analysis. The minimization problem on the right-hand side of (2.5) is equivalent to

$$\min_{\gamma \in \mathbb{R}} \left\{\gamma t + \iota_{[0,\infty)}(\gamma) + \frac{a}{2}(\gamma - \lambda)^2\right\}, \tag{2.6}$$

where $\iota_C(\gamma)$ is zero for $\gamma \in C$ and infinity for $\gamma \notin C$; see, for example, Rockafellar and Wets [29, p. 7]. The calculation in (2.6) yields the Moreau envelope function for the proper convex function $h(\gamma) = \gamma t + \iota_{[0,\infty)}(\gamma)$ [29, p. 4, Def. 1.22, and p. 40], t considered fixed; call the resulting envelope function $e_{1/a}(\lambda)$. The function of interest here is γt; with $t = |\theta|$, γt evidently corresponds to the linear term in the objective function (2.1). The role of $\iota_{[0,\infty)}(\gamma)$ is projection, and its presence ensures that the solution respects the desired constraint, that is, $\gamma \in [0, \infty)$. The envelope function $e_{1/a}(\lambda)$ is convex and continuously differentiable with gradient $\nabla e_{1/a}(\lambda) = a(\lambda - \mathcal{P}_{1/a}(\lambda))$, where $\mathcal{P}_{1/a}(\lambda)$ is the so-called proximal mapping [29, Thm. 2.26]. Importantly, $\mathcal{P}_{1/a}(\lambda)$ is also the solution to the minimization problem (2.6) [29, Def. 1.22], which by Theorem 2.3 equals $\hat{\gamma}_\lambda$. Hence, $\nabla e_{1/a}(\lambda) = a(\lambda - \hat{\gamma}_\lambda)$; the desired equivalence in (2.5) now follows upon integrating $\nabla e_{1/a}(s) = a(s - \hat{\gamma}_s)$ for $s \in [0, \lambda]$ and observing that the resulting integral equals the same expression for $\lambda \int_0^t \left(1 - \frac{x}{a\lambda}\right)_+ dx$ given in (B.1) of Appendix B.

In view of (2.1), the recovery of the LASSO solution (i.e., soft thresholding) at $a = \infty$ is initially surprising, for (2.1) reduces to the LASSO objective function if one sets $a = 0$ and treats γ as a fixed nonnegative constant. However, careful inspection of (2.1) shows that any value of γ other than $\gamma = \lambda$ will result in an infinite objective function as $a \to \infty$; for $\gamma = \lambda$, we evidently recover the LASSO objective function and corresponding solution at $a = \infty$.

The recovery of soft thresholding at $a = \infty$ also has an interesting quasi-Bayesian interpretation. Considering (1.4) for $\lambda > 0$ and recalling that $a = 2\alpha$, the prior distribution on γ (i.e., $\tilde{\pi}(\gamma|\alpha, \lambda)$) evidently becomes increasingly concentrated about $\gamma = \lambda$ as $\alpha \to \infty$. Asserting that this also implies $\pi(\theta, \gamma|\alpha, \lambda) \to \pi(\theta|\gamma = \lambda, \alpha, \lambda)$ as $\alpha \to \infty$, the MAP estimation problem leading to (2.1) again reduces to that for the univariate LASSO problem. It is interesting that the hard thresholding estimator arises as the solution when $a = 1$, the boundary where (2.1) transitions from being jointly convex to non-convex. However, in contrast to the case where $a = \infty$, a similarly intuitive Bayesian interpretation of this result does not appear to be available. Nevertheless, the indicated formulation of MCP shows that a and λ may be reasonably interpreted as hyperparameters, suggesting new methods of tuning parameter selection.

2.2. Thresholding rules derived as MAP estimators under marginal priors

The estimator (2.3) is derived as a MAP estimator through jointly optimizing the negative log-posterior (2.1) in θ and γ. From a frequentist perspective, such a procedure is sensible; one merely transfers the need to estimate the LASSO penalty parameter γ to the need to estimate (or otherwise specify) the hyperparameter λ. However, from a Bayesian point of view, γ is really a nuisance parameter and it is arguably more natural to derive the MAP estimator, as well as other Bayesian estimators (e.g., posterior means and medians), using the marginal prior distribution for θ [e.g., 20, 3, 27]. In the case of (1.1), equivalently (1.3), the relevant proper marginal prior is given by (1.2). It is straightforward to verify that this marginal prior distribution reduces to a double exponential density as $\alpha \to \infty$,; hence, as in the previous section, the MAP estimator under (1.2) also converges to the soft-thresholding estimator as $\alpha \to \infty$. More generally, for $s \geq 0$, define

$$p_{\lambda,\alpha}(s) = -\log\left[1 - r_{\alpha,\lambda}(s)\, M\{r_{\alpha,\lambda}(s)\}\right] + c_{\alpha,\lambda}, \tag{2.7}$$

where $\alpha, \lambda > 0$ are constants and $c_{\alpha,\lambda} = \log\left[1 + \lambda(2\alpha)^{1/2} M\{-\lambda(2\alpha)^{1/2}\}\right]$ ensures that $p_{\lambda,\alpha}(s) \geq 0$ whenever $s \geq 0$. Then, under the normal model of Section 2.1, and with $\theta|\alpha, \lambda$ having the prior distribution (1.2), the computation of the MAP estimator for θ is equivalent to minimizing $\frac{1}{2}(\theta - z)^2 + p_{\lambda,\alpha}(|\theta|)$ for $\theta \in \mathbb{R}$, the constant $c_{\alpha,\lambda}$ being irrelevant.

Properties of (2.7), presented with a view towards deriving thresholding rules satisfying the requirements of Antoniadis and Fan [2], are recorded in the theorem below; proof may be found Appendix C.

Theorem 2.4. *For $s \geq 0$, with $\lambda, \alpha > 0$, the penalty (2.7) is nonnegative, increasing, and strictly concave. Moreover, (2.7) is continuously differentiable for $s \in (0, \infty)$, with a decreasing first derivative $p^{(1)}_{\lambda,\alpha}(s)$ that satisfies $p^{(1)}_{\lambda,\alpha}(0+) > 0$ and decays to zero as $O(s^{-1})$ as $s \to \infty$. For $\alpha > 1/2$, $-s - p^{(1)}_{\lambda,\alpha}(s)$ is strictly unimodal, with $|s| + p^{(1)}_{\lambda,\alpha}(|s|)$ minimized at $s = 0$.*

The proof of this result relies heavily on properties of Mills Ratio; see Lemma C.1. The important implication of Theorem 2.4 is that the resulting thresholding rule $\eta_\pi(z;\lambda,a)$, though not available in closed form, satisfies all of the properties in Theorem 1 and Lemma 1 of Antoniadis and Fan [2], hence all of the desirable risk and estimation properties in Theorems 2-6 of that same work; see also [16]. For example, similarly to (2.3), $\eta_\pi(z;\lambda,\alpha)$ is continuous in the data z, thresholds z to zero for $|z| \le p^{(1)}_{\lambda,\alpha}(0)$, and satisfies $\eta_\pi(z;\lambda,\alpha) = z - p^{(1)}_{\lambda,\alpha}(|z|) + o(p^{(1)}_{\lambda,\alpha}(|z|))$ as $|z| \to \infty$, demonstrating shrinkage as well as "near unbiasedness" for large signals. By Lemma C.1(vii), $\pi(\theta|\lambda,\alpha) \propto [2 + \{r_{\alpha,\lambda}(|\theta|)\}^2]^{-1} + O(|\theta|^{-4})$ as $|\theta| \to \infty$, implying Cauchy-like tail behavior as $|\theta| \to \infty$. As result, one can also expect excellent Bayesian robustness properties, including the avoidance of excessive shrinkage away from the origin [e.g. 12, 3]. Posterior median thresholding rules and posterior mean estimators derived under the analogous class of mixture priors can also be expected to share similarly desirable risk properties to those in [23].

Remark 2.5. As noted in Remark 2.2, one can obtain an expression for the marginal prior "density" of θ under the improper prior (1.4); see (A.2). In particular, using (A.2) to construct a penalty leads to the objective function

$$G_m(\theta;z) = \frac{1}{2}(\theta - z)^2 + p^*_{\lambda,a}(|\theta|),$$

where $p^*_{\lambda,a}(s) = -\log M\{r_{a/2,\lambda}(s)\}, s \ge 0$ is both strictly concave and increasing on $\mathbb{R}$, a result of the fact that $M(\cdot)$ is strictly log-convex on $\mathbb{R}$ [4, Thm. 2.8]. In fact, similarly to Theorem 2.4, it can be shown that $p^*_{\lambda,a}(s), s \in \mathbb{R}$ also satisfies the key requirements of [2]. Therefore, as in Theorem 2.4 the resulting thresholding rule is expected to exhibit several desirable theoretical properties.

2.3. *MAP estimation and MCP: Extensions for multivariate problems*

Theorem 2.3 confirms that minimizing (2.4) for $\theta \in \mathbb{R}$ is equivalent to joint minimization of (2.1) for $\theta \in \mathbb{R}$ and $\gamma \ge 0$. Existence and uniqueness, hence equivalence, of the minimizers of (2.1) and (2.4) are ensured by the strict convexity of these objective functions that results from imposing the additional condition that $a > 1$. Under suitable regularity conditions, this result can be generalized in a very substantial way. As a prelude to this result, we note the following corollary to Theorem 2.3.

Corollary 2.6. *For constants $t_i \ge 0$ and $w_i > 0$, $i = 1 \dots L$,*

$$\min_{\boldsymbol{\gamma} \in \mathbb{R}^L_+} \left\{ \sum_{i=1}^L w_i \left(\gamma_i t_i + \frac{a}{2}(\gamma_i - \lambda)^2 \right) \right\} = \lambda \sum_{i=1}^L w_i \int_0^{t_i} \left(1 - \frac{x}{a\lambda}\right)_+ dx, \qquad (2.8)$$

where $\lambda, a > 0$, $\boldsymbol{\gamma} = (\gamma_1 \dots \gamma_L)'$, and $\mathbb{R}^L_+$ denotes the nonnegative orthant of $\mathbb{R}^L$.

Since $\sum_{i=1}^{L} w_i \left(\gamma_i t_i + \frac{a}{2}(\gamma_i - \lambda)^2\right)$ is a separable sum of positive functions, the proof of this result is an easy consequence of Theorem 2.3. The equivalence result in (2.8) is also what is needed to prove the following general result for MCP estimators; see Appendix D for proof.

Theorem 2.7. *Let $\boldsymbol{\theta} \in \Omega \subset \mathbb{R}^p$ and, for $i = 1 \dots L$, write $\boldsymbol{\theta} = (\boldsymbol{\theta}_1', \dots, \boldsymbol{\theta}_L')'$, where $\boldsymbol{\theta}_i$ has dimension $p_i \leq p$ and $\sum_{i=1}^{L} p_i = p$. Suppose $g(\boldsymbol{\theta})$ is a strictly convex, twice continuously differentiable function on Ω, where Ω is a bounded convex set. Let $H(\boldsymbol{\theta})$ denote the Hessian matrix of $g(\boldsymbol{\theta})$ and assume $\lambda_{min} > 0$, where λ_{min} is the minimum eigenvalue of $H(\boldsymbol{\theta})$ over $\boldsymbol{\theta} \in \Omega$. Finally, let $w_i > 0, i = 1 \dots L$ be constants and define $w_{max} = \max_{i=1\dots n} w_i$. Then, for $\lambda > 0$, $a > w_{max}/\lambda_{min} > 0$, it follows that*

$$\min_{\boldsymbol{\theta} \in \Omega} \; g(\boldsymbol{\theta}) + \lambda \sum_{i=1}^{L} w_i \int_0^{\|\boldsymbol{\theta}_i\|} \left(1 - \frac{x}{a\lambda}\right)_+ dx \tag{2.9}$$

is equivalent to

$$\min_{\boldsymbol{\theta} \in \Omega, \boldsymbol{\gamma} \in \mathbb{R}_+^L} \; g(\boldsymbol{\theta}) + \sum_{i=1}^{L} w_i \left\{ \gamma_i \|\boldsymbol{\theta}_i\| + \frac{a}{2}(\gamma_i - \lambda)^2 \right\}. \tag{2.10}$$

The stated equivalence holds in much greater generality, in the sense that the set of global minima for general $g(\boldsymbol{\theta})$ and/or $a > 0$ failing to satisfy the indicated eigenvalue constraint also coincide with each other [e.g., 29, Prop. 1.35].

The class of problems represented by (2.9), hence (2.10), includes many interesting special cases. For example, with $p_i = 1$ for each i, hence $L = p$, taking $g(\boldsymbol{\theta}) = \frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|^2$ for a response vector $\mathbf{y}$ and design matrix $\mathbf{X}$ corresponds to minimax concave penalized estimation for a linear model [39]. Similarly, with $p_i \geq 1$ for each i and $p_j > 1$ for at least one j, we have $L < p$ and hence a "grouped" version of minimax concave penalization; see, for example, Breheny and Huang [9]. Taking $g(\boldsymbol{\theta})$ to be the negative log-likelihood function for a generalized linear model extends this framework to a much wider class of problems. While a geometric-based interpretation and algorithm exists for fitting linear models with the minimax concave penalty [39], the fitting of generalized linear models with the minimax concave penalty relies heavily on iterative optimization algorithms. From an algorithmic point of view, (2.10) provides a direct route for solving such problems iteratively, proceeding by estimating $\boldsymbol{\theta}$ for fixed $\boldsymbol{\gamma}$ and then estimating $\boldsymbol{\gamma}$ for fixed $\boldsymbol{\theta}$, the latter existing in closed form (see Theorem 2.3). In the case where $L = p$, $p_i = 1$ for each i, and $g(\boldsymbol{\theta})$ corresponds to the negative log-likelihood function for a generalized linear regression model, the representation (2.10) directly justifies the use of the local linear approximation algorithm suggested in [41] for minimax concave penalization, and can be fruitfully combined with majorization-minimization algorithms [e.g., 32] and related coordinate-wise optimization methods [e.g., 10]. For example, the MIST algorithm builds directly on the work of Zou and Li [41] using a suitable modification of the majorization-minimization algorithm that facilitates coordinatewise optimization. In the case of a generalized linear model, the majorization function

used in MIST is exactly of the form (2.10); see Schifano, Strawderman and Wells [32] for comparisons to the original LLA algorithm in the case of the minimax concave penalty. Coordinate descent methods for solving (2.9) have been proposed for linear models with the minimax concave penalty [25]; we are not currently aware of published work that extends these results to directly solving (2.9) in the case of generalized linear models. It would also be interesting to study how such an algorithm would perform in comparison to a coordinate descent algorithm specifically designed for (2.10).

The results and discussion above are focused solely on the equivalence of the optimization problems (2.9) and (2.10). In order to interpret (2.10) as a MAP estimation problem in a generalized Bayes context, the function $g(\boldsymbol{\theta})$ should correspond to a suitable negative loglikelihood function for the response $\mathbf{Z}$ and, similarly to Theorem 2.1, we must show that $m(\mathbf{Z}) < \infty$, where

$$m(\mathbf{Z}) = \int_{\boldsymbol{\theta}\in\Omega} e^{-g(\boldsymbol{\theta})} \left[\int_{\boldsymbol{\gamma}\in\mathbb{R}_+^L} \exp\left\{ -\sum_{i=1}^{L} w_i \left\{ \gamma_i \|\boldsymbol{\theta}_i\| + \frac{a}{2}(\gamma_i - \lambda)^2 \right\} \right\} d\boldsymbol{\gamma} \right] d\boldsymbol{\theta}. \quad (2.11)$$

Calculations similar to those used to establish (A.2) may be used to prove that

$$\int_{\boldsymbol{\gamma}\in\mathbb{R}_+^L} \exp\left\{ -\sum_{i=1}^{L} w_i \left\{ \gamma_i \|\boldsymbol{\theta}_i\| + \frac{a}{2}(\gamma_i - \lambda)^2 \right\} \right\} d\boldsymbol{\gamma} \propto \prod_{i=1}^{L} M\{r_{w_i a/2,\lambda}(w_i\|\boldsymbol{\theta}_i\|)\};$$

hence, $m(\mathbf{Z}) < \infty$ and propriety of the resulting posterior distribution follows if

$$\int_{\boldsymbol{\theta}\in\Omega} \exp\{-g(\boldsymbol{\theta})\} \left[\prod_{i=1}^{L} M\{r_{w_i a/2,\lambda}(w_i\|\boldsymbol{\theta}_i\|)\} \right] d\boldsymbol{\theta} < \infty.$$

Because $r_{w_i a/2,\lambda}(w_i\|\boldsymbol{\theta}_i\|) > -\lambda(w_i a)^{1/2}$ for all $\boldsymbol{\theta} \in \Omega$, $0 < M(s) < \infty$ for all $s > -\infty$, and $M(s)$ is strictly decreasing, a sufficient condition for (2.11) to be finite is that

$$\int_{\boldsymbol{\theta}\in\Omega} \exp\{-g(\boldsymbol{\theta})\} d\boldsymbol{\theta} < \infty;$$

that is, $m(\mathbf{Z}) < \infty$ if the posterior under a flat improper prior on $\boldsymbol{\theta}$ also exists. There are several important examples for which such results have already been established. For example, this clearly holds true when the response vector $\mathbf{Z}|\boldsymbol{\theta} \sim N_p(\mathbf{X}\boldsymbol{\theta}, \mathbf{I}_p)$ and the design matrix $\mathbf{X}$ has full column rank; see, for example, Box and Tiao [8, Sec. 2.7.1]. As another example, Chen and Shao [14, Thm. 2.1] prove that the posterior is proper under a flat prior on $\boldsymbol{\theta}$ for general quantal response models, i.e., $P\{Z_i = 1|x_i\} = F(x_i'\boldsymbol{\theta})$ for some cumulative distribution function $F(\cdot)$, provided $F(\cdot)$ has at least $k = \dim(\boldsymbol{\theta})$ moments, the design matrix $\mathbf{X}$ has full column rank, and a certain verifiable constraint involving the binary response vector $\mathbf{Z}$ and $\mathbf{X}$ holds. Under related conditions on the design matrix and sample configuration, Chen, Ibrahim and Shao [13, Thms. 1 & 2] establish propriety of the posterior for Cox's regression model [15] under a flat prior on the regression parameter $\boldsymbol{\theta}$ using either the partial likelihood function or the

full likelihood function, the latter additionally involving a gamma process prior specification on the cumulative hazard function.

Weaker sufficient conditions for posterior propriety can be obtained using properties of Mill's ratio [e.g., 30]. Let $\tilde{\Omega} = \{\boldsymbol{\theta} \in \Omega : \|\boldsymbol{\theta}\|_i > \lambda a,\ i = 1 \dots L\}$. In particular, a sufficient condition for (2.11) to be finite is that

$$\int_{\boldsymbol{\theta} \in \tilde{\Omega}} \exp\{-g(\boldsymbol{\theta})\} \left[\prod_{i=1}^{L} q_i(\|\boldsymbol{\theta}_i\|) \right] d\boldsymbol{\theta} < \infty,$$

where, for $s \geq 0$,

$$q_i(s) = \frac{4}{3r_{w_i a/2,\lambda}(w_i s) + \sqrt{[r_{w_i a/2,\lambda}(w_i s)]^2 + 8}}$$

and $q_i(s) = O(s^{-1})$ for large s.

3. Discussion

In various forms, the class of normal scale mixture priors (1.1) has long played an important role in decision theory and shrinkage estimation. This paper demonstrates that several interesting thresholding estimators with good frequentist properties are connected either directly or indirectly to this class of priors, including the minimax concave penalized estimator of [39] (i.e., derived as a MAP under the improper modification (1.4)), thresholding estimators based on generalizations of the quasi-Cauchy prior of [23] (i.e., MAP estimators derived using the marginal prior (1.2)), and a wide class of multivariate generalizations of these results.

Theorem 2.3 further shows that penalties constructed via infimal convolutions of convex functions have natural connections to MAP estimators derived under the hierarchical prior (1.4). Certainly, not all penalty functions of interest can be represented in this way. Moreover, not all penalty functions that can be represented in this way will necessarily have an attractive Bayesian interpretation. For example, consider for $t = |\theta| \geq 0$ the penalty

$$p_{\lambda,a}(t) = \lambda t\, 1_{[t<\lambda]} + \left(\frac{a\lambda t}{a-1} - \frac{(t^2+\lambda^2)}{2(a-1)} \right) 1_{[\lambda \leq t < a\lambda]} + (a+1)\frac{\lambda^2}{2}\, 1_{[t \geq a\lambda]};$$

as discussed in Fan and Li [16], this "smoothly clipped absolute deviation" penalty [16] is designed to exhibit certain sparsity, continuity, and unbiasedness properties. However, beyond these key features, the motivation underlying $p_{\lambda,a}(\cdot)$ is largely heuristic. Define the function $g_t(\gamma) = c_t\gamma^2 + d_t\gamma + 0.5(a+1)(\gamma - \lambda)^2$ for $d_t = (\lambda + t)\, 1_{[t \geq a\lambda]}$ and

$$c_t = \frac{t(a+1)}{\lambda(a+1) - 2t} 1_{[t<\lambda]} + \left[\frac{(a+1)}{2} \left\{ \frac{\lambda^2(a^2-1)}{(t-a\lambda)^2} - 1 \right\} \right] 1_{[\lambda \leq t < a\lambda]}.$$

Arguments like those used to prove Theorem 2.3 show $\min_{\gamma \geq 0} g_t(\gamma) = p_{\lambda,a}(t)$ for $t \geq 0$. Formulated similarly to (1.4), these results suggest a joint prior

distribution of the form $\pi(\theta, \gamma|a, \lambda) \propto \exp\{-(c_{|\theta|}\gamma^2 + d_{|\theta|}\gamma) - 0.5(a+1)(\gamma-\lambda)^2\}$. Inspecting $c_{|\theta|}$ and $d_{|\theta|}$, this prior is observed to have a discontinuity at $|\theta| = a\lambda$ and respectively approaches zero as $|\theta| \uparrow a\lambda$ and a density proportional to $\exp\{-0.5(a+1)(\gamma^2+\lambda^2)\}$ as $|\theta| \downarrow a\lambda$. Such behavior presents an interesting contrast to the minimax concave penalty, which has a simpler and arguably more attractive Bayesian motivation while sharing the same sparsity, continuity, and unbiasedness properties.

Hierarchical priors often lead to Bayes estimators with good robustness and frequentist properties [e.g., admissibility and minimaxity; 5]. However, the implications of using penalty functions constructed from hierarchical priors have only received limited attention in the literature on penalized estimation. This paper has demonstrated strong links between hierarchical priors associated with Bayesian procedures having good decision-theoretic and shrinkage properties and frequentist sparse regularization procedures exhibiting good risk performance. The connections outlined here, including those with convex regularization and proximal operator theory, extend well beyond the univariate setting (e.g., Section 2.3) and suggest several novel avenues of investigation in penalized estimation problems.

Acknowledgement

We thank two referees, the associate editor and editor for their helpful comments and encouraging feedback.

Appendix A: Proof of Theorem 2.1

Let $z \in \mathbb{R}$ and $\lambda > 0$ be finite and assume $a > 1$. We wish to prove that (2.2) is finite; that is, $m(z) < \infty$, where

$$m(z) \;=\; (a/2)^{1/2} \int_{\mathbb{R}} \int_0^{\infty} \phi(z-\theta) \exp\{-\gamma|\theta| \;-\; (a/2)(\gamma-\lambda)^2\} d\gamma d\theta. \qquad \text{(A.1)}$$

We may write $m(z) = \int_{\mathbb{R}} \phi(z-\theta)\tilde{\pi}(\theta)d\theta$, where

$$\tilde{\pi}(\theta) = (a/2)^{1/2} \int_0^{\infty} \exp\{-\gamma|\theta|\} \exp\{-(a/2)(\gamma-\lambda)^2\} d\gamma.$$

For any finite a, straightforward computations show that

$$\tilde{\pi}(\theta) = \sqrt{\pi}\, \phi(r_{a/2,\lambda}(0)) M\{r_{a/2,\lambda}(|\theta|)\} \qquad \text{(A.2)}$$

where $r_{a/2,\lambda}(s)$ is defined as in (1.2). Letting $a \to \infty$, it can be shown that (A.2) converges to $\sqrt{\pi}\, e^{-\lambda|\theta|}$.

Supposing first that $a \to \infty$, it is immediately clear that $m(z)$ is finite. Assume now that a is finite. Observe that $r_{a/2,\lambda}(|\theta|) > 0$ if and if only $|\theta| > a\lambda$.

Defining $\mathcal{C} = \{\theta : |\theta| \le a\lambda\}$ and $c_{a,\lambda} = \phi(r_{a/2,\lambda}(0))\sqrt{\pi}$, we may rewrite (A.1) as the sum of three terms:

$$(I) = \int_{\theta\in\mathcal{C}} \phi(z-\theta)\tilde{\pi}(\theta)d\theta,$$

$$(II) = c_{a,\lambda}\int_{a\lambda}^{\infty} \phi(z-\theta)M\{r_{a/2,\lambda}(|\theta|)\}d\theta,$$

$$(III) = c_{a,\lambda}\int_{-\infty}^{-a\lambda} \phi(z-\theta)M\{r_{a/2,\lambda}(|\theta|)\}d\theta.$$

The function $\tilde{\pi}(\theta)$ is evidently bounded on $\mathcal{C}$; hence, term (I) is finite. For term (II), we may use the fact that $M\{r_{a/2,\lambda}(s)\}$ is a strictly decreasing function (e.g., see Lemma C.1 in Appendix C) to conclude that

$$(II) \;\le\; c_{a,\lambda}M(0)\int_{a\lambda}^{\infty}\phi(z-\theta)d\theta \;=\; c_{a,\lambda}M(0)\Phi(z-\lambda a) < \infty.$$

Arguing similarly for term (III),

$$(III) \;\le\; c_{a,\lambda}M(0)\int_{-\infty}^{-a\lambda}\phi(z-\theta)d\theta \;=\; c_{a,\lambda}M(0)\Phi(-z-\lambda a) < \infty.$$

As all three terms are finite, it follows that $m(z) < \infty$, completing the proof.

Appendix B: Proof of Theorem 2.3

Assume $\lambda > 0$ and $a > 0$. Fixing $t \ge 0$, define $g_t(\gamma) = \gamma t + \frac{a}{2}(\gamma-\lambda)^2$ and

$$h(t) \;=\; \lambda\int_0^t \Big(1-\frac{x}{a\lambda}\Big)_+ dx \;=\; \Big(\lambda t - \frac{t^2}{2a}\Big)1_{[t<a\lambda]} + \frac{a\lambda^2}{2}1_{[t\ge a\lambda]}. \tag{B.1}$$

The proof will follow if it can be shown that $\min_{\gamma\ge 0} g_t(\gamma) = h(t)$ for each $t \ge 0$.

Suppose $t = 0$. Then, $g_0(\gamma)$ is minimized at $\gamma = \lambda$; hence, $\min_{\gamma\ge0} g_0(\gamma) = g_0(\lambda) = 0 = h(0)$, proving equivalence for $t = 0$. Now, assume $t > 0$ and observe that $g_t(0) = \frac{a\lambda^2}{2}$. Since $a > 0$, $g_t(\gamma)$ will achieve its minimum value when $\gamma \in [0,\infty)$ either at $\gamma = 0$ or at a finite $\gamma = \tilde{\gamma} > 0$. Suppose first that $\lambda - t/a \le 0$; then, $t \ge a\lambda$ and upon noting that $g_t(\gamma) = g_t(0) + 0.5a\gamma^2 + \gamma(t-a\lambda)$ it is easily seen that $\gamma = 0$ must minimize $g_t(\gamma)$ when $t \ge a\lambda$. Now, suppose $\lambda - t/a > 0$; then, $t \in (0, a\lambda)$ and it is easily shown that $g_t(\gamma) < g_t(0)$ when $0 < \gamma < 2(\lambda - t/a)$. Since $g_t(\gamma)$ is twice continuously differentiable in γ with second derivative $a > 0$, it follows that $\tilde{\gamma} = \lambda - t/a$ is the unique minimizer of $g_t(\gamma)$ for $\gamma \ge 0$ when $0 < t < a\lambda$. In summary, the above arguments show

$$\min_{\gamma\ge0} g_t(\gamma) \;\;=\;\; g_t(\lambda - t/a)\;1_{[t<a\lambda]} \;+\; g_t(0)\;1_{[t\ge a\lambda]} \tag{B.2}$$

and that $\hat{\gamma} = (\lambda - t/a)_+$ minimizes $g_t(\gamma)$ for $\gamma \ge 0$. Evaluating the functions on the right hand side of (B.2) yields (B.1), proving the desired equivalence.

Appendix C: Proof of Theorem 2.4

We begin with a key lemma summarizing several properties of Mills Ratio, i.e., $M(s) = (1-\Phi(s))/\phi(s)$. Throughout, let $q^{(j)}(s)$, $j \geq 1$ denote the j^{th} derivative of an arbitrary function $q(s)$ with respect to s.

Lemma C.1. *For $s \in \mathbb{R}$, (i) $M(s) > 0$ is strictly convex and decreasing; (ii) $M^{(1)}(s) = sM(s) - 1 < 0$ and $M^{(2)}(s) = sM^{(1)}(s) + M(s) > 0$; (iii) $\log(-M^{(1)}(s))$ is strictly convex; (iv) $M^{(1)}(s)M^{(3)}(s) \geq [M^{(2)}(s)]^2$; and, (v) $0 < \frac{d}{ds}(1/M(s)) < 1$. Finally, as $s \to \infty$, (vi) $M(s) = s^{-1} - s^{-3} + O(s^{-5})$; (vii) $1 - sM(s) = (s^2+2)^{-1} + O(s^{-4})$; and, (viii) $\{sM(s)-1\}^{-1} = -s^2 - 3 + O(s^{-2})$.*

Detailed proof of this result is omitted. Result (i) is well known [e.g., 4, Sec. 1]. Result (ii) is an immediate consequence; results (iii) and (iv) then follow as easy consequences of the derivative formulas in (ii) and properties of $M(\cdot)$ established in Baricz [4, Thm. 2.5, Thm. 3.2]. Result (v) is due to [30]. Results (vi)-(viii) are easily proved using well-known asymptotic expansions for $M(s)$; see, for example, Abramowitz and Stegun [1, Eqns. 26.2.12 & 26.2.13].

Turning to the proof of the theorem, and noting that Lemma C.1(ii) implies $p_{\lambda,\alpha}(s)$ is in fact well-defined for all $s \in \mathbb{R}$, it is easy to check that $p_{\lambda,\alpha}(s) > 0$ for $s \geq 0$ and continuously differentiable for $s \in \mathbb{R}$. Moreover, Lemma C.1(ii) and (iii) and the fact that $r_{\lambda,\alpha}(s) = s/\sqrt{2\alpha} - \lambda\sqrt{2\alpha}$ is an increasing function imply that (2.7) is strictly concave function for $s \in \mathbb{R}$. Differentiating (2.7), we further have

$$p^{(1)}_{\lambda,\alpha}(s) \quad = \quad -\frac{M^{(2)}\{r_{\lambda,\alpha}(s)\}}{(2\alpha)^{1/2} M^{(1)}\{r_{\lambda,\alpha}(s)\}}. \tag{C.1}$$

By Lemma C.1(ii), (C.1) is positive for $s \in \mathbb{R}$; since $r_{\lambda,\alpha}(s)$ is strictly increasing, it follows that $p_{\lambda,\alpha}(s)$ is strictly increasing and satisfies $p^{(1)}_{\lambda,\alpha}(0+) = p^{(1)}_{\lambda,\alpha}(0) > 0$. Straightforward computations further show

$$p^{(2)}_{\lambda,\alpha}(s) \; = \; \frac{[M^{(2)}\{r_{\lambda,\alpha}(s)\}]^2 - M^{(1)}\{r_{\lambda,\alpha}(s)\}\, M^{(3)}\{r_{\lambda,\alpha}(s)\}}{2\alpha[M^{(1)}\{r_{\lambda,\alpha}(s)\}]^2}.$$

By Lemma C.1(iv), this last expression is never positive; as $r_{\lambda,\alpha}(s)$ is also strictly increasing, it follows that (C.1) is a decreasing function for $s \in \mathbb{R}$.

It remains to show (i) $-s - p^{(1)}_{\lambda,\alpha}(s)$ is strictly decreasing and unimodal, hence achieves a maximum value of $-p'_{\lambda,\alpha}(0)$ at $s = 0$; (ii) $|s| + p^{(1)}_{\lambda,\alpha}(|s|)$ is minimized at $s = 0$, with minimum value $p^{(1)}_{\lambda,\alpha}(0)$; and, (iii) $p^{(1)}_{\lambda,\alpha}(s) = O(s^{-1})$ as $s \to \infty$. Result (iii) follows upon noting that the expansions in Lemma C.1(vi)-(viii) can be used directly in conjunction with (C.1) to prove that $p^{(1)}_{\lambda,\alpha}(s) = O(s^{-1})$ as $s \to \infty$. Results (i) and (ii) evidently follow if it can be shown that $d(s) = s + p^{(1)}_{\lambda,\alpha}(s)$ is strictly increasing for $s \geq 0$, hence achieving its minimum for $s \geq 0$ at $s = 0$. Towards this end, note that for $s \geq 0$, we have from (C.1), Lemma C.1(ii), and

$r_{\lambda,\alpha}(s) = s/\sqrt{2\alpha} - \lambda\sqrt{2\alpha}$ that

$$d(s) = s + p^{(1)}_{\lambda,\alpha}(s) \;=\; s\left(1 - \frac{1}{2\alpha}\right) + \lambda - \frac{M\{r_{\lambda,\alpha}(s)\}}{(2\alpha)^{1/2} M^{(1)}\{r_{\lambda,\alpha}(s)\}}.$$

Taking the derivative of both sides with respect to s and simplifying, we find

$$d^{(1)}(s) \;=\; \frac{M\{r_{\lambda,\alpha}(s)\}M^{(2)}\{r_{\lambda,\alpha}(s)\}}{2\alpha[M^{(1)}\{r_{\lambda,\alpha}(s)\}]^2} + 1 - \frac{1}{\alpha}.$$

Suppose that $M(u)M^{(2)}(u) > [M^{(1)}(u)]^2, \quad u \in \mathbb{R}$. Then, assuming $\alpha > \frac{1}{2}$, it immediately follows that $d^{(1)}(s) > 0$ for $s \geq 0$, proving that $d(s) = s + p^{(1)}_{\lambda,\alpha}(s)$ is strictly increasing for $s \geq 0$. Consequently, it only remains to prove $M(u)M^{(2)}(u) > [M^{(1)}(u)]^2, \quad u \in \mathbb{R}$. Using Lemma C.1(ii), it is easily shown that this inequality is equivalent to $M^2(u) + uM(u) - 1 > 0$. However, this is guaranteed by Lemma C.1(v). In particular, differentiating $1/M(u)$, algebra results in the inequality $M^2(u) + M^{(1)}(u) > 0$; Lemma C.1(ii) then yields $M^2(u) + uM(u) - 1 > 0$ for $u \in \mathbb{R}$, completing the proof.

Appendix D: Proof of Theorem 2.7

The stated assumptions imply that $\lambda_{min} > 0$ and hence that $\tilde{g}(\boldsymbol{\theta}) = g(\boldsymbol{\theta}) - 0.5\,\delta\,\boldsymbol{\theta}'\boldsymbol{\theta}$ is strictly convex on Ω provided that $\delta \leq \lambda_{min}$ [e.g., 40, Lemma 1]. Define for $y_1, y_2 \geq 0$ and $i = 1 \ldots L$ the functions $q_i(y_1, y_2) = 0.5\,(\delta/w_i)\,y_1^2 + y_1 y_2 + 0.5a\,(y_2 - \lambda)^2$. Then, since $\boldsymbol{\theta}'\boldsymbol{\theta} = \sum_{=1}^{L} \|\boldsymbol{\theta}_i\|^2$, the objective function in (2.10) may be written as $k(\boldsymbol{\theta}, \gamma) = \tilde{g}(\boldsymbol{\theta}) + \sum_{i=1}^{L} w_i\, q_i(\|\boldsymbol{\theta}_i\|, \gamma_i)$. It is straightforward to check that $q_i(y_1, y_2)$ is strictly convex for $(y_1, y_2) \in \mathbb{R}^2_+$ and $i = 1 \ldots L$ provided that $a > w_{max}/\delta$, hence monotone increasing in each coordinate [34, Thm. 4.2].

Define $\mathbf{z} = (\boldsymbol{\theta}', \boldsymbol{\gamma}')'$, a $(p + L) \times 1$ vector. For $i = 1 \ldots L$, let B_{1i} be a set of $p_i \times (p + L)$ matrices such that $B_{1i}\mathbf{z} = \boldsymbol{\theta}_i$; similarly, define B_{2i} such that $B_{2i}\mathbf{z} = \gamma_i$ and B such that $B\mathbf{z} = \boldsymbol{\theta}$. Since $q_i(\|B_{1i}\,\mathbf{z}\|, B_{2i}\,\mathbf{z}) = q_i(\|\boldsymbol{\theta}_i\|, \gamma_i)$, $k(\boldsymbol{\theta}, \boldsymbol{\gamma}) = \tilde{g}(B\mathbf{z}) + \sum_{i=1}^{L} w_i\, q_i(\|B_{1i}\,\mathbf{z}\|, B_{2i}\,\mathbf{z})$. Since $B\mathbf{z}$ is convex for $\mathbf{z} \in \Omega \times \mathbb{R}^L_+$, $\tilde{g}(B\mathbf{z})$ is also convex there. The functions $\|B_{1i}\,\mathbf{z}\|$ and $B_{2i}\,\mathbf{z}$ for $i = 1 \ldots L$ are convex for $\mathbf{z} \in \Omega \times \mathbb{R}^L_+$; hence, $q_i(\|B_{1i}\,\mathbf{z}\|, B_{2i}\,\mathbf{z})$ is strictly convex there for $i = 1 \ldots L$ and so is $\sum_{i=1}^{L} w_i q_i(\|B_{1i}\,\mathbf{z}\|, B_{2i}\,\mathbf{z})$. The above establishes strict convexity of $k(\boldsymbol{\theta}, \boldsymbol{\gamma}) = k(\mathbf{z})$ for $\mathbf{z} \in \Omega \times \mathbb{R}^L_+$, hence the existence of a unique minimizer. Set $\hat{\boldsymbol{\gamma}}(\boldsymbol{\theta}) = \operatorname{argmin}_{\boldsymbol{\gamma} \in \mathbb{R}^L_+} k(\boldsymbol{\theta}, \boldsymbol{\gamma})$; result (2.8) now implies (2.9) equals $k(\boldsymbol{\theta}, \hat{\boldsymbol{\gamma}}(\boldsymbol{\theta}))$, a strictly convex function for $\boldsymbol{\theta} \in \Omega$ if $a/w_{max} > \delta^{-1} \geq \lambda^{-1}_{min}$ [e.g., 29, Prop. 2.22]. But, as (2.9) and (2.10) are both strictly convex on their respective domains and have the same global minimum [29, Prop. 1.35], their solutions must also coincide.

References

[1] Abramowitz, M. and Stegun, I. (1970). *Handbook of mathematical functions*. Dover Publications Inc., New York.

[2] Antoniadis, A. and Fan, J. (2001). Regularization of Wavelet Approximations. *J. Am. Statist. Assoc.* **96** 939-955. MR1946364

[3] Armagan, A., Dunson, D. and Lee, J. (2011). Generalized double Pareto shrinkage. *ArXiv e-prints*. MR2803744

[4] Baricz, A. (2008). Mills' ratio: Monotonicity patterns and functional inequalities. *J. Math. Anal. Applic.* **340** 1362-1370. MR2390935

[5] Berger, J. O. and Robert, C. (1990). Subjective hierarchical Bayes estimation of a multivariate normal mean: on the frequentist interface. *Ann. Statist.* **18** 617–651. MR1056330

[6] Berger, J. O. and Strawderman, W. E. (1996). Choice of hierarchical priors: admissibility in estimation of normal means. *Ann. Statist.* **24** 931–951. MR1401831

[7] Berger, J. O., Strawderman, W. E. and Tang, D. (2005). Posterior Propriety and Admissibility of Hyperpriors in Normal Hierarchical Models. *Ann. Statist.* **33** 606–646. MR2163154

[8] Box, G. E. P. and Tiao, G. C. (1992). *Bayesian Inference in Statistical Analysis (1973 ed., Wiley Classics Library)*. John Wiley and Sons, New York.

[9] Breheny, P. and Huang, J. (2009). Penalized methods for bi-level variable selection. *Stat. Interface* **2** 369–380. MR2540094

[10] Breheny, P. and Huang, J. (2011). Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. *Ann. Appl. Stat.* **5** 232–253. MR2810396

[11] Bruce, A. G. and Gao, H. Y. (1996). *Applied Wavelet Analysis with S-Plus*. Springer, New York.

[12] Carvalho, C. M., Polson, N. G. and Scott, J. G. (2010). The horseshoe estimator for sparse signals. *Biometrika* **97** 465-480. MR2650751

[13] Chen, M.-H., Ibrahim, J. G. and Shao, Q.-M. (2006). Posterior Propriety and Computation for the Cox Regression Model with Applications to Missing Covariates. *Biometrika* **93** pp. 791-807. MR2285072

[14] Chen, M.-H. and Shao, Q.-M. (2001). Propriety of Posterior Distribution for Dichotomous Quantal Response Models. *Proceedings of the American Mathematical Society* **129** pp. 293-302. MR1694452

[15] Cox, D. R. (1972). Regression Models and Life-Tables. *Journal of the Royal Statistical Society. Series B (Methodological)* **34** pp. 187-220. MR0341758

[16] Fan, J. and Li, R. (2001). Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties. *J. Am. Statist. Assoc.* **96** 1348–1360. MR1946581

[17] Fourdrinier, D., Strawderman, W. E. and Wells, M. T. (1998). On the construction of Bayes minimax estimators. *Ann. Statist.* **26** 660–671. MR1626063

[18] Gao, H. and Bruce, A. G. (1997). Waveshrink with firm shrinkage. *Statist. Sinica* **7** 855–874. MR1488646

[19] Gomez-Sanchez-Manzano, E., Gomez-Villegas, M. A. and Marin, J. M. (2008). Multivariate exponential power distributions as mixtures of normal distributions with Bayesian applications. *Comm. Stat. Thry. Meth.* **37** 972-985.

[20] Griffin, J. E. and Brown, P. J. (2007). Bayesian adaptive Lassos with non-convex penalization. Technical Report, Dept. of Statistics, University of Warwick.

[21] Griffin, J. E. and Brown, P. J. (2010). Inference with normal-gamma prior distributions in regression problems. *Bayesian Analysis* **6** 171–188. MR2596440

[22] Hans, C. (2009). Bayesian Lasso regression. *Biometrika* **96** 835–845. MR2564494

[23] Johnstone, I. M. and Silverman, B. W. (2004). Needles and straw in haystacks: empirical Bayes estimates of possibly sparse sequences. *Ann. Statist.* **32** 1594–1649. MR2089135

[24] Kass, R. E. and Wasserman, L. (1996). The Selection of Prior Distributions by Formal Rules. *Journal of the American Statistical Association* **91** pp. 1343-1370.

[25] Mazumder, R., Friedman, J. H. and Hastie, T. (2011). SparseNet: Coordinate Descent With Nonconvex Penalties. *Journal of the American Statistical Association* **106** 1125-1138. MR2894769

[26] Park, T. and Casella, G. (2008). The Bayesian Lasso. *J. Am. Statist. Assoc.* **103** 681–686. MR2524001

[27] Polson, N. G. and Scott, J. G. (2011). Shrink Globally, Act Locally: Sparse Bayesian Regularization and Prediction (with discussion). In *Bayesian Statistics 9* (J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman and M. West, eds.) 501–525. Oxford University Press.

[28] Robert, C. P. (2007). *The Bayesian Choice*. Springer-Verlag, New York. MR2723361

[29] Rockafellar, R. T. and Wets, R. J. B. (2004). *Variational Analysis*. Springer-Verlag, Berlin. MR1491362

[30] Sampford, M. R. (1953). Some Inequalities on Mill's Ratio and Related Functions. *Ann. Math. Statist.* **24** 130–132. MR0054890

[31] Schifano, E. D. (2010). Topics in Penalized Estimation PhD thesis, Cornell University. MR2801759

[32] Schifano, E. D., Strawderman, R. L. and Wells, M. T. (2010). Majorization-minimization algorithms for nonsmoothly penalized objective functions. *Electron. J. Stat.* **4** 1258–1299. MR2738533

[33] Strawderman, W. E. (1971). Proper Bayes minimax estimators of the normal multivariate normal distribution. *Ann. Math. Statist.* **42** 385-388. MR0397939

[34] Strawderman, R. L. and Wells, M. T. (2012). On Hierarchical Prior Specifications and Penalized Likelihood. In *Contemporary Developments in*

Bayesian Analysis and Statistical Decision Theory: A Festricht for William E. Strawderman, (D. Fourdrinier, E. Marchand and A. Ruhkin, eds.) **8** 154-180. Institute of Mathematical Statistics, Hayward, CA.

[35] Takada, Y. (1979). Stein's positive part estimator and Bayes estimator. *Ann. Inst. Statist. Math.* **31** 177-183. MR0550792

[36] Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *J. R. Statist. Soc. B* **58** 267–288. MR1379242

[37] Tipping, M. E. (2001). Sparse Bayesian learning and the relevance vector machine. *J. Mach. Learn. Res.* **1** 211–244. MR1875838

[38] Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *J. R. Statist. Soc. B* **68** 49–67. MR2212574

[39] Zhang, C.-H. (2010). Nearly Unbiased Variable Selection Under Minimax Concave Penalty. *Ann. Statist.* **38** 894–942. MR2604701

[40] Zlobec, S. (2003). Estimating convexifiers in continuous optimization. *Math. Comm.* **8** 129-137. MR2026391

[41] Zou, H. and Li, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models. *Ann. Statist.* **36** 1509–1533. MR2435443