IMS Collections
Contemporary Developments in Bayesian Analysis and Statistical Decision Theory: A Festschrift for William E. Strawderman
Vol. 8 (2012) 154–180
© Institute of Mathematical Statistics, 2012
DOI: 10.1214/11-IMSCOLL811

On hierarchical prior specifications and penalized likelihood*

Robert L. Strawderman[1] and Martin T. Wells[1]

Cornell University

Abstract: Using a Bayesian model with a class of hierarchically specified scale-mixture-of-normals priors as motivation, we consider a generalization of the grouped LASSO in which an additional penalty is placed on the penalty parameter of the L_2 norm. We show that the resulting MAP estimator obtained by jointly minimizing the corresponding objective function in both the mean and penalty parameter is a thresholding estimator that generalizes (i) the grouped lasso estimator of Yuan and Lin (2006) and (ii) the univariate minimax concave penalization procedure of Zhang (2010) to the setting of a vector of parameters. An exact formula for the risk and a corresponding SURE formula are obtained for the proposed class of estimators.

A new universal threshold is proposed under appropriate sparsity assumptions; in combination with the proposed class of estimators, we subsequently obtain a new and interesting motivation for the class of positive part estimators. In particular, we establish that the original positive part estimator corresponds to a suboptimal choice of this thresholding parameter. Numerical comparisons between the proposed class of estimators and the positive part estimator show that the former can achieve further, significant reductions in risk near the origin.

Contents

*This paper is dedicated to Bill Strawderman. Bill has advanced statistical science in deep and original ways, and the statistics community has been inestimably enriched by his research contributions, intellectual leadership, and gracious mentorship of many junior colleagues. On a more personal note, Rob wishes to express his sincere gratitude for all that his father Bill has done, both directly and indirectly, to support him throughout his life and career. The support of NSF Grant 06-12031 and NIH Grant R01-GM083606-01 are gratefully acknowledged.

[1]Department of Statistical Science, Comstock Hall, Ithaca NY 14853 e-mail: rls54@cornell.edu; mtw1@cornell.edu

AMS 2000 subject classifications: 62C10, 62C12, 62C20, 62F10, 62F15

Keywords and phrases: hierarchical models, grouped lasso, lasso, maximum aposteriori estimate, minimax concave penalty, penalized likelihood, positive-part estimator, regularization, restricted parameter space, Stein estimation

1. Introduction

Let $\boldsymbol{X} \sim N_p(\boldsymbol{\theta}, \boldsymbol{I}_p)$ and suppose it is desired to estimate the unknown vector $\boldsymbol{\theta} \in \mathbb{R}^p$. The standard estimator of $\boldsymbol{\theta}$ is $\delta_0(\boldsymbol{X}) = \boldsymbol{X}$, in the sense that it is the maximum likelihood estimator, the uniformly minimum variance unbiased estimator, and the least squares estimator. Under a wide variety of loss functions, $\delta_0(\boldsymbol{X})$ is also the minimum risk equivariant estimator, a minimax estimator, and it is admissible for $p = 1$ or 2. However, [38] showed that $\boldsymbol{X}$ is also inadmissible if $p \geq 3$ for the squared error loss $L(\boldsymbol{\theta}, \delta) = \|\delta(\boldsymbol{X}) - \boldsymbol{\theta}\|_2^2$. This result was remarkable at the time and has since led to a vast number of developments in multiparameter estimation, a field that Bill Strawderman has deeply influenced and in which he continues to expand the research frontier. An important aspect of this so-called "Stein-phenomenon" is that it illustrates the inherent difference in the problems of simultaneously estimating $\boldsymbol{\theta}$ versus a single component, say θ_i. Indeed, for each $i = 1 \ldots p$, $\delta_{0i}(\boldsymbol{X}) = X_i$ is an admissible estimator of θ_i whatever the value of p.

The risk of the MLE $\delta_0(\boldsymbol{X})$ is p. [27] showed that

$$\delta_{JS}^c(\boldsymbol{X}) = \left(1 - \frac{c}{\|\boldsymbol{X}\|_2^2}\right)\boldsymbol{X} \tag{1.1}$$

dominates $\delta_0(\boldsymbol{X})$ when $p \geq 3$ provided that $0 < c < 2(p-2)$; moreover, they demonstrated that the risk of $\delta_{JS}(\boldsymbol{X}) \equiv \delta_{JS}^{p-2}(\boldsymbol{X})$ equals 2 at $\boldsymbol{\theta} = 0$ for all $p \geq 3$, indicating the substantial gains in risk near the origin for large p. However, $\delta_{JS}^c(\boldsymbol{X})$ is known to be inadmissible since it can be dominated by Stein's positive part estimator [1, 2]. James-Stein estimators are such that, when $\|\boldsymbol{X}\|_2^2 < c$, the multiplier of $\boldsymbol{X}$ becomes negative and, furthermore, $\lim_{\|\boldsymbol{X}\|_2 \to 0} \|\delta_{JS}^c(\boldsymbol{X})\|_2 = \infty$. It follows that, for any $K > 0$, there exists $\eta > 0$ such that $\|\boldsymbol{X}\|_2 < \eta$ implies $\|\delta_{JS}^c(\boldsymbol{X})\|_2 > K$. Hence an observation that would lead to almost certain acceptance of $H_0 : \boldsymbol{\theta} = 0$ can give rise to an estimate very far from zero. Furthermore, $\delta_{JS}^c(\boldsymbol{X})$ is not coordinatewise-monotone in $\boldsymbol{X}$ in the sense that a larger value of $\boldsymbol{X}$ in a particular coordinate may lead to a smaller estimate of the mean of that coordinate. Such behavior is clearly undesirable.

One possible remedy is to modify the James-Stein estimator and use the positive-part of its multiplier, namely

$$\delta_{JS+}^c(\boldsymbol{X}) = \left(1 - \frac{c}{\|\boldsymbol{X}\|_2^2}\right)_+ \boldsymbol{X} \tag{1.2}$$

where $(t)_+ = \max(t, 0)$. This is a particular example of a Baranchik-type estimator [1, 2]. Results of [10] imply the positive-part version is itself inadmissible, although this result was assumed to be true much earlier. Results due to [8] and [11] imply that it is impossible to find an estimator that simultaneously dominates the positive-part estimator and whose unbiased estimator of risk is uniformly smaller (i.e., in $\boldsymbol{X}$) than that of the positive-part estimator. In practical terms, this shows that

improving upon the positive-part estimator is difficult and the usual tools of the trade for constructing improved estimators fail. In a landmark paper, [36] introduce an explicit class of estimators that dominate (1.2).

In this article, we develop a competing class of competitors to (1.2). The motivation for this work stems from that of [42], who proves that (1.2) also arises as the maximum aposteriori estimator (MAP) under a certain class of hierarchically specified but improper prior distributions that, in qualitative terms, behaves similarly to the class of proper hierarchical prior distributions introduced by [40] in his seminal work on proper Bayes minimax estimation of $\boldsymbol{\theta}$. The results of [42] are reviewed in Section 2; there, we first briefly review some key results from decision theory on Bayes minimax estimators. We also introduce some new hierarchical prior specifications that lead to marginal priors on $\boldsymbol{\theta}$ that are equivalent to those obtained using the scale mixtures of multivariate normal distributions respectively considered in [40] and [42]. In Section 3, we then connect this work to current research on penalized likelihood estimation, establishing in particular relationships between the popular lasso and grouped lasso estimators and MAP estimation under these new hierarchical prior specifications. In Section 4, we exploit these connections and the ideas of [42] to motivate a new class of penalized likelihood estimators that can be interpreted as MAP estimators under specific marginalizations of these prior specifications. Formulas for the theoretical and empirical risk of these estimators are derived in Section 5, and include numerical studies of performance. We close this article in Section 6 with a discussion.

2. Estimation with hierarchical prior specifications

2.1. Minimax estimators derived from Bayes principles

The classical Stein estimate and its positive part modification can be motivated in a number of ways, perhaps most commonly as empirical Bayes estimates (i.e., posterior means) under a normal hierarchical model in which $\boldsymbol{\theta} \sim \mathrm{N}_p(\mathbf{0}, \psi \boldsymbol{I}_p)$ and ψ, viewed as a hyperparameter, is estimated [e.g. 17, 9, 22]. [10] proved that any admissible estimator of the mean must be generalized Bayes, that is, minimizes the posterior expected squared error loss under a possibly improper prior. It is well known that neither (1.1) nor (1.2) are admissible estimators [e.g. 10]; however, like the MLE $\delta_0(\boldsymbol{X})$, both estimators are minimax [e.g. 2]. [40, 41], addressing an earlier conjecture due to Stein on the existence of proper Bayes minimax estimators for $\boldsymbol{\theta}$, proves that such estimators only exist for $p \geq 5$.

The class of proper Bayes minimax estimators constructed in [40] relies on the use of a hierarchically specified class of proper prior distributions $\pi_S(\boldsymbol{\theta}, \kappa)$. In particular, $\pi_S(\boldsymbol{\theta}, \kappa)$ is specified according to

$$\boldsymbol{\theta}|\kappa \sim \mathrm{N}_p(\mathbf{0}, g(\kappa)\boldsymbol{I}_p), \quad \pi_S(\kappa) = (1-a)\kappa^{-a}\, \mathbb{1}_{[0<\kappa<1]}, \tag{2.1}$$

where $g(\kappa) = (1-\kappa)/\kappa$ and the constant a satisfies $0 \leq a < 1$ (i.e., $\pi_S(\kappa)$ is a Beta$(1-a, 1)$ probability distribution); see [3, 4] for related generalizations and [5] and [6] for further work on hierarchically specified priors in normal models. Suppose $a = 1/2$; then, with $\psi = g(\kappa) > 0$ in (2.1), we obtain the equivalent specification

$$\boldsymbol{\theta}|\psi \sim \mathrm{N}_p(\mathbf{0}, \psi \boldsymbol{I}_p), \quad \pi_S(\psi) = \frac{1}{2}\left(\frac{1}{1+\psi}\right)^{\frac{3}{2}}\, \mathbb{1}_{[\psi>0]}. \tag{2.2}$$

Two interesting alternative formulations of (2.2), given below for the case $p = 1$ and generalized later for arbitrary p, are provided in Theorem 2.1 below. In what follows, and hereafter, we let Gamma(τ, ξ) denote for $\tau, \xi > 0$ a random variable with probability density function

$$g(x|\tau,\xi) = \frac{\xi^\tau}{\Gamma(\tau)} x^{\tau-1} e^{-x\xi}\, \mathbb{1}_{[x>0]}$$

and Exp(ξ) correspond to the choice $\tau = 1$ (i.e., an exponential random variable in its rate parameterization).

Theorem 2.1. *For $p = 1$, the marginal prior distribution on θ induced by (2.2) is equivalent to that obtained under the specification*

$$\theta|\psi,\lambda \sim \mathrm{N}(0,\psi), \quad \psi|\lambda \sim \mathrm{Exp}\left(\frac{\lambda^2}{2}\right), \quad \lambda|\varpi \sim \mathrm{HN}(\varpi^{-1}), \tag{2.3}$$

where $\varpi = 1$ and HN(ζ) *denotes for $\zeta > 0$ the half-normal density*

$$f(x|\zeta) = \sqrt{\frac{2}{\pi\zeta}} \exp\left\{-\frac{x^2}{2\zeta}\right\} \mathbb{1}_{[x>0]}.$$

The marginal prior distribution on θ induced by (2.2) is also equivalent to that obtained under the alternative specification

$$\theta|\lambda \sim \mathrm{DoubExp}(\lambda), \quad \lambda|\varpi \sim \mathrm{HN}(\varpi^{-1}), \tag{2.4}$$

where $\varpi = 1$ and DoubExp(λ) denotes a random variable with the double exponential probability density function

$$f(y|\lambda) = \frac{\lambda}{2} e^{-\lambda|y|}\, \mathbb{1}_{[y\in\mathbb{R}]}.$$

Proof. Following [24, 25], define

$$\theta|\psi,\omega \sim \mathrm{N}(0,\psi), \quad \psi|\omega \sim \mathrm{Exp}(\omega), \quad \omega|\delta,\varpi \sim \mathrm{Gamma}(1/2,\varpi). \tag{2.5}$$

as a hierarchically specified prior distribution for θ, ψ and ω. The resulting marginal prior distribution for θ, obtained by integrating out ψ and ω, is exactly the quasi-Cauchy distribution of [28, 29]; see [24, 25] for details. [13] show that this distribution also coincides with the marginal prior distribution for θ induced by taking $a = 1/2$ in (2.1). The transformation $\lambda = \sqrt{2\omega}$ in (2.5) leads directly to (2.3) upon setting $\varpi = 1$; (2.4) is then obtained by integrating out ψ in (2.3). □

2.2. The positive part estimator as a MAP estimator

In a very interesting paper, [42] proves that the minimax estimator (1.2) is also the maximum aposteriori (MAP) estimator under a certain class of hierarchically specified improper prior distributions, say $\pi_T(\boldsymbol{\theta}, \kappa) = \pi(\boldsymbol{\theta}|\kappa)\pi_T(\kappa)$. For the specific choice $c = p - 2$ in (1.2), Takada's prior reduces to

$$\boldsymbol{\theta}|\kappa \sim \mathrm{N}_p(\mathbf{0}, g(\kappa)\boldsymbol{I}_p), \quad \pi_T(\kappa) \propto (1-\kappa)^{p/2}\kappa^{-1}\, \mathbb{1}_{[0<\kappa<1]}. \tag{2.6}$$

The prior (2.6) evidently behaves similarly to Strawderman's proper prior (2.1) (i.e., for $a = 1/2$) and the improper generalizations of this prior considered in [3, 4]

and [20], particularly so as $\kappa \to 0$. Notably, the numerator $(1-\kappa)^{p/2}$ in $\pi_T(\kappa)$ explicitly offsets the contribution of $(1-\kappa)^{-p/2}$ arising from the determinant of the variance matrix $g(\kappa)\boldsymbol{I}_p$ in the conditional prior specification $\boldsymbol{\theta}|\kappa$. Under the monotone decreasing variable transformation $\psi = g(\kappa) > 0$, (2.6) implies an alternative representation that is analogous to (2.2):

$$\boldsymbol{\theta}|\psi \sim \mathrm{N}_p(\mathbf{0}, \psi \boldsymbol{I}_p), \quad \pi_T(\psi) \propto \psi^{\frac{p}{2}} \left(\frac{1}{1+\psi}\right)^{\frac{p}{2}+1} \mathbb{1}_{[\psi>0]}. \tag{2.7}$$

We observe that the proper prior $\pi_S(\psi)$ in (2.2) and improper prior $\pi_T(\psi)$ (2.7) nearly coincide when $p = 1$; in particular, multiplying the former by $\psi^{1/2}$ yields the latter. In view of the fact that (2.2) and (2.3) lead to the same marginal prior on $\boldsymbol{\theta}$ when $p = 1$, one is led to question whether a deeper connection between these two prior specifications might exist. Supposing $p \geq 1$, consider the following straightforward generalization of (2.3):

$$\boldsymbol{\theta}|\psi, \lambda \sim \mathrm{N}_p(\mathbf{0}, \psi \boldsymbol{I}_p), \ \psi|\lambda \sim \text{Gamma}\left(\frac{p+1}{2}, \frac{\lambda^2}{2}\right), \ \lambda|\varpi \sim \mathrm{HN}(\varpi^{-1}). \tag{2.8}$$

Integrating λ out of the higher level prior specification

$$\psi|\lambda \sim \text{Gamma}\left(\frac{p+1}{2}, \frac{\lambda^2}{2}\right), \quad \lambda|\varpi \sim \mathrm{HN}(\varpi^{-1}),$$

the resulting marginal (proper) prior for ψ reduces to

$$\pi(\psi|\varpi) \propto \psi^{-1/2} \psi^{p/2} \left(\frac{1}{1+\frac{\psi}{\varpi}}\right)^{\frac{p}{2}+1} \mathbb{1}_{[\psi>0]}. \tag{2.9}$$

For $\varpi = 1$ and any $p \geq 1$, the proper prior (2.9) is observed to be equal to the improper prior $\pi_T(\psi)$ in (2.7), multiplied by $\psi^{-1/2}$, and reduces Strawderman's prior (2.2) when $p = 1$.

3. Penalized likelihood and hierarchical prior specifications

Expressed in modern terms, [42] proves that the positive part estimator (1.2) is the solution to a certain penalized likelihood estimation problem in which the penalty (or regularization) term is determined by the prior (2.6). Penalized likelihood estimation, and more generally problems of regularized estimation, have become very important conceptual paradigms in both statistics and machine learning. Such methods suggest principled estimation and model selection procedures for a variety of high-dimensional problems. Regularization by squared Euclidean norms has been thoroughly studied. In recent years, regularization through the use of other norms has generated considerable interest; a particularly prevalent example is ℓ_1 norm regularization, that is, the so-called "lasso" problem [43]. Problems involving regularization (and penalization) using norms other than the squared Euclidean norm typically cannot be solved using simple linear algebra; tools for solving both convex and also non-convex optimization minimization problems are needed.

In recent years, the statistical literature on penalized likelihood estimators has exploded, in part due to success in constructing procedures for regression problems in which one can simultaneously select variables and estimate their effects. The

class of penalty functions leading to procedures with good asymptotic frequentist properties have singularities at the origin; important examples of separable penalties include the lasso [43], smoothly clipped absolute deviation [SCAD; 18], and minimax concave [MCP; 48] penalties. In fact, most such penalties utilized in the literature behave similarly to the lasso penalty near the origin, differing more in their respective behaviors away from the origin, where control of estimation bias for those parameters not estimated to be zero becomes the driving concern. In a regression context, the main purpose of using a singular, separable penalty is to permit one to estimate regression coefficients as being either nonzero or exactly equal to zero, thereby permitting simultaneity in both estimation and variable selection. Generalizations of the lasso penalty have recently been proposed to deal with correlated groupings of parameters, such as those that might arise in problems with features that can be sensibly ordered [e.g., fused lasso; 44] or separated into distinct subgroups [e.g., grouped lasso; 46]. In such problems, the use of these penalties serves a related purpose. For example, in the case of the grouped lasso, the goal is still to permit the possibility of simultaneous selection and estimation; however, unlike the standard lasso penalty, the process of selection occurs at the group level (i.e., all coefficients in a group are estimated as zero, or none are).

The lasso was initially formulated as a least squares estimation problem subject to a ℓ_1 constraint on the parameter vector. The more well-known penalized likelihood formulation arises from a Lagrange multiplier formulation of this constrained optimization problem. Since the underlying objective function is separable in the parameters, the underlying estimation problem is evidently directly related to the now-classical problem of estimating a bounded normal mean. From a decision theoretic point of view, if $X \sim N(\theta, 1)$ for $|\theta| \leq \lambda$ then the projection of the usual estimator dominates the unrestricted MLE but cannot be minimax for quadratic loss because it is not a Bayes estimator. [14] showed that the unique minimax estimator of θ is the Bayes estimator corresponding to a two point prior on $\{-\lambda, \lambda\}$ for λ sufficiently small. [14] further showed that the uniform boundary Bayes estimator, $\lambda \tanh(\lambda x)$, is the unique minimax estimator if $\lambda < \lambda_0 \approx 1.0567$. They also considered three-point priors supported on $\{-\lambda, 0, \lambda\}$ and obtained sufficient conditions for such a prior to be least favorable. [30] considered the multivariate extension, $\boldsymbol{X} \sim N_p(\boldsymbol{\theta}, \boldsymbol{I}_p)$ with $\|\boldsymbol{\theta}\|_2 \leq \lambda$ and showed that the Bayes estimator with respect to a boundary uniform prior dominates the MLE whenever $\lambda \leq \sqrt{p}$ under squared error loss.

It has long been recognized that the class of penalized likelihood estimators also has a Bayesian interpretation. For example, in the canonical version of the "lasso" problem, minimizing

$$\frac{1}{2}\|\boldsymbol{X} - \boldsymbol{\theta}\|_2^2 + \lambda\|\boldsymbol{\theta}\|_1, \quad ||\boldsymbol{\theta}||_1 = \sum_{i=1}^{p} |\theta_i| \tag{3.1}$$

with respect to $\boldsymbol{\theta}$ is easily seen to be equivalent to computing the MAP estimator of $\boldsymbol{\theta}$ under a model specification in which $\boldsymbol{X} \sim \mathrm{N}_p(\boldsymbol{\theta}, \boldsymbol{I}_p)$ and $\boldsymbol{\theta}$ has a prior distribution satisfying $\theta_i \overset{iid}{\sim} \mathrm{DoubExp}(\lambda)$. It is easily shown that the solution to (3.1) is $\widehat{\theta}_i(\boldsymbol{X}) = \mathrm{sign}(X_i)(|X_i| - \lambda)_+$, $i = 1 \ldots p$. The critical hyperparameter λ, though regarded as fixed for the purposes of estimating $\boldsymbol{\theta}$, is typically estimated in some adhoc manner (e.g., cross validation), resulting in an estimator with an empirical Bayes flavor.

In the machine learning literature, the double exponential density has been widely used as a sparsity-inducing prior in various contexts [e.g., 32, 19, 37]. As suggested in (2.3) and (2.4), a DoubExp(λ) distribution has a hierarchical repre-

sentation, being obtained by treating λ as fixed and then integrating out ψ in (2.3). [19] directly exploits this representation in deriving an EM algorithm for computing the MAP estimator in a regression version of (3.1). [19] also introduces a variation on this scheme in which the exponential prior distribution on ψ in (2.3) is replaced by a version of Jeffrey's prior. A related model in this class is known as the Relevance Vector Machine [7, 45], where the marginal prior on $\boldsymbol{\theta}$ is constructed from a product of independent Student $t-$distributions. In hierarchical form, this corresponds to a prior constructed from independent normal priors with distinct scale parameters, each having a gamma density; compare (2.8).

As suggested in Theorem 2.1 and elsewhere, the double exponential prior inherent in the lasso minimization problem (3.1) has broad connections to estimation under hierarchical prior specifications that lead to scale mixtures of normal distributions. As pointed out above, the conditional prior distribution of $\theta|\lambda$ obtained by integrating out ψ in (2.3) is exactly DoubExp(λ). More generally, the conditional distribution for $\boldsymbol{\theta}|\lambda$ under the hierarchical prior specification (2.8) is a special case of the class of multivariate exponential power distributions [cf. 23, Thm. 2.1]; in particular, we obtain

$$\pi(\boldsymbol{\theta}|\lambda) \propto \lambda^p \exp\{-\lambda\|\boldsymbol{\theta}\|_2\}, \tag{3.2}$$

a direct generalization of the double exponential distribution that arises when $p = 1$. Treating λ as fixed hyperparameter, computation of the resulting MAP estimator under the previous model specification $\boldsymbol{X} \sim \mathrm{N}_p(\boldsymbol{\theta}, \boldsymbol{I}_p)$ reduces to determining the value of $\boldsymbol{\theta}$ that minimizes

$$\frac{1}{2}\|\boldsymbol{X} - \boldsymbol{\theta}\|_2^2 + \lambda\|\boldsymbol{\theta}\|_2. \tag{3.3}$$

The resulting estimator is easily shown to be

$$\delta_{GL}(\boldsymbol{X}) = \left(1 - \frac{\lambda}{\|\boldsymbol{X}\|_2}\right)_+ \boldsymbol{X}, \tag{3.4}$$

an estimator that coincides with the solution to the canonical version of the grouped lasso problem involving a single group of parameters [46] and equals $\widehat{\theta}(X) = \mathrm{sign}(X)(|X| - \lambda)_+$ for the case where $p = 1$. Interestingly, (3.4) is also similar in form to, but distinct from, (1.2), a relationship that will be discussed in greater depth below.

In summary, the lasso and (canonical) grouped lasso estimators can be viewed as MAP estimators under the prior specification (3.2), where λ is treated as known. As summarized earlier in Section 2.2, [42] also proves that the positive part estimator (1.2) is the MAP estimator under a hierarchically specified prior that, for $c = p-2$, is given by (2.6) or, equivalently, (2.7). Evidently, these two classes of estimation problems are at least loosely related through the connections between the prior specifications (2.7), (2.8) and (3.2) described earlier. However, an interesting and noteworthy distinction is that the positive part estimator arises as the MAP estimator when the corresponding posterior is maximized *jointly* in both $\boldsymbol{\theta}$ and κ. The positive part estimator also has excellent risk properties that, over time, have proved to be challenging to dominate. In the prior formulation (2.9), hence (2.7), the influence of the prior on λ in (2.8) is arguably implicitly captured through the marginal prior on ψ. In (3.2), it is instead the influence of ψ that is being implicitly captured through consideration of the conditional prior $\boldsymbol{\theta}|\lambda$ implied by (2.8), with λ regarded as a known hyperparameter. Instead of treating λ as fixed, and in

the spirit of [42], one may consider the possibility of generalizing the class of lasso and grouped lasso estimators by maximizing the posterior distribution in both $\boldsymbol{\theta}$ and λ under a joint prior distribution given by (3.2) and a suitable class of prior distributions on λ. We expand on this idea in Section 4 using prior distributions motivated by the hierarchical structure (2.8). For simplicity, we continue to focus on the canonical (i.e., single group) version of the grouped lasso estimation problem; extensions to the more practical setting of multiple groups of parameters are possible and shall be considered elsewhere.

4. The WS estimator

Consider the problem of estimating $\boldsymbol{\theta}$ in the canonical setting $\boldsymbol{X} \sim \mathrm{N}_p(\boldsymbol{\theta}, \boldsymbol{I}_p)$. In view of the fact that (2.8) leads to (3.2) upon integrating out ψ, our starting point is the (possibly improper) generalized class of joint prior distributions $\pi(\boldsymbol{\theta}, \lambda|\alpha, \beta)$, which we define in the following hierarchical fashion:

$$\pi(\boldsymbol{\theta}|\lambda, \alpha, \beta) \propto \lambda^p \exp\{-\lambda\|\boldsymbol{\theta}\|_2\}\,,\ \pi(\lambda|\alpha, \beta) \propto \lambda^{-p} \exp\{-\alpha(\lambda-\beta)^2\}, \tag{4.1}$$

where $\alpha, \beta > 0$ are hyperparameters. Equivalently,

$$\pi(\boldsymbol{\theta}, \lambda|\alpha, \beta) \propto \exp\{-\lambda\|\boldsymbol{\theta}\|_2\} \exp\{-\alpha(\lambda-\beta)^2\}. \tag{4.2}$$

The prior on λ is evidently an improper modification of that given in (2.8), in which a location parameter β is introduced and a multiplicative λ^{-p} is included in order to offset the contribution λ^p in (3.2). This construction mimics the idea underlying the prior used by [42] to motivate (1.2) as a MAP estimator; in addition, as will soon be evident, this choice also turns out to be very convenient from a computational point view.

Remark 4.1. *In comparison with* (2.9)*, the (improper) prior on* ψ *that is induced by replacing* $\pi(\lambda|\varpi)$ *in* (2.8) *with* $\pi(\lambda|\alpha, \beta = 0)$ *in* (4.1) *is given by*

$$\pi(\psi|\alpha) \propto \psi^{-1/2}\psi^{p/2}\left(\frac{1}{1+\frac{\psi}{\alpha}}\right)\mathbb{1}_{[\psi>0]}.$$

Interestingly, this prior places increasingly higher weight on larger values of ψ*; in contrast, the prior* (2.7) *is unimodal, achieving its maximum value at* $\psi = (p-1)/3$.

Considering (4.2) as motivation for defining a new class of hierarchical penalty functions, we propose to compute the MAP estimator for $(\boldsymbol{\theta}, \lambda)$ through minimizing the objective function

$$G(\boldsymbol{\theta}, \lambda) \quad = \quad \frac{1}{2}\|\boldsymbol{X}-\boldsymbol{\theta}\|_2^2 + \lambda\|\boldsymbol{\theta}\|_2 + \alpha(\lambda-\beta)^2 \tag{4.3}$$

jointly in $\boldsymbol{\theta} \in \mathbb{R}^p$ and $\lambda > 0$, where $\alpha > 1/2$ and $\beta > 0$ are fixed. The resulting estimator for $\boldsymbol{\theta}$, hereafter referred to as the **W**hole vector **Shrinkage** (WS) estimator, takes the closed form

$$\delta^{(\alpha,\beta)}(\boldsymbol{X}) = w_{\alpha,\beta}(\|\boldsymbol{X}\|_2)\boldsymbol{X}, \tag{4.4}$$

where

$$w_{\alpha,\beta}(s) = \begin{cases} 0 & s \le \beta \\ \nu_\alpha\left(1-\frac{\beta}{s}\right) & \beta < s \le 2\alpha\beta \\ 1 & s > 2\alpha\beta \end{cases}$$

for $\nu_\alpha = 2\alpha/(2\alpha-1)$. Equivalently, we may write

$$w_{\alpha,\beta}(s) = \begin{cases} \nu_\alpha \left(1 - \frac{\beta}{s}\right)_+ & s \le 2\alpha\beta \\ 1 & s > 2\alpha\beta \end{cases},$$

demonstrating that (4.4) has the flavor of a range-modified positive part estimator. We now state this result as a theorem and provide its proof.

Theorem 4.2. *Let $p \ge 1$, $\alpha > 1/2$ and $\beta > 0$. Then, (4.3) is strictly convex for $(\boldsymbol{\theta}, \lambda) \in \mathbb{R}^p \times \mathbb{R}_+$; moreover, this function has a unique minimum at $\boldsymbol{\theta} = \delta^{(\alpha,\beta)}(\mathbf{X})$ and $\lambda = \lambda(\alpha, \beta, \mathbf{X})$, where*

$$\lambda(\alpha, \beta, \mathbf{X}) = \beta - \frac{\|\delta^{(\alpha,\beta)}(\mathbf{X})\|_2}{2\alpha}. \tag{4.5}$$

Proof. Throughout, we suppose that $\boldsymbol{X} \neq \mathbf{0}$, as this occurs with zero probability. For bookkeeping purposes, we also work with the following equivalently rescaled version of (4.3):

$$G_0(\boldsymbol{\theta}, \lambda_0) \quad = \quad \|\boldsymbol{X} - \boldsymbol{\theta}\|_2^2 + \lambda_0 \|\boldsymbol{\theta}\|_2 + \alpha_0(\lambda_0 - \beta_0)^2,$$

where $\lambda_0 = 2\lambda$, $\beta_0 = 2\beta$, and $\alpha_0 = \alpha/2$. The objective function $G_0(\boldsymbol{\theta}, \lambda_0)$ is clearly continuous and bounded below for $(\boldsymbol{\theta}, \lambda_0) \in \mathbb{R}^p \times \mathbb{R}_+$. We shall now establish the strict convexity of $G_0(\boldsymbol{\theta}, \lambda_0)$ on this same set under the restrictions $\alpha_0 > 1/4$ and $\beta_0 > 0$. Convexity is not only desirable from the perspective of minimization; it also ensures that the solution is continuous with respect to the regularization parameter. We may write $G_0(\boldsymbol{\theta}, \lambda_0)$ as $\|\boldsymbol{X}\|_2^2 + \|\boldsymbol{\theta}\|_2^2 - 2\langle \boldsymbol{X}, \boldsymbol{\theta}\rangle + \lambda_0\|\boldsymbol{\theta}\|_2 + \alpha_0(\lambda_0 - \beta_0)^2$, where $\langle \boldsymbol{X}, \boldsymbol{\theta}\rangle$ denotes the inner product. Since $-2\langle \boldsymbol{X}, \boldsymbol{\theta}\rangle$ is convex, the convexity of $G_0(\boldsymbol{\theta}, \lambda_0)$ is then determined by the convexity of

$$F(\boldsymbol{\theta}, \lambda_0) = \|\boldsymbol{\theta}\|_2^2 + \lambda_0\|\boldsymbol{\theta}\|_2 + \alpha_0(\lambda_0 - \beta_0)^2$$

for $\boldsymbol{\theta} \in \mathbb{R}^p$ and $\lambda_0 \in \mathbb{R}_+$ when $\alpha_0 > 1/4$ and $\beta_0 \in \mathbb{R}_+$.

Letting $\mathbf{z} = (\boldsymbol{\theta}', \lambda_0)'$, observe that $F(\boldsymbol{\theta}, \lambda_0) = g(h_1(\mathbf{z}), h_2(\mathbf{z}))$, where $h_1(\mathbf{z}) = \|\mathbf{A}_1\mathbf{z}\|_2 = \|\boldsymbol{\theta}\|_2$ for the $p \times (p+1)$ matrix $\mathbf{A}_1 = (\boldsymbol{I}_p\ \mathbf{0})$, $h_2(\mathbf{z}) = \mathbf{A}_2\,\mathbf{z} = \lambda_0$ for the $1 \times (p+1)$ vector $\mathbf{A}_2 = (\mathbf{0}', 1)'$ and

$$g(a, b) = \frac{1}{2}(a\ b)\begin{pmatrix} 2 & 1 \\ 1 & 2\alpha_0 \end{pmatrix}\begin{pmatrix} a \\ b \end{pmatrix} + \alpha_0(\beta_0^2 - 2b\beta_0), \quad (a, b) \in \mathbb{R}_+^2.$$

The function $g(a, b)$ is strictly convex for $(a, b) \in \mathbb{R}_+^2$ when $\alpha_0 > 1/4$. For such α_0, it is therefore monotonically increasing in each coordinate. The functions $h_i(\mathbf{z}), i = 1, 2$ are each convex for $\mathbf{z} \in \mathbb{R}^p \times \mathbb{R}_+$. Consequently, $F(\boldsymbol{\theta}, \lambda_0)$ is strictly convex for $(\boldsymbol{\theta}, \lambda_0) \in \mathbb{R}^p \times \mathbb{R}_+$, proving that $G_0(\boldsymbol{\theta}, \lambda_0)$ is strictly convex there.

An important consequence of these results is that $G_0(\boldsymbol{\theta}, \lambda_0)$ has a unique solution provided that $\alpha_0 > 1/4$. To deduce the form of this solution, we must find $(\boldsymbol{\theta}^*, \lambda_0^*)$ such that $\mathbf{0} \in \partial G_0(\boldsymbol{\theta}^*, \lambda_0^*)$, where ∂G_0 denotes the subdifferential of $G_0(\boldsymbol{\theta}, \lambda_0)$. We prove below that the solution exists in closed form and equals that specified in the statement of the theorem.

Since $G_0(\boldsymbol{\theta}, \lambda_0)$ is differentiable in λ_0, we find that

$$\lambda_0^* = \max\left\{\beta_0 - \frac{\|\boldsymbol{\theta}^*\|_2}{2\alpha_0}, 0\right\} \tag{4.6}$$

regardless of $\boldsymbol{\theta}^*$. The solution $\boldsymbol{\theta}^*$ (i.e., as a function of $\boldsymbol{X}$) can now be determined by considering the possible values of $\|\boldsymbol{\theta}^*\|_2$.

Suppose first that $\|\boldsymbol{\theta}^*\|_2 = 0$; then, $\boldsymbol{\theta}^* = \mathbf{0}$ and, from (4.6), $\lambda_0^* = \beta_0 > 0$. For $\boldsymbol{\theta}^* = \mathbf{0}$ to be the unique solution, $G_0(\boldsymbol{\theta}, \lambda_0)$ further implies $G_0(\mathbf{0}, \beta_0) < G_0(\boldsymbol{\theta}, \beta_0)$ for all $\boldsymbol{\theta} \neq \mathbf{0}$; equivalently, for $\boldsymbol{\theta}^* = \mathbf{0}$, $\boldsymbol{X}$ must satisfy

$$\|\boldsymbol{X}\|_2^2 < \|\boldsymbol{X} - \boldsymbol{\theta}\|_2^2 + \beta_0 \|\boldsymbol{\theta}\|_2$$

for all $\boldsymbol{\theta} \neq \mathbf{0}$. Since

$$\|\boldsymbol{X} - \boldsymbol{\theta}\|_2^2 + \beta_0 \|\boldsymbol{\theta}\|_2 = \|\boldsymbol{X}\|_2^2 - 2\langle \boldsymbol{\theta}, \boldsymbol{X} \rangle + \|\boldsymbol{\theta}\|_2^2 + \beta_0 \|\boldsymbol{\theta}\|_2,$$

this follows if

$$\|\boldsymbol{\theta}\|_2^2 + \beta_0 \|\boldsymbol{\theta}\|_2 > 2\langle \boldsymbol{\theta}, \boldsymbol{X} \rangle.$$

By Cauchy-Schwarz,

$$|\langle \boldsymbol{\theta}, \boldsymbol{X} \rangle| \leq \|\boldsymbol{\theta}\|_2 \|\boldsymbol{X}\|_2,$$

where equality holds if $\boldsymbol{\theta} = \boldsymbol{X}$. Thus, we require

$$\|\boldsymbol{\theta}\|_2^2 + \beta_0 \|\boldsymbol{\theta}\|_2 > 2\|\boldsymbol{\theta}\|_2 \|\boldsymbol{X}\|_2,$$

leading to the inequality

$$\|\boldsymbol{\theta}\|_2 + \beta_0 > 2\|\boldsymbol{X}\|_2.$$

Since this must be satisfied for all $\boldsymbol{\theta} \neq \mathbf{0}$, it follows that $\boldsymbol{\theta}^* = \mathbf{0}$ is the solution when $\|\boldsymbol{X}\|_2 \leq \beta_0/2$.

Now, assume that $\boldsymbol{\theta}^* \neq 0$ and hence that $\|\boldsymbol{X}\|_2 > \beta_0/2$. Suppose first that $0 < \|\boldsymbol{\theta}^*\|_2 \leq 2\alpha_0\beta_0$. Using the definition of the subdifferential, it follows immediately that $\boldsymbol{\theta}^*$ must satisfy

$$\boldsymbol{X} = \left(1 + \frac{\beta_0}{2\|\boldsymbol{\theta}^*\|_2} - \frac{1}{4\alpha_0}\right) \boldsymbol{\theta}^*. \tag{4.7}$$

By (4.7),

$$\|\boldsymbol{X}\|_2 = \left(1 + \frac{\beta_0}{2\|\boldsymbol{\theta}^*\|_2} - \frac{1}{4\alpha_0}\right) \|\boldsymbol{\theta}^*\|_2 = \frac{4\alpha_0 - 1}{4\alpha_0} \|\boldsymbol{\theta}^*\|_2 + \frac{\beta_0}{2};$$

upon rearranging terms, we find

$$\|\boldsymbol{\theta}^*\|_2 = \frac{4\alpha_0}{4\alpha_0 - 1} \left(\|\boldsymbol{X}\|_2 - \frac{\beta_0}{2}\right),$$

where $\|\boldsymbol{\theta}^*\|_2 > 0$ due to the fact that $\|\boldsymbol{X}\|_2 > \beta_0/2$. Substituting this expression for $\|\boldsymbol{\theta}^*\|_2$ in (4.7) and solving for $\boldsymbol{\theta}^*$ yields

$$\boldsymbol{\theta}^* = \frac{4\alpha_0}{4\alpha_0 - 1} \left(\frac{\|\boldsymbol{X}\|_2 - \beta_0/2}{\|\boldsymbol{X}\|_2}\right)_+ \boldsymbol{X},$$

where it is noted that this reduces to $\boldsymbol{X}$ for $\|\boldsymbol{\theta}^*\|_2 = 2\alpha_0\beta_0$. Finally, assume $\|\boldsymbol{\theta}\|_2 > 2\alpha_0\beta_0$. Then, from (4.6), we must have $\lambda^* = 0$, implying that $\boldsymbol{\theta}^*$ solves $\|\boldsymbol{X} - \boldsymbol{\theta}\|^2 + \alpha_0\beta_0$, that is, $\boldsymbol{\theta}^* = \boldsymbol{X}$.

Combining the cases outlined above, and using the facts that $\lambda_0 = 2\lambda$, $\beta_0 = 2\beta$, and $\alpha_0 = \alpha/2$, it follows that $\boldsymbol{\theta}^* = \delta^{(\alpha,\beta)}(\boldsymbol{X})$ in (4.4) and λ^* given in (4.5) uniquely minimize (4.3). □

Remark 4.3. *The optimization of* (4.3) *for* $\alpha = 0$ *may at first glance appear to correspond to the minimization problem* (3.1)*, in which* λ *is considered a fixed constant and optimization takes place over* $\boldsymbol{\theta}$ *only. However, recalling that* (4.3) *is optimized jointly in* $(\boldsymbol{\theta}, \lambda)$*, the correct correspondence is in fact obtained by letting* $\alpha \to \infty$*. For example, with* $p = 1$*, it is easy to see that* (4.4) *reduces to the familiar soft-thresholding estimator* $sign(X)(|X| - \beta)_+$ *as* $\alpha \to \infty$*, with* β *replacing* λ *as the penalty parameter. This is perfectly sensible when viewed from a Bayesian perspective: as* $\alpha \to \infty$*, the prior probability mass on* λ *becomes increasingly concentrated at* $\lambda = \beta$*.*

Some interesting special cases of the estimator (4.4) arise when considering specific values of α, β and p. For example, letting $\alpha \to \infty$, we obtain (for $\beta > 0$)

$$\delta^{(\beta)}(\boldsymbol{X}) = \left(1 - \frac{\beta}{\|\boldsymbol{X}\|_2}\right)_+ \boldsymbol{X}; \tag{4.8}$$

upon setting $\beta = \lambda$, we evidently recover (3.4); subsequently setting $\lambda = \sqrt{p-2}$, one then obtains an obvious modification of (1.2) for the case where $c = p - 2$:

$$\delta^*_{PP}(\boldsymbol{X}) = \left(1 - \frac{\sqrt{p-2}}{\|\boldsymbol{X}\|_2}\right)_+ \boldsymbol{X}. \tag{4.9}$$

Further specifications of α and β will be considered in later sections. In the special case $p = 1$, the estimator (4.4) reduces to

$$\delta^M(X) = \begin{cases} 0 & \text{if } |X| \le \beta \\ \frac{2\alpha}{2\alpha - 1}(X - \text{sign}(X)\beta) & \text{if } \beta < |X| \le 2\alpha\beta \\ X & \text{if } |X| > 2\alpha\beta \end{cases}. \tag{4.10}$$

As shown in [35], (4.10) is also the solution to the penalized minimization problem

$$\frac{1}{2}(X - \theta)^2 + \rho(\theta; \alpha, \beta),$$

where $\beta > 0$, $\alpha > 1/2$ and

$$\rho(t; \alpha, \beta) = \beta \int_0^{|t|} (1 - \frac{z}{2\alpha\beta})_+ dz, \quad t \in \mathbb{R}.$$

This optimization problem is the univariate equivalent of the penalized likelihood estimation problem considered in [47, 48], who refers to $\rho(t; \alpha, \beta)$ as the *minimax concave penalty* (MCP). It follows that (4.10) is equivalent to the univariate MCP thresholding operator; consequently, (4.4) may be regarded as a generalization of this operator for thresholding a vector of parameters. [47, 48] shows that the lasso, SCAD and MCP belong to a family of quadratic spline penalties with certain sparsity and continuity properties. MCP turns out to be the simplest penalty that results in an estimator that is nearly unbiased, sparse and continuous. As demonstrated above, MCP also has an interesting Bayesian motivation under a hierarchical modeling strategy. Simulation evidence for the advantages of MCP over other penalties can be found in [47, 48] and [31]. [35] undertakes a more detailed study of the connections between MCP and the hierarchically penalized estimator for the case of $p = 1$, as well as compares this estimator to several others through consideration of frequentist and Bayes risks.

In addition to the various special cases outlined above, one may also consider an elastic-net-type extension of (4.3) [cf. 49]. In particular, consider minimizing the extended objective function

$$\frac{1}{2}\|\boldsymbol{X}-\boldsymbol{\theta}\|^2+\lambda_1\|\boldsymbol{\theta}\|_2+\lambda_2\|\boldsymbol{\theta}\|_2^2+\alpha(\lambda_1-\beta)^2 \tag{4.11}$$

jointly in $\boldsymbol{\theta}\in\mathbb{R}^p$ and $\lambda_1,\lambda_2\in\mathbb{R}_+$ for $\beta>0$ and $\alpha>1/(2[1+2\lambda_2])$. The resulting estimator takes the closed form

$$\delta^{(\alpha,\beta,\lambda_2)}(\boldsymbol{X})=(1+2\lambda_2)^{-1}\delta^{(\alpha(1+2\lambda_2),\beta)}(\boldsymbol{X}),$$

where $\delta^{(\alpha,\beta)}(\cdot)$ is defined in (4.4). The proposed estimator can be considered as a hierarchical extension of that proposed in [49], who generalize the lasso procedure using a linear combination of ℓ_1 and ℓ_2 penalties.

5. Exact and unbiased estimators of the risks of the positive part and WS estimators

In Section 5.1, we utilize the techniques of [2] and [40] to derive expressions for the risk of the WS estimator (4.4) and also the positive part estimator in the case where $\boldsymbol{X}\sim N_p(\boldsymbol{\theta},\boldsymbol{I}_p)$; some associated numerical results and comparisons are provided in Section 5.2, where the impact of α is also explored. Following [39], we then develop an unbiased estimator of risk in Section 5.3, that is, a SURE formula [cf. 16]. Methods for selecting β, including a universal threshold criterion and a SURE-based criterion, are proposed in Sections 5.4.1 and 5.4.2. Finally, in Section 5.5, we compare the theoretical risk of the positive part estimator (1.2) ($c=p-2$) to the risks (theoretical or simulated, as appropriate) of various versions of the WS estimator (4.4).

5.1. Exact risks

A formula for the risk of (1.2) with $c>0$ can be obtained in a manner analogous to [2] and [40]; in particular, one can show that

$$E\left[\|\delta^c_{JS+}(\boldsymbol{X})-\boldsymbol{\theta}\|_2^2\right] \;=\; p+\exp\left\{-\frac{1}{2}\|\boldsymbol{\theta}\|_2^2\right\}\sum_{k=0}^{\infty}\frac{\|\boldsymbol{\theta}\|_2^{2k}}{2^k k!}J_p(k,c), \tag{5.1}$$

where $Y_k\sim\chi^2_{p+2k}$ and

$$J_p(k,c)=E\left[Y_k q^2(Y_k)-4kq(Y_k)-p+2k\right]$$

for $q(s)=(1-c/s)_+$. Using the facts that

$$\begin{aligned}J_p(k,c) \;=\;& (-p+2k)P\{Y_k\le c\}+E[Y_k\,\mathbb{1}_{[Y_k>c]}]\\ &+\;(c^2+4kc)E[Y_k^{-1}\,\mathbb{1}_{[Y_k>c]}]\\ &-\;(p+2(c+k))E[Y_k\,\mathbb{1}_{[Y_k>c]}]\end{aligned}$$

and

$$E[Y_k^r I\{a\le Y_k\le b\}]=\frac{\Gamma\left(\frac{p+2(r+k)}{2}\right)}{\Gamma\left(\frac{p+2k}{2}\right)}2^r P\{a\le\chi^2_{p+2(r+k)}\le b\} \tag{5.2}$$

for $0 \le a < b$ and any r such that $p + 2(r + k)$ is a positive integer, some easy calculations show

$$\begin{aligned} J_p(k,c) &= (2k-p)P\{\chi^2_{p+2k} \le c\} - (p+2(c+k))P\{\chi^2_{p+2k} > c\} \\ &+ (p+2k)P\{\chi^2_{p+2k+2} > c\} + \frac{c^2+4kc}{p+2k-2}P\{\chi^2_{p+2k-2} > c\}. \end{aligned}$$

[34] provides an alternative, finite series representation for the risk of this estimator in the case of unknown variance.

A formula for the risk of (4.4), assuming $\alpha > 1/2$ and $\beta > 0$ can be derived similarly to (5.1); in particular, straightforward calculations yield

$$E\left[\|\delta^{(\alpha,\beta)}(\boldsymbol{X}) - \boldsymbol{\theta}\|_2^2\right] = p + \exp\left\{-\frac{1}{2}\|\boldsymbol{\theta}\|_2^2\right\} \sum_{k=0}^{\infty} \frac{\|\boldsymbol{\theta}\|_2^{2k}}{2^k k!} H_p(k,\alpha,\beta), \tag{5.3}$$

where $Y_k \sim \chi^2_{p+2k}$ and

$$H_p(k,\alpha,\beta) = E\left[Y_k w^2_{\alpha,\beta}(Y_k) - 4k w_{\alpha,\beta}(Y_k) - p + 2k\right].$$

As in (5.1), the resulting risk formula involves an infinite series of weighted chi-square probabilities, now depending on α and β, that one may simplify using (5.2) and the fact that

$$\begin{aligned} H(k,\alpha,\beta) &= (-p+2k)P\left\{Y_k \le \beta^2\right\} + E\left[(Y_k - (p+2k))\ \mathbb{1}_{[Y_k > 4\alpha^2\beta^2]}\right] \\ &+ \left(\nu_\alpha^2\beta^2 - p + 2k(1-2\nu_\alpha)\right) P\left\{\beta^2 < Y_k \le 4\alpha^2\beta^2\right\} \\ &+ E\left[\left(\nu_\alpha^2 Y_k - 2\nu_\alpha^2\beta\, Y_k^{1/2} + 4k\nu_\alpha\beta\, Y_k^{-1/2}\right) \mathbb{1}_{[\beta^2 < Y_k \le 4\alpha^2\beta^2]}\right]. \end{aligned}$$

In both cases, the risk may be accurately computed numerically using a truncated series approximation, at least for moderate values of $\|\boldsymbol{\theta}\|_2$.

5.2. Some numerical insights

The risk formulas in the previous section are very similar in form but difficult to compare directly, even when done term-by-term. In Figures 1 and 2, we therefore provide plots of the risk (5.1) and the partially optimized risk

$$R(\alpha, \|\boldsymbol{\theta}\|_2) = \min_{\beta \ge 0} E\left[\|\delta^{(\alpha,\beta)}(\boldsymbol{X}) - \boldsymbol{\theta}\|_2^2\right]$$

in an effort to better understand (i) the role and impact of α and (ii) how, for $\alpha = \infty$, the optimized version of (4.8) (equivalently, (3.4)) compares to (1.2) ($c = p - 2$) in terms of risk. It is important to note that the optimal solution $\beta(\alpha)$ depends not only on α but also on $\|\boldsymbol{\theta}\|_2$; hence, these plots are not intended to depict the performance of a computable, data-based estimator of $\boldsymbol{\theta}$. Nevertheless, the results do lead to some useful and interesting insights, as discussed below.

We consider $\alpha = 1, 2, 4, 6, 10, \infty$, $\|\boldsymbol{\theta}\|_2 \in [0, 10]$ and $p \in \{3, 5, 7, 9\}$. In the context of the hierarchical prior specification (4.1), increasing the value of α corresponds to using an increasingly informative prior on λ, concentrating increasing amounts of mass near $\lambda = \beta$. For each p and $\|\boldsymbol{\theta}\|_2$, $R(\alpha, \|\boldsymbol{\theta}\|_2)$ is observed to decrease as

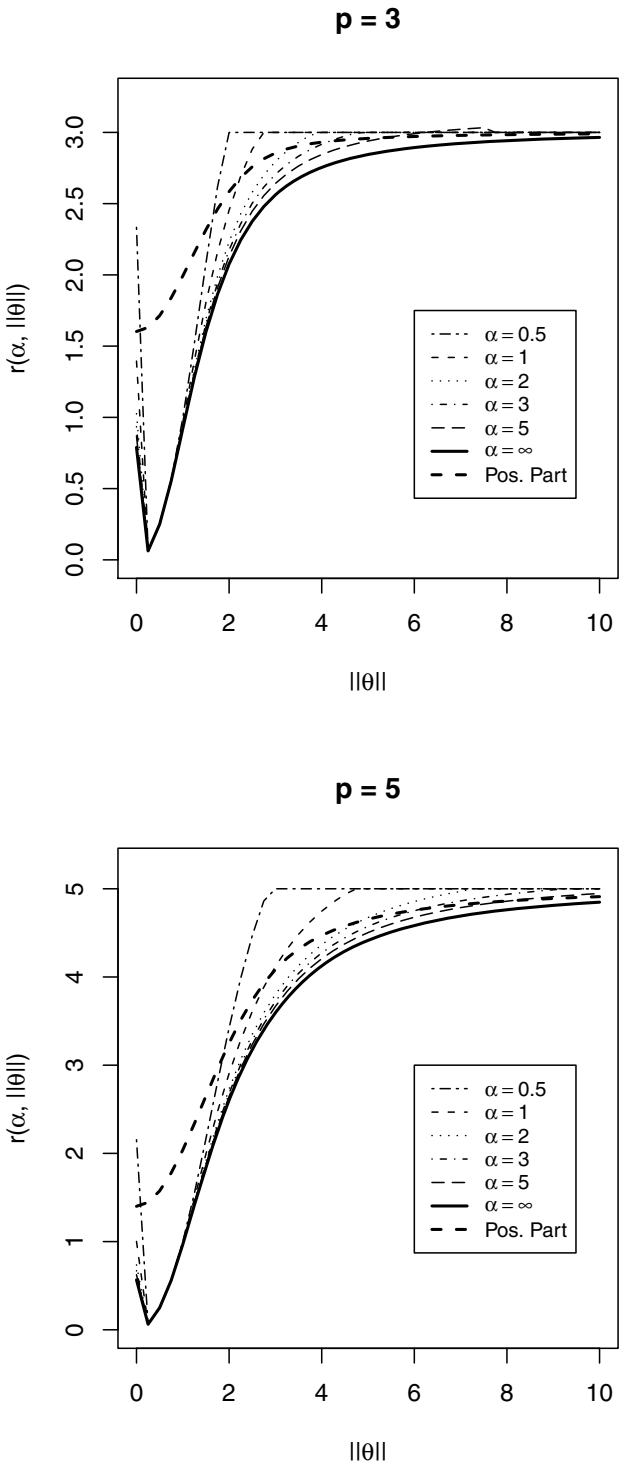


FIG 1. *Plots of $\beta-$optimized risk for selected values of α ($p = 3$ and $p = 5$). Solid dark line corresponds to $\alpha = \infty$, solid red line corresponds to (1.2) with $c = p - 2$.*

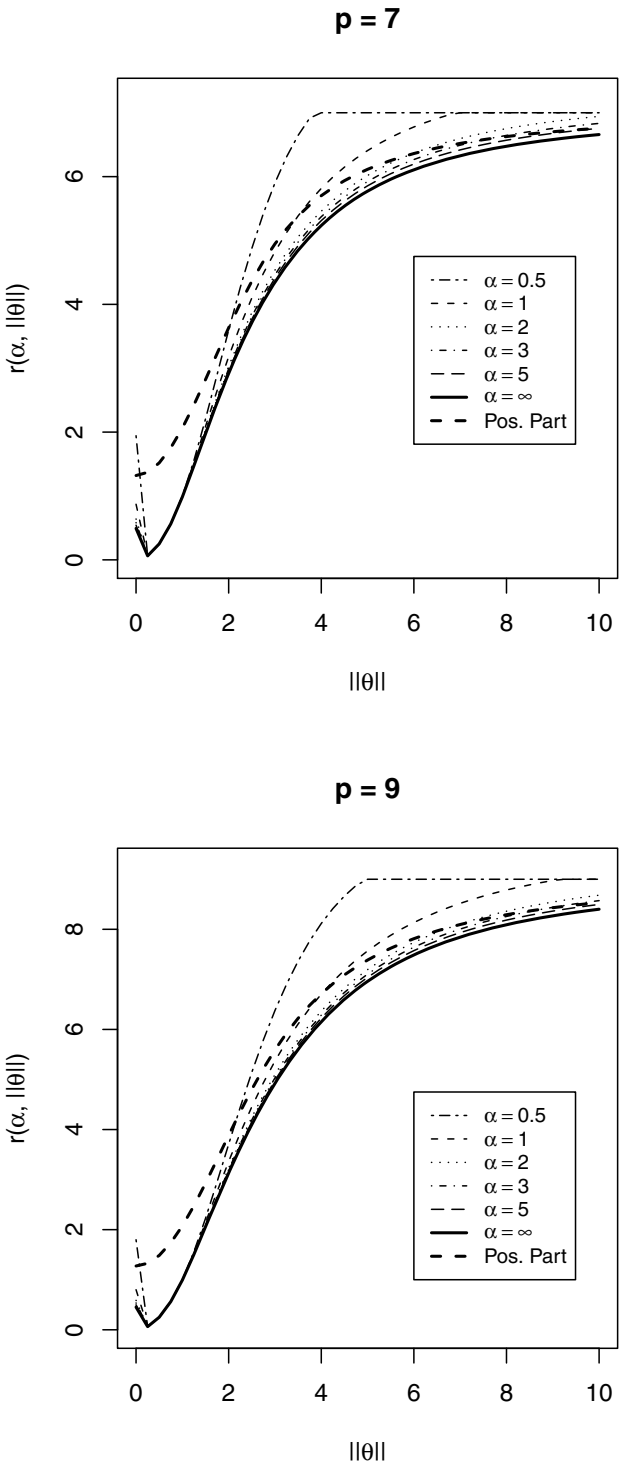


FIG 2. *Plots of $\beta-$optimized risk for selected values of α ($p = 9$ and $p = 13$). Solid dark line corresponds to $\alpha = \infty$, solid red line corresponds to (1.2) with $c = p - 2$.*

α increases, with the largest changes occurring between $\alpha = 1$ and $\alpha = 4$ and the impact of α on risk becoming increasingly smaller as α continues to increase. Interestingly, a precipitous drop in the optimized risk also occurs as one moves away from the origin, followed by a gradual rise to p.

Plots of $\beta(\alpha)$ by $\|\boldsymbol{\theta}\|_2$ (not shown) also demonstrate some interesting patterns. For example, for a given value of p, there exists a sharp peak in the values of $\beta(\alpha)$ for values of $\|\boldsymbol{\theta}\|_2$ near zero; in addition, the magnitude and location of this peak are approximately independent of α, consistent with the behavior of the optimized risk in Figures 1 and 2. We further observe that $\beta(\alpha)$ decays as $\|\boldsymbol{\theta}\|_2$ increases away from the value corresponding to the peak value of $\beta(\alpha)$. For $\alpha = \infty$, this decay occurs in a monotone fashion as $\|\boldsymbol{\theta}\|_2$ increases; such behavior is not unexpected, since increasing values of $\|\boldsymbol{\theta}\|_2$ should generate increasingly large values of $\|\boldsymbol{X}\|_2$, thereby decreasing the necessity for thresholding $\boldsymbol{X}$. However, for $\alpha < \infty$, there exists a finite $\|\boldsymbol{\theta}\|_2$ such that the optimal β eventually reverses its decline and starts to increase slowly. In view of (4.4), such behavior does not appear to be numerical artifact but instead a consequence of fixing α. In particular, the eventual increase in $\beta(\alpha)$ likely reflects the need to expand the region of shrinkage in (4.4) in order to continue to effectively manage the bias-variance tradeoff.

In general, the risk plots also suggest that it may be possible to achieve a substantial improvement on the positive part estimator (1.2) ($c = p - 2$), at least for smaller values of $\|\boldsymbol{\theta}\|_2$, through optimization of β. While $\alpha \to \infty$ appears to offer an optimal choice from the perspective of risk minimization, relatively small gains in risk are also observed beyond $\alpha = 4$, particularly as $\|\boldsymbol{\theta}\|_2$ and/ or p increases. Of course, the degree of improvement observed in Figures 1 and 2 must be treated with skepticism because the value of β used (i.e., $\beta(\alpha)$) depends directly on $\|\boldsymbol{\theta}\|_2$. In Section 5.5, we therefore evaluate the risk-based performance of (4.4) using $\alpha = 4$ and $\alpha = \infty$ and several possible choices of β, including a data-based method in which β is selected by minimizing an unbiased estimate of the risk (5.3) derived in Section 5.3 below.

Remark 5.1. *The consideration of a fixed, finite α is consistent with [47, 48], who explores the performance of the MCP penalty in this case and demonstrates improved variable selection performance in comparison to the lasso ($\alpha = \infty$) and SCAD penalty functions. The use of a data-based criterion for selecting α is also possible and worthy of consideration. However, the results thus far suggest that larger, fixed choices of α may result in estimators with similar risk profiles. As a result, the use of a purely risk-based criterion for selecting α may not be very informative. Further investigation into the role and importance of α is worthwhile and may dictate better choices of criterion functions for selecting this parameter.*

5.3. Unbiased estimators of risk

In general, when estimating $\boldsymbol{\theta}$ by some estimator $\delta(\boldsymbol{X})$ under a given loss function $L(\boldsymbol{\theta}, \delta(\boldsymbol{X}))$, classical decision theory advocates that such a decision rule should be used if it has suitable properties with respect to the frequentist risk $R(\boldsymbol{\theta}, \delta) = E[L(\boldsymbol{\theta}, \delta(\boldsymbol{X}))]$. However, having observed $\boldsymbol{X}$, instances arise in practice in which $\delta(\boldsymbol{X})$ is to be accompanied by an assessment of its loss $L(\boldsymbol{\theta}, \delta(\boldsymbol{X}))$. This loss function, which depends on $\boldsymbol{\theta}$, is not an observable quantity since $\boldsymbol{\theta}$ is unknown. A common approach to this assessment is to consider the estimation of $L(\boldsymbol{\theta}, \delta(\boldsymbol{X}))$ (equivalently, the corresponding risk function) by a so-called "loss estimator" $\Lambda_0(\boldsymbol{X})$ that depends

only on $\boldsymbol{X}$. There is now a sizable literature dealing with loss estimation; [21] provide a recent review.

In a classic article, [39] developed an unbiased estimator of the risk under the quadratic loss $||\delta(\boldsymbol{X})-\theta||_2^2$ for (nearly) arbitrary estimators of the form $\delta(\boldsymbol{X}) = \boldsymbol{X}+ g(\boldsymbol{X})$. In particular, under certain differentiability conditions that will be recalled below, he shows that

$$\Lambda_0(\boldsymbol{X}) = p + 2\,\mathrm{div} g(\boldsymbol{X}) + ||g(\boldsymbol{X})||_2^2$$

satisfies $E_\theta[\Lambda_0(\boldsymbol{X})] = E[||\delta(\boldsymbol{X}) - \boldsymbol{\theta}||_2^2] \equiv R(\boldsymbol{\theta}, \delta)$, where the expectation is taken under $\boldsymbol{\theta}$ and $\mathrm{div} g(X)$ stands for the divergence of $g(\boldsymbol{X})$. The resulting unbiased estimator of the risk, called the Stein's Unbiased Risk Estimate (SURE), was used primarily for the purpose of constructing estimates that improve on the MLE $\boldsymbol{X}$ in the case where $p \geq 3$ (i.e., to devise sufficient conditions such that $R(\boldsymbol{\theta}, \delta) \leq p$).

[16] demonstrated the importance and utility of SURE as a tool for threshold selection in the context of function estimation using wavelets, opening the door to a much wider range of possible applications. A formula for the SURE associated with (1.2) can be found in Cai and Zhou [12, Eqn. 5]. In this subsection, we shall derive an explicit expression of this SURE for the estimator $\delta^{(\alpha,\beta)}(\boldsymbol{X})$ in (4.4). In a subsequent section, we shall make use of this SURE expression to select the hyperparameter β for a given value of α. Recalling (4.4) and writing $\delta^{(\alpha,\beta)}(\boldsymbol{X}) = \boldsymbol{X} + g(\boldsymbol{X})$, we have

$$g(\boldsymbol{X}) = \begin{cases} -\boldsymbol{X} & \|\boldsymbol{X}\|_2 \leq \beta \\ \frac{1}{2\alpha-1}\boldsymbol{X} - \frac{2\alpha\beta}{2\alpha-1}\frac{\boldsymbol{X}}{\|\boldsymbol{X}\|_2} & \beta < \|\boldsymbol{X}\|_2 \leq 2\alpha\beta \\ 0 & \|\boldsymbol{X}\|_2 > 2\alpha\beta. \end{cases}$$

The function $g(\cdot)$ is a weakly differentiable function from $\mathbb{R}^p \to \mathbb{R}$; that is, one can show that there exist p functions $h_1(\cdot), \ldots, h_p(\cdot)$ that are locally integrable on $\mathbb{R}^p$ such that

$$\int_{\mathbb{R}^p} g(x)\,\frac{\partial \varphi}{\partial x_i}(x)\,dx = -\int_{\mathbb{R}^p} h_i(x)\,\varphi(x)\,dx, \quad i = 1, \ldots, p$$

for any infinitely differentiable function φ on $\mathbb{R}^p$ with compact support. The functions $h_i(\cdot)$ are defined to be the i-th partial weak derivatives of $g(\cdot)$. Using Stein's identity [39], we have,

$$SURE(\delta^{(\alpha,\beta)}(\boldsymbol{X})) = p + 2\,\mathrm{div} g(\boldsymbol{X}) + ||g(\boldsymbol{X})||_2^2. \tag{5.4}$$

A straightforward calculation shows

$$\begin{aligned} ||g(\boldsymbol{X})||^2 &= \sum_{i=1}^{p}\left[-X_i\,\mathbb{1}_{[\|\boldsymbol{X}\|_2\leq\beta]} + \left(\frac{X_i}{2\alpha-1} - \frac{2\alpha\beta}{2\alpha-1}\frac{X_i}{\|\boldsymbol{X}\|_2}\right)\mathbb{1}_{[\beta<\|\boldsymbol{X}\|_2\leq 2\alpha\beta]}\right]^2 \\ &= \|\boldsymbol{X}\|_2^2\left[-\mathbb{1}_{[\|\boldsymbol{X}\|_2\leq\beta]} + \left(\frac{1}{2\alpha-1} - \frac{2\alpha\beta}{2\alpha-1}\frac{1}{\|\boldsymbol{X}\|_2}\right)\mathbb{1}_{[\beta<\|\boldsymbol{X}\|_2\leq 2\alpha\beta]}\right]^2 \\ &= \|\boldsymbol{X}\|_2^2\left[\mathbb{1}_{[\|\boldsymbol{X}\|_2\leq\beta]} + \frac{(\|\boldsymbol{X}\|_2 - 2\alpha\beta)^2}{(2\alpha-1)^2\|\boldsymbol{X}\|_2^2}\,\mathbb{1}_{[\beta<\|X\|_2\leq 2\alpha\beta]}\right]. \end{aligned} \tag{5.5}$$

In addition, for $\mathbf{z} \in \mathbb{R}^p$, one can show that $\mathrm{div}(\mathbf{z}/\|\mathbf{z}\|_2) = (p-1)/\|\mathbf{z}\|_2$; therefore,

$$\begin{aligned} \mathrm{div} g(\boldsymbol{X}) &= -p\,\mathbb{1}_{[\|\boldsymbol{X}\|_2\leq\beta]} + \left(\frac{p}{2\alpha-1} - \frac{2\alpha\beta}{2\alpha-1}\frac{(p-1)}{\|\boldsymbol{X}\|_2}\right)\mathbb{1}_{[\beta<\|\boldsymbol{X}\|_2\leq 2\alpha\beta]} \\ &= -p\,\mathbb{1}_{[\|\boldsymbol{X}\|_2\leq\beta]} + \frac{p\|\boldsymbol{X}\|_2 - 2\alpha\beta(p-1)}{\|\boldsymbol{X}\|_2(2\alpha-1)}\,\mathbb{1}_{[\beta<\|\boldsymbol{X}\|_2\leq 2\alpha\beta]}. \end{aligned} \tag{5.6}$$

Substituting (5.5) and (5.6) into the SURE formula (5.4) and collecting terms, it follows that

$$SURE(\delta^{(\alpha,\beta)}(\boldsymbol{X})) = p + (\|\boldsymbol{X}\|_2^2 - 2p)\,\mathbb{1}_{[\|\boldsymbol{X}\|_2 \le \beta]} + V_p(\alpha, \beta, \|X\|_2) \tag{5.7}$$

where

$$V_p(\alpha, \beta, \|\boldsymbol{X}\|_2) = \left[\frac{2p}{2\alpha - 1} + \frac{(\|\boldsymbol{X}\|_2 - 2\alpha\beta)^2}{(2\alpha-1)^2} - \frac{4\alpha\beta(p-1)}{(2\alpha-1)\|\boldsymbol{X}\|_2}\right] \mathbb{1}_{[\beta < \|\boldsymbol{X}\|_2 \le 2\alpha\beta]}.$$

For $\alpha = \infty$, (5.7) is identical, except that we replace $V_p(\alpha, \beta, \|\boldsymbol{X}\|_2)$ with its limit

$$V_p(\infty, \beta, \|\boldsymbol{X}\|_2) = \left[\beta^2 - \frac{2\beta(p-1)}{\|\boldsymbol{X}\|_2}\right] \mathbb{1}_{[\beta < \|\boldsymbol{X}\|_2]}.$$

5.4. Tuning parameter selection

5.4.1. A "universal threshold" criterion

Suppose $\boldsymbol{\theta} = \mathbf{0}$. Then, one may ask whether it is possible to select β in such a way that ensures that $P(\delta^{(\alpha,\beta)}(\boldsymbol{X}) = \mathbf{0}) \approx 1$. Such an idea underlies the development of the "universal threshold" originally developed for use in wavelet thresholding applications [15, 16]. In the present case, and assuming $\|\boldsymbol{\theta}\|_2 = 0$,

$$P\left(\delta^{(\alpha,\beta)}(\boldsymbol{X}) = \mathbf{0}\right) = P\left(\|\boldsymbol{X}\|_2 \le \beta\right) = 1 - P\left(\|\boldsymbol{X}\|_2^2 > \beta^2\right),$$

where $\|\boldsymbol{X}\|_2^2 \sim \chi_p^2$. An asymptotic approach to this problem is to determine a threshold ω_p such that $\omega_p = \beta_p^2$ and

$$P\left(\chi_p^2 > \omega_p\right) \to 0$$

as $p \to \infty$. Notably, the choice $\omega_p = p$ (equivalently, $\beta_p = \sqrt{p}$) does not work since one can easily prove directly that

$$\lim_{p\to\infty} P\left(\chi_p^2 > p\right) = \frac{1}{2}.$$

For any $x > 0$, recall that

$$P\left(\chi_p^2 > x\right) = \frac{\Gamma\left(\frac{p}{2}, \frac{x}{2}\right)}{\Gamma\left(\frac{p}{2}\right)},$$

where $\Gamma(v, z) = \int_z^\infty s^{v-1}e^{-s}ds$. Using inequalities devised in [33] for the incomplete gamma function, the following bounds are easily obtained:

$$\frac{\left(\frac{\omega_p}{2}\right)^{\frac{p}{2}-1} e^{-\omega_p/2}}{\Gamma\left(\frac{p}{2}\right)} < P\left(\chi_p^2 > \omega_p\right) < b\,\frac{\left(\frac{\omega_p}{2}\right)^{\frac{p}{2}-1} e^{-\omega_p/2}}{\Gamma\left(\frac{p}{2}\right)}$$

for $p > 2$, $b > 1$ and $\omega_p > b(p-2)/(b-1)$. Recalling that $\omega_p = p$ gives $P\left(\chi_p^2 > \omega_p\right) \to 1/2$, consider $\omega_p = hp$ for $h > 1$. Substituting this choice into the lower and upper bounds given above, it is easily shown that both converge to zero. It follows that $P\left(\chi_p^2 > \omega_p\right) \to 0$ as $p \to \infty$ if one takes $\omega_p = hp$ and we can select $h > 1$ and $b > 1$

such that $hp > b(p-2)/(b-1)$. When $p > 2$, these conditions are satisfied for any $h > 1$ provided that $b \geq h/(h-1)$.

The above arguments establish that the smallest threshold ω_p leading to a unit probability of thresholding for $\boldsymbol{\theta} = \mathbf{0}$ as $p \to \infty$ satisfies $\omega_p = hp$ (equivalently, $\beta_p = \sqrt{hp}$) for $h > 1$. An interesting choice of h is simply $\nu_\alpha = 2\alpha/(2\alpha - 1)$ for any finite $\alpha > 1/2$.

In addition to being an interesting result in its own right, the above calculations lead to a new and fascinating justification for the class of positive part estimators (1.2). Consider, in particular, estimating β using

$$\widehat{\beta}(\boldsymbol{X}) = \frac{hp}{\|\boldsymbol{X}\|_2}, \tag{5.8}$$

for some fixed $h > 0$. Observe that

$$P\left(\delta^{(\alpha,\widehat{\beta}(\boldsymbol{X}))}(\boldsymbol{X}) = \mathbf{0}\right) = P\left(\|\boldsymbol{X}\|_2 \leq \widehat{\beta}(\boldsymbol{X})\right) = 1 - P\left(\|\boldsymbol{X}\|_2^2 > hp\right);$$

provided $h > 1$, we see from earlier calculations that this probability also converges to one as $p \to \infty$ when $\boldsymbol{\theta} = \mathbf{0}$. Notice, however, that

$$\delta^{(\alpha,\widehat{\beta}(\boldsymbol{X}))}(\boldsymbol{X}) = w_{\alpha,\widehat{\beta}(\boldsymbol{X})}(\|\boldsymbol{X}\|_2)\boldsymbol{X}, \tag{5.9}$$

where

$$w_{\alpha,\widehat{\beta}(\boldsymbol{X})}(\|\boldsymbol{X}\|_2) = \begin{cases} 0 & \|\boldsymbol{X}\|_2^2 \leq hp \\ \nu_\alpha\left(1 - \frac{hp}{\|\boldsymbol{X}\|_2^2}\right) & hp < \|\boldsymbol{X}\|_2^2 \leq 2\alpha hp \\ 1 & \|\boldsymbol{X}\|_2^2 > 2\alpha hp \end{cases}$$

for ν_α defined as before. For $\alpha = \infty$, this estimator reduces to (4.8) with $\beta = \widehat{\beta}(\boldsymbol{X})$; equivalently, we obtain (1.2) for $c = hp$. In other words, for certain choices of c, (1.2) can be interpreted as the solution to (3.2) with $\lambda = \widehat{\beta}(\boldsymbol{X})$, an "optimal" choice of thresholding parameter under sparsity of the mean vector. Interestingly, the estimator (1.2) for $c = p - 2$ corresponds to selecting $h = 1 - 2p^{-1} < 1$ in (5.9), an estimator that is arguably suboptimal in the sense just described.

5.4.2. SURE-based selection

Fixing α and considering (5.7) as a function of β leads to an empirical procedure for selecting β, in particular taken to be a minimizer of (5.7). Evidently, (5.7) is a discontinuous function of β for $\alpha < \infty$; moreover, a unique minimizer typically does not exist. In the spirit of [16], we therefore propose to select the smallest value of β minimizing (5.7) for $\beta \in [0, \sqrt{hp}]$, where the upper bound corresponds to the threshold calculation of Section 5.4.1. Numerically, the optimal β, say $\widehat{\beta}_{SURE(\alpha)}$, is then easily determined, at least approximately, upon evaluating (5.7) over a fine grid spanning this interval.

5.5. Simulation results

In this section, we consider the risk-based performance of (4.4) for the specific choices $\alpha = 4$ and $\alpha = \infty$. We further considered several choices for β, including

- $\beta = \sqrt{hp}$ for $h = 8/7$ (universal threshold, Section 5.4.1, using $h = \nu_4$)

- $\beta = \sqrt{p-2}$ (i.e., estimator (4.9))
- $\beta =$ (5.8) (i.e., estimator (5.9))
- $\beta = \widehat{\beta}_{SURE(\alpha)}$ (Section 5.4.2)

For each α, theoretical risk calculations are done for the first two choices of β, whereas simulated risks are obtained for the latter two selections. We note that the risk of (5.9) for $\alpha = \infty$ can also be computed using (5.1). We include for comparison the theoretical risk for the standard form of the positive part estimator (1.2) (i.e., $c = p - 2$).

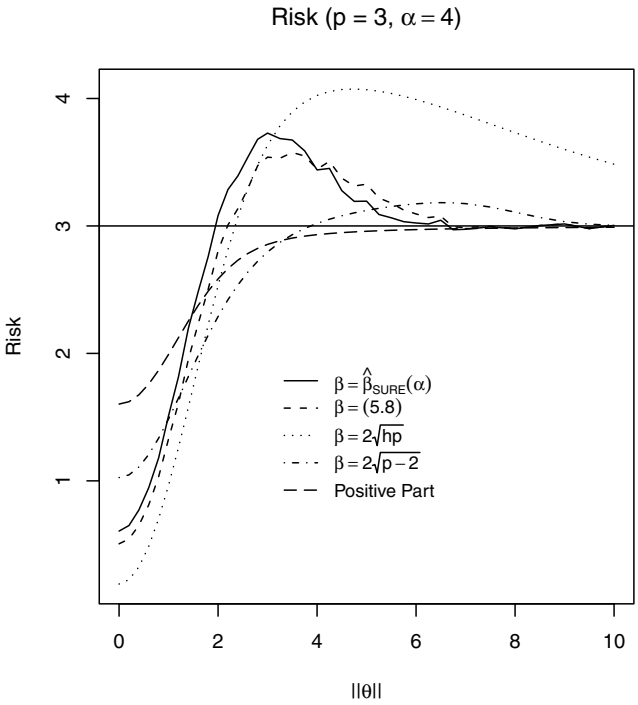
In general, Figures 3–6 suggest that the estimators using $\beta =$ (5.8) and $\beta = \widehat{\beta}_{SURE(\alpha)}$ perform similarly for $\alpha = 4$ and $\alpha = \infty$, particularly for $\|\boldsymbol{\theta}\|_2 \leq 4$. For larger $\|\boldsymbol{\theta}\|_2$, $\beta = \widehat{\beta}_{SURE(\alpha)}$ tends to result in somewhat lower risk. For both $\alpha = 4$ and $\alpha = \infty$, the selection $\beta = \sqrt{hp}$ also tends to yield the smallest risk for $\|\boldsymbol{\theta}\|_2 < 2$, seemingly consistent with its derivation; however, its performance soon degrades, thresholding $\boldsymbol{X}$ more often than necessary. The effect of selecting $\alpha < \infty$ is perhaps most clearly seen for the selection $\beta = \sqrt{p-2}$, where the risk remains under control for larger values of $\|\boldsymbol{\theta}\|_2$ when compared to setting $\alpha = \infty$.

Of significant interest here is the performance of all such estimators relative to (1.2) for $c = p - 2$. All choices of β lead to substantial reductions in risk for small values of $\|\boldsymbol{\theta}\|_2$, and those utilizing a data-based choice of β (i.e., $\beta =$ (5.8) and $\beta = \widehat{\beta}_{SURE(\alpha)}$) also perform similarly to (1.2) for large values of $\|\boldsymbol{\theta}\|_2$. However, an apparent difficulty remains in beating (1.2) for moderate values of $\|\boldsymbol{\theta}\|_2$. While the exact reasons for this difficulty are unknown, we conjecture that this may well be a consequence of the fact that (1.2) and (1.1) are (nearly) equivalent for values of $\|\boldsymbol{\theta}\|_2$ sufficiently far from the origin and that the choice $c = p - 2$ in (1.1) corresponds to the estimator having the smallest risk among all estimators of the form (1.1) [e.g. 26].

6. Discussion

Motivated by [40] and [42], and through consideration of hierarchical prior constructions derived from scale mixtures of normal distribution, we have attempted to point out and exploit interesting connections between the (arguably) original version of the large p - small n problem that permeates much of Bill Strawderman's work and today's ever-growing literature on solving such problems from the perspective of penalized likelihood. We have also demonstrated, in particular, that it is possible to significantly improve on the venerable positive part estimator using a class of estimators derived from (4.4). The theoretical risk plots in Figures 1 and 2 suggest that an (unrealistically) accurate choice of β can produce an estimator that beats the positive part estimator over a wide range of $\boldsymbol{\theta}$; the simulated risks in Figures 3–6 demonstrate that significant gains remain possible using a data-based choice of β, at least for $\boldsymbol{\theta}$ closer to the origin, and with increasingly little penalty as p grows.

Further work in the general area of hierarchical prior design and methods of hyperparameter selection in the context of penalized likelihood problems seems warranted. As one example: in deriving (4.4), the form of the prior, while loosely similar to that considered in (2.8), was specifically selected to make closed form computations feasible. The fact that this conveniently chosen hierarchical prior leads to the MCP thresholding function of [47, 48] in the case where $p = 1$ is both fascinating and unexpected; that the MCP thresholding function has also



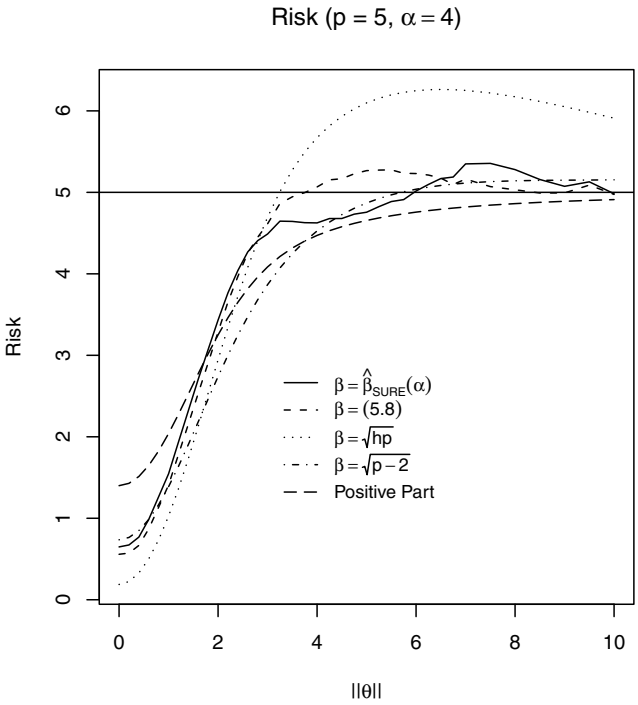
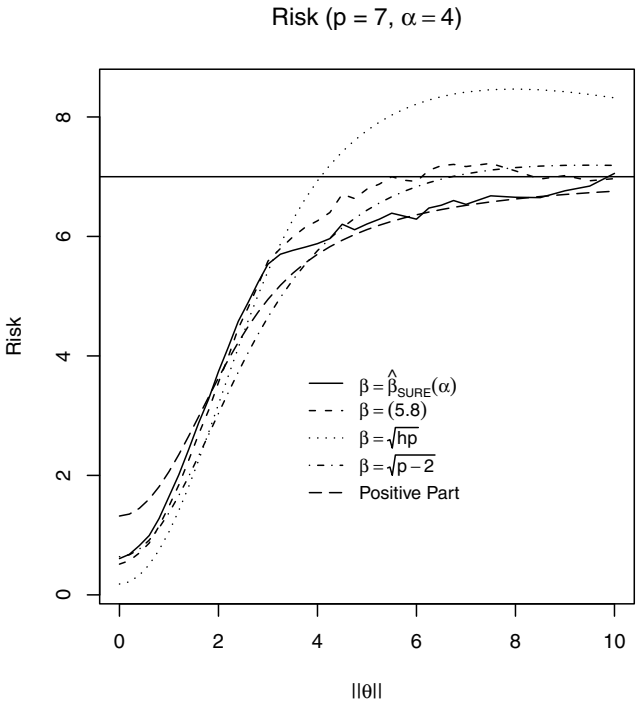


FIG 3. *Risk plots for $\alpha = 4$ ($p = 3$ and $p = 5$). The risks for $\beta = (5.8)$ and $\beta = \widehat{\beta}_{SURE(\alpha)}$ are simulated using 2500 randomly generated datasets for each $\|\boldsymbol{\theta}\|_2$; all others are computed using the theoretical risk formulas from Section 5.1.*



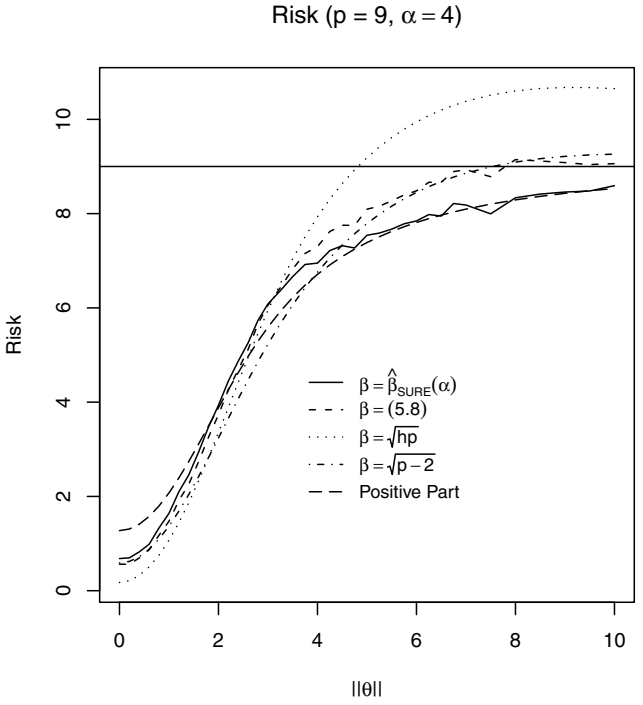
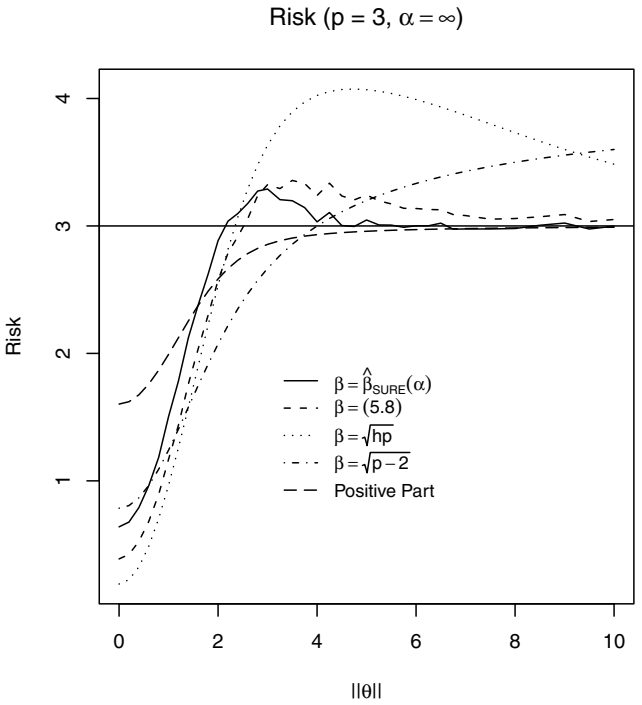


FIG 4. *Risk plots for $\alpha = 4$ ($p = 7$ and $p = 9$). The risks for $\beta = (5.8)$ and $\beta = \widehat{\beta}_{SURE(\alpha)}$ are simulated using 2500 randomly generated datasets for each $\|\boldsymbol{\theta}\|_2$; all others are computed using the theoretical risk formulas from Section 5.1.*



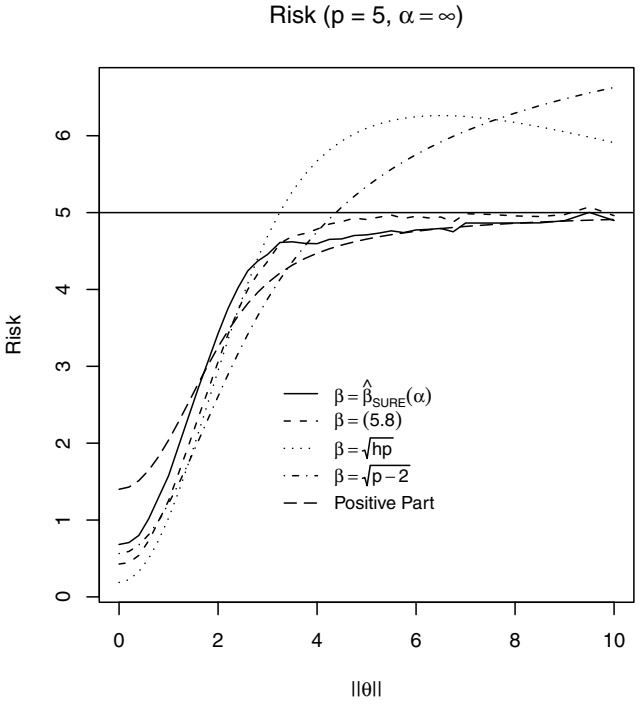
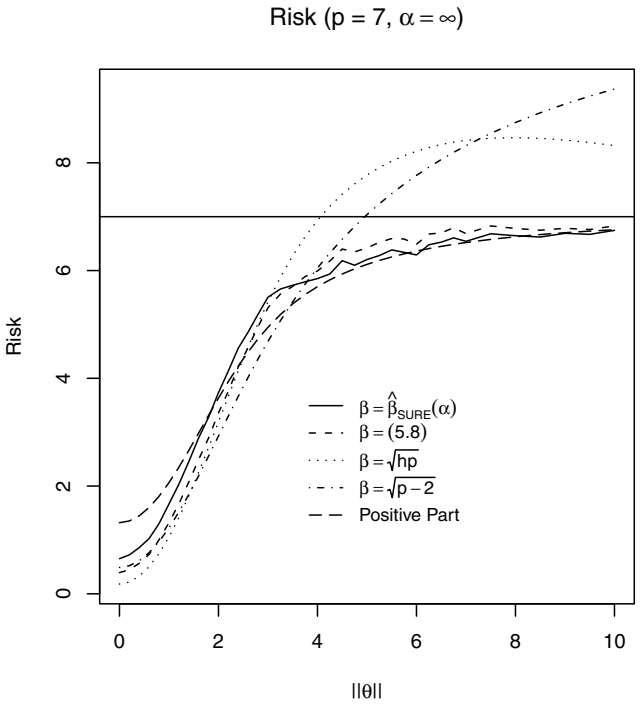


FIG 5. *Risk plots for $\alpha = \infty$ ($p = 3$ and $p = 5$). The risks for $\beta = (5.8)$ and $\beta = \widehat{\beta}_{SURE(\alpha)}$ are simulated using 2500 randomly generated datasets for each $\|\boldsymbol{\theta}\|_2$; all others are computed using the theoretical risk formulas from Section 5.1.*



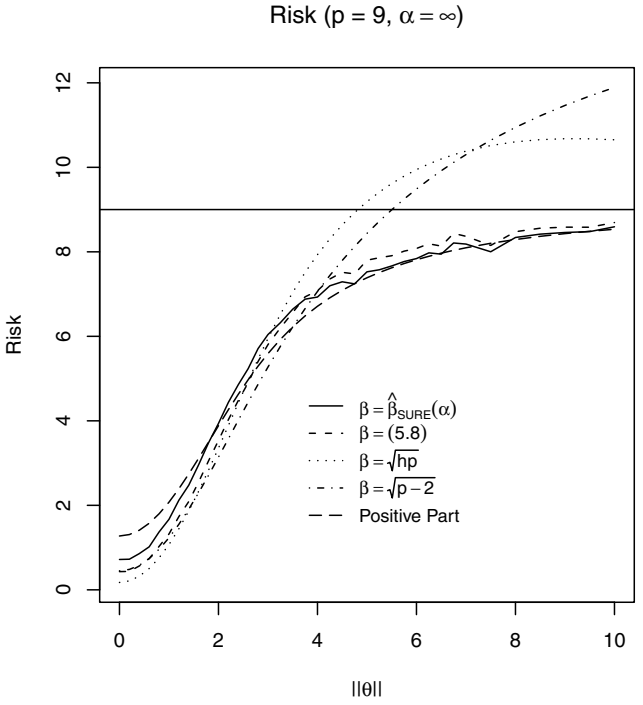


FIG 6. *Risk plots for $\alpha = \infty$ ($p = 7$ and $p = 9$). The risks for $\beta = (5.8)$ and $\beta = \widehat{\beta}_{SURE(\alpha)}$ are simulated using 2500 randomly generated datasets for each $\|\boldsymbol{\theta}\|_2$; all others are computed using the theoretical risk formulas from Section 5.1.*

proved to be a very effective alternative in comparison with both the lasso and SCAD penalties in the regression context suggests, at the very least, a continued potential for numerous fruitful avenues of further investigation. As one specific example: the direct generalization of (4.4) to the more practical setting of multiple groups of parameters has already generated some new and interesting estimators with properties that we intend to investigate and report elsewhere.

References

[1] Baranchik, A. J. (1964). Multiple regression and estimation of the mean of a multivariate normal distribution Technical Report No. 51, Department of Statistics, Stanford University.

[2] Baranchik, A. J. (1970). A family of minimax estimators of the mean of a multivariate normal distribution. *Annals of Mathematical Statistics* **41** 642–645.

[3] Berger, J. O. (1976). Admissible minimax estimation of a multivariate normal mean with arbitrary quadratic loss. *Annals of Statistics* **4** 223–226.

[4] Berger, J. O. (1980). A robust generalized Bayes estimator and confidence region for a multivariate normal mean. *Annals of Statistics* **8** 716–761.

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

[6] Berger, J. O., Strawderman, W. E. and Tang, D. (2005). Posterior propriety and admissibility of hyperpriors in normal hierarchical models. *Annals of Statistics* **33** 606–646.

[7] Bishop, C. M. and Tipping, M. E. (2000). Variational relevance vector machines. In *UAI '00: Proceedings of the 16th Conference on Uncertainty in Artificial Intelligence* (C. Boutilier and M. Goldszmidt, eds.) 46–53. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

[8] Bock, M. E. (1988). Shrinkage estimator: pseudo-Bayes estimators for normal mean vectors. In *Statistical Decision Theory and Related Topics 4*, (S. S. Gupta and J. . Berger, eds.) **1** 281–298. Springer-Verlag, New York.

[9] Brandwein, A. C. and Strawderman, W. E. (1990). Stein Estimation: the spherically symmetric case. *Statistical Science* **5** 356–369.

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

[11] Brown, L. D. (1988). The differential inequality of a statistical estimation problem. In *Statistical Decision Theory and Related Topics 4*, (S. S. Gupta and J. Berger, eds.) **1** 299–324. Springer-Verlag, New York.

[12] Cai, T. T. and Zhou, H. H. (2009). A data-driven block thresholding approach to wavelet estimation. *Ann. Statist.* **37** 569–595.

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

[14] Casella, G. and Strawderman, W. E. (1981). Estimating a bounded normal mean. *Annals of Statistics* **9** 870-878.

[15] Donoho, D. and Johnstone, I. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika* **81** 425–455.

[16] Donoho, D. and Johnstone, I. (1995). Adapting to unknown smoothness via wavelet shrinkage. *Journal of the American Statistical Association* **90** 1200–1224.

[17] Efron, B. and Morris, C. (1973). Stein's estimation rule and its competitors–an empirical Bayes approach. *Journal of the American Statistical Association* **68** 117–130.

[18] Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* 456.

[19] Figueiredo, M. A. T. (2003). Adaptive sparseness for supervised learning. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* **25** 1150–1159.

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

[21] Fourdrinier, D. and Wells, M. T. (2010). On loss estimation. *Statistical Science* (to appear).

[22] Ghosh, M. (1992). Hierarchical and empirical Bayes multivariate estimation. In *Current Issues in Statistical Inference: Essays in Honor of D. Basu. IMS Lecture Notes Monogr. Ser.* **17** 151–177. Institute of Mathematical Statistics, Hayward, CA.

[23] 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. *Communications in Statistics - Theory and Methods* **37** 972–985.

[24] Griffin, J. E. and Brown, P. J. (2005). Alternative prior distributions for variable selection with very many more variables than observations Technical Report, Department of Statistics, University of Warwick.

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

[26] Gupta, A. K. and Peña, E. A. (1991). A simple motivation for James-Stein estimators. *Statistics and Probability Letters* **12** 337–340.

[27] James, W. and Stein, C. (1961). Estimation with quadratic loss. In *Proc. 4th Berkeley Sympos. Math. Statist. and Prob., Vol. I* 361–379. Univ. California Press, Berkeley.

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

[29] Johnstone, I. M. and Silverman, B. W. (2005). Empirical Bayes selection of wavelet thresholds. *Annals of Statistics* **33** 1700–1752.

[30] Marchand, E. and Perron, F. (2001). Improving on the MLE of a bounded normal mean. *Annals of Statistics* **29** 1078–1093.

[31] Mazumder, R., Friedman, J. and Hastie, T. (2009). SparseNet: coordinate descent with non-convex penalties Technical Report, Department of Statistics, Stanford University.

[32] Moulin, P. and Liu, J. (1999). Analysis of multiresolution image denoising schemes using generalized Gaussian and complexity priors. *IEEE Transactions on Information Theory* **45** 909–919.

[33] Natalini, P. and Palumbo, B. (2000). Inequalities for the incomplete gamma function. *Mathematical Inequalities & Applications* **3** 69–77.

[34] Robert, C. (1988). An explicit formula for the risk of the positive-part James-Stein estimator. *The Canadian Journal of Statistics* **16** 161–168.

[35] Schifano, E. D. (2010). Topics in Penalized Estimation PhD Dissertation, Cornell University, Department of Statistical Science.

[36] Shao, P. Y. and Strawderman, W. E. (1994). Improving on the James-Stein positive-part estimator. *Annals of Statistics* **22** 1517–1538.
[37] Shevade, S. K. and Keerthy, S. S. (2003). A simple and efficient algorithm for gene selection using sparse logistic regression. *Bioinformatics* **19** 2246-2253.
[38] Stein, C. (1956). Inadmissibility of the usual estimator for the mean of a multivariate distribution. In *Proc. 1st Berkeley Sympos. Math. Statist. and Prob., Vol. I* 197-206. Univ. California Press, Berkeley.
[39] Stein, C. (1981). Estimation of the mean of multivariate normal distribution. *Annals of Statistics* **9** 1135–1151.
[40] Strawderman, W. E. (1971). Proper Bayes minimax estimators of the multivariate normal mean. *Annals of Mathematical Statistics* **42** 385–388.
[41] Strawderman, W. E. (1972). On the existence of proper Bayes minimax estimators of the mean of a multivariate normal distribution. In *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability (Univ. California, Berkeley, Calif., 1970/1971), Vol. I: Theory of Statistics* 51–55. Univ. California Press, Berkeley.
[42] Takada, Y. (1979). Steins positive part estimator and Bayes estimator. *Annals of the Institute of Statistical Mathematics* **31** 177–183.
[43] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B* 1 267–288.
[44] 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.
[45] Tipping, M. E. (2001). Sparse Bayesian learning and the relevance vector machine. *Journal of Machine Learning Research* **1** 211–244.
[46] Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society. Series B* **68** 49–67.
[47] Zhang, C.-H. (2007). Penalized linear unbiased selection. Technical Report, Department of Statistics, Rutgers University.
[48] Zhang, C. H. (2010). Nearly unbiased variable selection under minimax concave penalty. *Annals of Statistics* **38** 894–942.
[49] Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B* 2 301–320.