Sparsity and the truncated ℓ^2-norm

Lee H. Dicker
Department of Statistics and Biostatistics
Rutgers University
ldicker@stat.rutgers.edu

Abstract

Sparsity is a fundamental topic in high-dimensional data analysis. Perhaps the most common measures of sparsity are the ℓ^p-norms, for $0 \leq p < 2$. In this paper, we study an alternative measure of sparsity, the truncated ℓ^2-norm, which is related to other ℓ^p-norms, but appears to have some unique and useful properties. Focusing on the n-dimensional Gaussian location model, we derive exact asymptotic minimax results for estimation over truncated ℓ^2-balls, which complement existing results for ℓ^p-balls. We then propose simple new adaptive thresholding estimators that are inspired by the truncated ℓ^2-norm and are adaptive asymptotic minimax over ℓ^p-balls ($0 \leq p < 2$), as well as truncated ℓ^2-balls. Finally, we derive lower bounds on the Bayes risk of an estimator, in terms of the parameter's truncated ℓ^2-norm. These bounds provide necessary conditions for Bayes risk consistency in certain problems that are relevant for high-dimensional Bayesian modeling.

1 INTRODUCTION

Many methods for prediction and estimation in high-dimensional data analysis are designed to exploit additional structure that may be present, but not completely obvious, in a given dataset. Important examples of this type of structure include the existence of an interesting low-dimensional latent subspace, which is crucial for principal component and factor analysis-based methods (Jolliffe, 2002), and sparsity, which is the main focus of this paper. Broadly speaking, sparsity measures the extent to which a (typically high-dimensional) signal may be described by relatively few elements of some prespecified basis; if the signal is sparse, then only a handful of basis vectors may be required to accurately describe it. Sparse methods — that is, methods designed to exploit sparsity, like thresholding (Donoho and Johnstone, 1994a, 1995), lasso (Tibshirani, 1996) and other penalized estimation procedures (Zhang, 2010; Fan and Lv, 2011) — have proven to be extremely successful tools for high-dimensional data analysis in a wide range of applications, e.g., (Donoho, 1995; Wright et al., 2009; Erlich et al., 2010).

The performance of sparse methods generally hinges on some measure of sparsity in the underlying signal. Perhaps the most common measures of sparsity are the ℓ^p-norms, with $0 \leq p < 2$ (Abramovich et al., 2006; Donoho, 2006) [here, we restrict our attention to quadratic loss functions; otherwise, an alternative range of ℓ^p-norms might be of interest (Donoho and Johnstone, 1994b)]. For $\boldsymbol{\theta} = (\theta_1, ..., \theta_n)^\top \in \mathbb{R}^n$ and $0 < p < \infty$, the ℓ^p-norm is defined by $||\boldsymbol{\theta}||_p = (\sum_{i=1}^n |\theta_i|^p)^{1/p}$; the ℓ^0-norm $||\boldsymbol{\theta}||_0 = |\{i;\ \theta_i \neq 0\}|$ is the number of nonzero components of $\boldsymbol{\theta}$. (For $0 \leq p < 1$, the ℓ^p-norm is not a genuine norm; however, following standard practice, we still refer to it as such.) The ℓ^0- and ℓ^1-norms deserve special attention. The ℓ^0-norm is especially appealing from an intuitive perspective and is closely identified with hard-thresholding estimators and "subset selection" methods (Foster and George, 1994; Breiman, 1996). The ℓ^1-norm plays a key role in soft-thresholding and lasso procedures; moreover, significant computational advantages are often associated with the ℓ^1-norm (Tropp, 2006).

In this paper, we present a collection of results for an alternative measure of sparsity: the truncated ℓ^2-norm,

$$||\boldsymbol{\theta}||_{t,\sigma^2} = \left\{ \sum_{i=1}^n (\theta_i^2 \wedge \sigma^2) \right\}^{1/2}.$$

Appearing in Proceedings of the 17th International Conference on Artificial Intelligence and Statistics (AISTATS) 2014, Reykjavik, Iceland. JMLR: W&CP volume 33. Copyright 2014 by the authors.

Here $\sigma^2 > 0$ is the noise-level for the problem under consideration and $\theta_i^2 \wedge \sigma^2 = \min\{\theta_i^2, \sigma^2\}$. In what follows we typically take $\sigma^2 = 1$; indeed, define $||\boldsymbol{\theta}||_t = ||\boldsymbol{\theta}||_{t,1}$. We will refer to $||\cdot||_t$ as both the truncated ℓ^2-norm and, for convenience, the ℓ^t-norm. As with the ℓ^p-norm for $0 \leq p < 1$, the ℓ^t-norm is not a true norm. Generally, a vector is considered to be (ℓ^p-) sparse if its ℓ^p-norm is small, for some $p \in [0,2) \cup \{t\}$ (in this notation, t serves as a token for indicating the ℓ^t-norm; that is, taking $p = t$ refers to the ℓ^t-norm).

The truncated ℓ^2-norm has appeared elsewhere in the literature on sparse estimation, often in the context of oracle inequalities (Donoho and Johnstone, 1994a; Foster and George, 1994; Zhang, 2005; Candès and Tao, 2007). However, in some aspects, existing theory for ℓ^t-sparsity is less fully developed than for other ℓ^p-norms. In this paper, we hope to demonstrate that significant new insights into high-dimensional sparse estimation problems can be gained by further studying the truncated ℓ^2-norm. Focusing on the n-dimensional Gaussian location model, which is a cornerstone for the development and understanding of more complex statistical models (Johnstone, 2013), the three main technical contributions of this paper are: (i) we derive exact asymptotic minimax results for estimation over truncated ℓ^2-balls, assuming $n \to \infty$ and the sparsity condition $n^{-1}||\boldsymbol{\theta}||_t^2 \to 0$; (ii) we identify simple new adaptive thresholding estimators that are inspired by the truncated ℓ^2-norm; and (iii) we derive lower bounds on the Bayes risk of an estimator, in terms of the parameter's truncated ℓ^2-norm.

Taking a broader view of this paper's potential implications, our results suggest that the truncated ℓ^2-norm gives necessary and sufficient conditions for the successful application of "sparse methods" in high-dimensions. Thus, estimating the truncated ℓ^2-norm, or some proxy [e.g., $\kappa_t(\boldsymbol{\theta})$ defined in (13) below], might provide a simple summary measure of sparsity for high-dimensional datasets. This could be useful for exploratory data analysis in high dimensions. Furthermore, the lower bounds on Bayes risk derived in this paper may provide guidance on identifying effective priors for Bayesian inference in high-dimensional data analysis.

1.1 Overview of the paper

Section 2 covers preliminaries. We introduce the statistical model, some basic definitions and notation, and discuss some well-known elementary results involving the truncated ℓ^2-norm.

Sections 3–4 contain minimax and adaptive thresholding results. These results are related to work by Donoho et al. (1992), Donoho and Johnstone (1994b), Johnstone and Silverman (2004), Zhang (2005), Abramovich et al. (2006), Jiang and Zhang (2009) and others, who studied asymptotic minimaxity over ℓ^p-balls ($0 \leq p < 2$). Here, we extend results from (Donoho et al., 1992; Donoho and Johnstone, 1994b) to truncated ℓ^2-balls and propose simple new estimators that are adaptive asymptotic minimax over ℓ^p-balls and truncated ℓ^2-balls. In these results, we assume a sparsity condition, i.e., that the relevant ℓ^p-balls are very small. We emphasize that the asymptotic minimax results in Sections 3–4 paper are sharp; that is, we obtain the constants, as well as the rates.

In Section 5, we present lower bounds for Bayes risk that depend on the truncated ℓ^2-norm. Lower bounds and Bayes risk often play a key role in minimax analyses; however, our approach in Section 5 is somewhat different. In particular, our results in Section 5 require no sparsity assumptions, are valid for a broad class of prior distributions, and give lower bounds on the Bayes risk in terms of a quantity closely related to the truncated ℓ^2-norm. These results may shed some light on the consequences for high-dimensional problems when sparsity assumptions do not hold. This, in turn, could have significant practical implications for conducting valid statistical inference in high dimensions and Bayesian modeling. Two examples related to high-dimensional Bayesian modeling — the Bayesian lasso (Park and Casella, 2008) and the horseshoe prior (Carvalho et al., 2010) — are discussed at the end of Section 5.

Section 6 contains a concluding discussion. Proofs may be found in the Supplementary Material.

2 PRELIMINARIES

2.1 Notation and Definitions

We assume that the observed data consists of a single n-dimensional Gausian random vector $\mathbf{x} \sim N(\boldsymbol{\theta}, I_n)$, with unknown mean parameter $\boldsymbol{\theta} = (\theta_1, ..., \theta_n)^\top \in \mathbb{R}^n$. For an estimator $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}(\mathbf{x})$ of $\boldsymbol{\theta}$, define the risk

$$R(\hat{\boldsymbol{\theta}}; \boldsymbol{\theta}) = R_n(\hat{\boldsymbol{\theta}}; \boldsymbol{\theta}) = \frac{1}{n} E_{\boldsymbol{\theta}} \left\{ ||\hat{\boldsymbol{\theta}}(\mathbf{x}) - \boldsymbol{\theta}||_2^2 \right\}. \quad (1)$$

The subscript $\boldsymbol{\theta}$ in the expectation on the right-hand side of (1) serves to emphasize that $\mathbf{x} \sim N(\boldsymbol{\theta}, I_n)$ and that the expecation is conditional on $\boldsymbol{\theta}$.

In Sections 3–4, the performance of an estimator $\hat{\boldsymbol{\theta}}$ will primarily be measured by its maximal risk over some subset $\Theta \subseteq \mathbb{R}^n$,

$$R(\hat{\boldsymbol{\theta}}; \Theta) = \sup_{\boldsymbol{\theta} \in \Theta} R(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}). \quad (2)$$

In Section 5, the Bayes risk plays a more prominent role; we defer the definition of Bayes risk until then.

For the maximal risk (2), we are primarily interested in cases where Θ is an ℓ^p-ball. For $\boldsymbol{\theta} \in \mathbb{R}^n$ and $p \in [0,\infty) \cup \{t\}$, define the standardized ℓ^p-norm

$$\eta_p(\boldsymbol{\theta}) = \eta_{p,n}(\boldsymbol{\theta}) = \begin{cases} n^{-1}||\boldsymbol{\theta}||_0 & \text{if } p = 0, \\ n^{-1}||\boldsymbol{\theta}||_p^p & \text{if } 0 < p < \infty, \\ n^{-1}||\boldsymbol{\theta}||_t^2 & \text{if } p = t. \end{cases}$$

Define the ℓ^p-ball with standardized radius $\eta \geq 0$ by

$$B_n^p(\eta) = \{\boldsymbol{\theta} \in \mathbb{R}^n;\ \eta_p(\boldsymbol{\theta}) \leq \eta\}.$$

As seen below in (10), focusing on the standardized ℓ^p-norm facillitates a more direct comparison between ℓ^p-balls with different p; additionally, it helps to emphasize connections between ℓ^p-norms and p-th moments of prior distributions (see Section 5).

We use the following asymptotic notation throughout the paper. Suppose that $\{a_\iota\}_{\iota \in I}$, $\{b_\iota\}_{\iota \in I}$ are two collections of numbers indexed by some set I. We write $a_\iota \sim b_\iota$, if $a_\iota / b_\iota \to 1$ under some specified limiting conditions $\iota \to \iota_0$; we write $a_\iota \lesssim b_\iota$, if $\limsup a_\iota / b_\iota \leq 1$. The notation $a_\iota = O(b_\iota)$ means that there is a constant $C \in \mathbb{R}$ such that $|a_\iota| \leq C b_\iota$ for all $\iota \in I$. .

2.2 Basic Results for ℓ^t-Balls

In this subsection, we discuss some well-known inequalities for soft-thresholding that illustrate the usefulness of the ℓ^t-norm and are closely related to many of the results that follow. For $x, y \in \mathbb{R}$, let $x \vee y = \max\{x, y\}$. Define the soft-thresholding function $s_\lambda(x) = \text{sign}(x)\{(|x| - \lambda) \vee 0\}$ ($x \in \mathbb{R}$, $\lambda \geq 0$) and the corresponding soft-thresholding estimator $\hat{\boldsymbol{\theta}}_\lambda = \hat{\boldsymbol{\theta}}_\lambda(\mathbf{x}) = (s_\lambda(x_1), ..., s_\lambda(x_n))^\top$. Soft-thresholding is one of the fundamental sparse estimators (Donoho and Johnstone, 1994b,a); it is closely related to lasso and has inspired many other thresholding and penalized estimation procedures (Tibshirani, 1996; Fan and Li, 2001). Notice that $||\hat{\boldsymbol{\theta}}_\lambda||_0 < n$ with positive probability; this provides one simple justification for referring to soft-thresholding as a "sparse estimator."

Let $\lambda_{\text{univ}} = \{2\log(n)\}^{1/2}$ and define the universal thresholding estimator $\hat{\boldsymbol{\theta}}_{\text{univ}} = \hat{\boldsymbol{\theta}}_{\lambda_{\text{univ}}}$ (Donoho and Johnstone, 1994a). Then

$$\begin{aligned} \frac{1}{2}\eta_t(\boldsymbol{\theta}) &\leq \inf_{\lambda \geq 0} R(\hat{\boldsymbol{\theta}}_\lambda; \boldsymbol{\theta}) && (3) \\ &\leq R(\hat{\boldsymbol{\theta}}_{\text{univ}}; \boldsymbol{\theta}) && (4) \\ &\leq \{2\log(n) + 1\}\left\{\frac{1}{n} + \eta_t(\boldsymbol{\theta})\right\} && (5) \end{aligned}$$

for $\boldsymbol{\theta} \in \mathbb{R}^n$. The inequality (3) follows from Lemma 8.3 of (Johnstone, 2013); (4) is trivial; (5) follows from Proposition 8.8 of (Johnstone, 2013). The inequalities (3)–(5) imply that the performance of soft-thresholding is determined, up to a log-factor, by the (standardized) ℓ^t-norm of $\boldsymbol{\theta}$.

To see where other sparse ℓ^p-norms fit in, notice that

$$\eta_t(\boldsymbol{\theta}) \leq \eta_p(\boldsymbol{\theta}), \quad \boldsymbol{\theta} \in \mathbb{R}^n,\ 0 \leq p < 2. \tag{6}$$

[In fact (6) holds for $0 \leq p \leq 2$, but the ℓ^2-norm is less relevant for the present discussion of sparsity.] It follows from (5) and (6) that if $\boldsymbol{\theta} \in \mathbb{R}^n$ and $0 \leq p < 2$, then

$$R(\hat{\boldsymbol{\theta}}_{\text{univ}}; \boldsymbol{\theta}) \leq \{2\log(n) + 1\}\left\{\frac{1}{n} + \eta_p(\boldsymbol{\theta})\right\}. \tag{7}$$

Hence, the risk of $\hat{\boldsymbol{\theta}}_{\text{univ}}$ is small when $\eta_p(\boldsymbol{\theta})$ is small. On the other hand, there is no corresponding lower bound on $\inf_{\lambda \geq 0} R(\hat{\boldsymbol{\theta}}_\lambda; \boldsymbol{\theta})$ involving the ℓ^p-norm for $0 \leq p < 2$. Thus, while good upper bounds on $R(\hat{\boldsymbol{\theta}}_{\text{univ}}; \boldsymbol{\theta})$ are available in terms of the ℓ^p-norm for all $p \in [0,2) \cup \{t\}$, lower-bounds like (3) are only valid for the truncated ℓ^2-norm.

Most of the results in this paper may be viewed as extensions or refinements of the inequalities (3)–(5). The results in Sections 3-4 are, in a sense, refinements of the upper bound (5). Indeed, the results in Section 3 largely parallel existing results for ℓ^p-norms ($0 \leq p < 2$), which may themselves be viewed as refinements of (7); the adaptation results in Section 4 further demonstrate how results for ℓ^p-norms ($0 \leq p < 2$) may follow easily from results on ℓ^t-norms, similar to how (7) follows from (5)–(6).

Our results for Bayes risk in Section 5 build on the lower bound (3). While (3) applies only to thresholding estimators, the lower bounds in Section 5 are valid for all estimators $\hat{\boldsymbol{\theta}}$ and involve a Bayesian analogue of $\eta_t(\boldsymbol{\theta})$.

3 MINIMAX RISK OVER ℓ^t-BALLS

For $p \in [0,\infty) \cup \{t\}$, define the minimax risk over $B_n^p(\eta)$,

$$r_n^p(\eta) = \inf_{\hat{\boldsymbol{\theta}}} R\{\hat{\boldsymbol{\theta}}; B_n^p(\eta)\} = \inf_{\hat{\boldsymbol{\theta}}} \sup_{\boldsymbol{\theta} \in B_n^p(\eta)} R(\hat{\boldsymbol{\theta}}; \boldsymbol{\theta}),$$

where the infimum is taken over all measurable estimators $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}(\mathbf{x})$. Our first result gives exact asymptotics for the minimax risk $r_n^t(\eta)$, as $n \to \infty$, $\eta \to 0$, and implies that soft-thresholding is asymptotically minimax. (Corresponding results are available for hard-thresholding estimators; we focus on soft-thresholding here because it is slightly more convenient mathematically.) For $\eta > 0$, define the thresholding level

$$\lambda_\eta = \{2\log(\eta^{-1})\}^{1/2}.$$

Theorem 1. *If $n \to \infty$, $\eta \to 0$ and $\eta \geq n^{-1}$, then*

$$r_n^t(\eta) \sim R\{\hat{\boldsymbol{\theta}}_{\lambda_\eta}; B_n^t(\eta)\} \sim 2\eta \log(\eta^{-1}).$$

Theorem 1 gives the exact asymptotic minimax risk over ℓ^t-balls of radius η, and implies that soft-thresholding at level $\lambda_\eta = \{2\log(\eta^{-1})\}^{1/2}$ is an asymptotically minimax estimator. The condition $\eta \to 0$ in Theorem 1 is a sparsity condition, which ensures that the signal $\boldsymbol{\theta}$ is ℓ^t-sparse; the condition $\eta \geq n^{-1}$ is a signal strength condition which ensures that the signal is not too small.

Theorem 1 is closely related to existing work by Donoho, Johnstone, and coauthors on ℓ^p-balls, with $0 \leq p < 2$. The results in (Donoho et al., 1992) imply that if $n \to \infty$, $\eta \to 0$ and $\eta \geq n^{-1}$, then

$$r_n^0(\eta) \sim R\{\hat{\boldsymbol{\theta}}_{\lambda_\eta}; B_n^0(\eta)\} \sim 2\eta \log(\eta^{-1}); \qquad (8)$$

in (Donoho and Johnstone, 1994b), it was shown that if $0 < p < 2$, $n \to \infty$, $\eta \to 0$ and $\eta n / \{\log(n)\}^{p/2} \to \infty$, then

$$r_n^p(\eta) \sim R\{\hat{\boldsymbol{\theta}}_{\lambda_\eta}; B_n^p(\eta)\} \sim \eta\{2\log(\eta^{-1})\}^{1-p/2}. \qquad (9)$$

The result for ℓ^0-balls (8) is key to proving Theorem 1. Indeed, (6) implies that

$$B_n^p(\eta) \subseteq B_n^t(\eta), \quad 0 \leq p < 2, \ \eta \geq 0. \qquad (10)$$

Hence,

$$2\eta\log(\eta^{-1}) \sim r_n^0(\eta) \leq r_n^t(\eta) \leq R\{\hat{\boldsymbol{\theta}}_{\lambda_\eta}; B_n^t(\eta)\}. \quad (11)$$

Theorem 1 follows upon proving that $R\{\hat{\boldsymbol{\theta}}_{\lambda_\eta}; B_n^t(\eta)\} \lesssim 2\eta\log(\eta^{-1})$. Details may be found in the Supplementary Material.

In addition to playing an important role in the proof of Theorem 1, the inclusion (10) provides some geometric insight into the relationship between ℓ^t- and ℓ^p-sparsity ($0 \leq p < 2$). Figure 1 provides a graphical illustration of the ℓ^p-balls for $p = t, 0, 1$, $n = 2$ and $\eta = 5/8$, which clearly depicts the inclusions (10).

Theorem 1 and (8) imply that it is just as difficult to estimate $\boldsymbol{\theta}$ over ℓ^t- as it is to estimate $\boldsymbol{\theta}$ over ℓ^0-balls (since the minimax risk in Theorem 1 and (8) are the same). On the other hand, Theorem 1 and (9) imply that it is slightly easier to estimate $\boldsymbol{\theta}$ over ℓ^p-balls for $0 < p < 2$ (by a factor of $\{2\log(\eta^{-1})\}^{p/2}$).

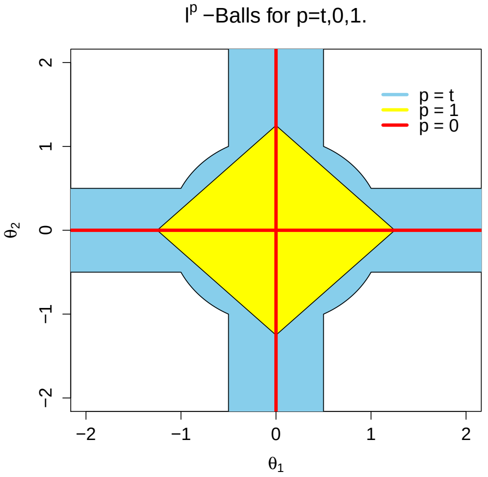

Figure 1: The ℓ^p-balls $B_n^p(\eta)$ for $p = t, 0, 1$, $n = 2$ and $\eta = 5/8$. Observe that $B_n^0(\eta), B_n^1(\eta) \subseteq B_n^t(\eta)$.

To see how Theorem 1 relates to the upper bound for universal thresholding in (5), observe that if $n \to \infty$, $\eta_t(\boldsymbol{\theta}) \to 0$ and $\eta_t(\boldsymbol{\theta}) \geq n^{-1}$, then Theorem 1 implies

$$\begin{aligned} R\{\hat{\boldsymbol{\theta}}_{\lambda_{\eta_t(\boldsymbol{\theta})}}; \boldsymbol{\theta}\} &\lesssim 2\eta_t(\boldsymbol{\theta})\log\{\eta_t(\boldsymbol{\theta})^{-1}\} \qquad (12)\\ &= \left[\frac{\log\{\eta_t(\boldsymbol{\theta})^{-1}\}}{\log(n)}\right] \times 2\eta_t(\boldsymbol{\theta})\log(n)\\ &\lesssim 2\eta_t(\boldsymbol{\theta})\log(n). \end{aligned}$$

It follows that, in the specified setting, the upper bound (12) improves upon the upper bound for universal thresholding in (5). Hence, it may be beneficial to threshold at level $\lambda = \lambda_{\eta_t(\boldsymbol{\theta})}$, as opposed to $\lambda = \lambda_{\text{univ}}$. However, in practice, it is typically unreasonable to assume that $\eta_t(\boldsymbol{\theta})$ is known; thus, it is not possible to implement the estimator $\hat{\boldsymbol{\theta}}_{\lambda_{\eta_t(\boldsymbol{\theta})}}$. This issue is addressed in the next section, where we discuss adaptive thresholding.

4 ADAPTIVE THRESHOLDING

Since $\eta_t(\boldsymbol{\theta})$ is typically unknown, a reasonable strategy is to replace $\eta_t(\boldsymbol{\theta})$ in $\hat{\boldsymbol{\theta}}_{\lambda_{\eta_t(\boldsymbol{\theta})}}$ with an estimate. One obstacle to this approach is that it is challenging to accurately estimate $\eta_t(\boldsymbol{\theta})$; according to Zhang (2005), $\eta_t(\boldsymbol{\theta})$ is only estimable at logarithmic rates. On the other hand, Zhang (2005) proposed a surrogate for $\eta_t(\boldsymbol{\theta})$ that is more easily estimated. Define

$$\kappa_t(\boldsymbol{\theta}) = 1 - \frac{1}{n}\sum_{i=1}^{n} e^{-\theta_i^2/4}. \qquad (13)$$

It is elementary to check that

$$\frac{e-1}{4e}\eta_t(\boldsymbol{\theta}) \leq \kappa_t(\boldsymbol{\theta}) \leq \eta_t(\boldsymbol{\theta}). \qquad (14)$$

The main implication of (14) is that if $\eta_t(\boldsymbol{\theta}) \to 0$, then $\kappa_t(\boldsymbol{\theta}) \to 0$ at the same rate, and conversely. Hence, by estimating $\kappa_t(\boldsymbol{\theta})$, we can estimate the rate of $\eta_t(\boldsymbol{\theta})$; it turns out that this is sufficient to construct an adaptive asymptotically minimax thresholding estimator, with properties similar to those of $\hat{\boldsymbol{\theta}}_{\lambda_{\eta_t(\boldsymbol{\theta})}}$.

Define

$$\hat{\kappa}_t = 1 - \frac{1}{n}\sum_{i=1}^n \sqrt{2}e^{-x_i^2/2}.$$

It is easily seen that $\hat{\kappa}_t$ is an unbiased estimator for $\kappa_t(\boldsymbol{\theta})$ with $\mathrm{Var}(\hat{\kappa}_t) = O(n^{-1})$. Now define the adaptive thresholding level

$$\hat{\lambda} = \left[2\log\left\{\left(\hat{\kappa}_t \vee \frac{1}{n}\right)^{-1}\right\}\right]^{1/2}. \qquad (15)$$

Theorem 2. *Let $\hat{\boldsymbol{\theta}}_{\hat{\lambda}}$ be the soft-thresholding estimator at level $\hat{\lambda}$, defined in (15). Suppose that $p \in [0,2) \cup \{t\}$ is fixed. If $n \to \infty$, $\eta \to 0$, and $\eta\{n/\log(n)\}^{1/2} \to \infty$, then*

$$R\{\hat{\boldsymbol{\theta}}_{\hat{\lambda}}; B_n^p(\eta)\} \sim r_n^p(\eta).$$

Theorem 2 implies that $\hat{\boldsymbol{\theta}}_{\hat{\lambda}}$ is asymptotically minimax over ℓ^p-balls, for all $p \in [0,2) \cup \{t\}$, assuming the sparsity condition $\eta \to 0$ and the signal strength condition $\eta\{n/\log(n)\}^{1/2} \to \infty$. The estimator $\hat{\boldsymbol{\theta}}_{\hat{\lambda}}$ is an adaptive thresholding estimator because the thresholding level adapts to the various ℓ^p-norms and the radius η.

The signal strength condition $\eta\{n/\log(n)\}^{1/2} \to \infty$ in Theorem 2 is considerably stronger than that in Theorem 1 and those required for (8)–(9). This is because of the error inherent in $\hat{\kappa}_t$ for estimating κ_t. It may be of interest to investigate how much the signal strength condition in Theorem 2 can be relaxed, either by more carefully analyzing $\hat{\boldsymbol{\theta}}_{\hat{\lambda}}$ or considering other related estimators.

Other adaptive estimators, which are asymptotically minimax over ℓ^p-balls for a range of values p and radii η, have been previously proposed. Some of these estimators are relatively complex (Johnstone and Silverman, 2004; Zhang, 2005; Jiang and Zhang, 2009). Abramovich et al. (2006) proposed adaptive thresholding estimators based on procedures for controlling the false discovery rate in multiple testing problems; their work is probably the most relevant for comparison to Theorem 2. The main result of Abramovich et al. (2006) (Theorem 1.1) requires $\eta \geq n^{-1}\{\log(n)\}^5$, which is a substantially weaker signal strength condition than that in Theorem 2; but their result also requires $\eta \leq n^{-\delta}$ for some fixed $\delta > 0$, which is a stronger sparsity condition than in Theorem 2.

The proof of Theorem 2 contains two key elements and may be found in the Supplementary Material. The first key point is the approximation

$$R(\hat{\boldsymbol{\theta}}_{\hat{\lambda}}; \boldsymbol{\theta}) \approx R\{\hat{\boldsymbol{\theta}}_{\lambda_{\eta_t(\boldsymbol{\theta})}}; \boldsymbol{\theta}\}. \qquad (16)$$

The second key is the following proposition, which is proved in the Supplementary Material.

Proposition 1. *Fix $0 \leq p < 2$. If $n \to \infty$ and $\eta \to 0$, then*

$$\sup_{\boldsymbol{\theta} \in B_n^p(\eta)} R\{\hat{\boldsymbol{\theta}}_{\lambda_{\eta_t(\boldsymbol{\theta})}}; \boldsymbol{\theta}\} \lesssim \eta\left\{2\log(\eta^{-1})\right\}^{1-p/2}.$$

Proposition 1 further illustrates the usefulness of the ℓ^t-norm vis-à-vis other ℓ^p-norms. Indeed, together with (8)–(9), Proposition 1 implies that if one thresholds at a level determined by the ℓ^t-norm, i.e., $\lambda = \lambda_{\eta_t(\boldsymbol{\theta})}$, then asymptotic minimaxity over ℓ^p-balls follows automatically, for each $0 \leq p < 2$. Theorem 2 follows by combining (16) with Proposition 1 and (8)–(9).

5 BAYES RISK AND LOWER BOUNDS

5.1 General Results

In the previous sections, we focused largely on soft-thresholding estimators and minimax risk over ℓ^p-balls ($p \in [0,2) \cup \{t\}$) with shrinking radius η. Taking a broader view, the techniques employed above — whereby one considers the maximal risk of a specific estimator (e.g., soft-thresholding) over some restricted parameter space (e.g., an ℓ^p-ball) — illustrate a common method for obtaining minimax upper bounds. Lower bounds on minimax risk are generally derived using Bayesian arguments, which involve bounding the Bayes risk for a sequence of approximately least favorable prior distributions (in the proof of Theorem 1, we avoid direct use of Bayesian arguments by appealing to the lower bound (11); the proof of (8), which is crucial for (11), relies heavily on Bayesian techniques).

Beyond its role in minimax arguments, a detailed analysis of Bayes risk may be useful for more practical purposes in high-dimensional Bayesian modeling. Conditioning on hyperparameters as necessary, many previously proposed prior distributions $\boldsymbol{\pi}$ for high-dimensional Bayesian inference are n-fold products of a symmetric one-dimensional distribution π, i.e., $\boldsymbol{\pi} = \otimes^{(n)}\pi$ (Park and Casella, 2008; Carvalho et al., 2010). The Bayes risk

$$\begin{aligned} r(\pi) &= \inf_{\hat{\theta}} \int_{\mathbb{R}} E_\theta\left[\{\hat{\theta}(x) - \theta\}^2\right] \, d\pi(\theta) \qquad (17) \\ &= \inf_{\hat{\boldsymbol{\theta}}} \int_{\mathbb{R}^n} E_{\boldsymbol{\theta}}\left\{||\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}||^2\right\} \, d\boldsymbol{\pi}(\boldsymbol{\theta}), \end{aligned}$$

where $x \sim N(\theta, 1)$ in the first expectation above, seems to be a reasonable measure for gauging the appropriateness of these priors and the effectiveness of the associated Bayes procedures in high dimensions. For instance, if $\boldsymbol{\tau} \in T \subseteq \mathbb{R}^k$ is a hyperparameter for further specifying $\pi = \pi(\cdot|\boldsymbol{\tau})$, then we should hope that for some choice of $\boldsymbol{\tau} \in T$, the Bayes risk $r\{\pi(\cdot|\boldsymbol{\tau})\}$ is very small; otherwise, it would be impossible to conduct effective inference using the class of priors $\{\pi(\cdot|\boldsymbol{\tau});\ \boldsymbol{\tau} \in T\}$. Moreover, in problems where sparsity is an important feature, it may be useful to characterize priors from $\{\pi(\cdot|\boldsymbol{\tau});\ \boldsymbol{\tau} \in T\}$ by some sparsity-related measure [e.g., $\eta_t(\pi)$, defined in (18) below] that can be linked to their Bayes risk. In the remainder of this subsection, we take some intial steps towards this goal; some examples are considered in the next subsection.

If $\boldsymbol{\theta} \sim \boldsymbol{\pi} = \otimes^{(n)}\pi$ and n is large, then

$$\eta_t(\boldsymbol{\theta}) = \frac{1}{n}||\boldsymbol{\theta}||_t^2 \approx \int_{\mathbb{R}} |\theta|^2 \wedge 1\ d\pi(\theta).$$

This motivates the following definition for the truncated ℓ^2-norm of π,

$$\eta_t(\pi) = \int_{\mathbb{R}} |\theta|^2 \wedge 1\ d\pi(\theta). \tag{18}$$

Similarly, define

$$\eta_p(\pi) = \int_{\mathbb{R}} |\theta|^p\ d\pi(\theta), \quad 0 < p < \infty.$$

A vector $\boldsymbol{\theta} \in \mathbb{R}^n$ is considered to be sparse if $\eta_p(\boldsymbol{\theta})$ is small ($p \in [0, 2) \cup \{t\}$); thus, it seems reasonable that a "sparse" prior π should have small $\eta_p(\pi)$. On the other hand, if $\eta_2(\pi)$ is also very small, then signal strength issues may limit the effectiveness of any estimation procedure (Donoho and Jin, 2004). Hence, we focus on the variance standardized ℓ^p-norm, $\nu_p(\pi) = \eta_p(\pi)/\eta_2(\pi)$, as a measure of sparsity in what follows. The next theorem is our main result in this section. It is proved in the Supplementary Material.

Theorem 3. *Let π be a probability distribution on $\mathbb{R}$ that is symmetric about 0 and has finite variance. Let $\nu_t(\pi) = \eta_t(\pi)/\eta_2(\pi)$ be the variance standardized ℓ^t-norm of π. Then*

$$r(\pi) \geq \frac{\eta_2(\pi)}{\sqrt{2}} \left[\left\{\frac{1}{4e^2}\nu_t(\pi)\right\} \wedge e^{-8/\nu_t(\pi)}\right].$$

Theorem 3 implies that if $\{\pi_n\}$ is a sequence of symmetric prior distributions with finite variance, then $r(\pi_n)/\eta_2(\pi_n) \to 0$ implies $\nu_t(\pi_n) \to 0$. Consequently, if the priors $\{\pi_n\}$ have bounded variance, then the $\nu_t(\pi_n) \to 0$ is a necessary condition for Bayes risk consistency, i.e., $r(\pi_n) \to 0$. In settings where $\eta_2(\pi_n) \to 0$, consistency holds automatically because $r(\pi_n) \leq \eta_2(\pi_n)$ [to see this, take $\hat{\theta} = \hat{\theta}_{\text{null}} = 0$ in (17)]; in these settings, Theorem 3 implies that $\nu_t(\pi_n) \to 0$ is a necessary condition for the existence of an estimator that substantially outperforms $\hat{\theta}_{\text{null}}$.

Theorem 3 does not apply to infinite variance priors and may be less informative for studying classes of prior distributions with unbounded variance. This is a significant issue that we address in a fairly ad hoc manner in the examples below; a more systematic approach is desirable and is a topic for future research.

More broadly, Theorem 3 gives a lower bound on the performance of Bayesian methods in terms of the prior distribution's ℓ^t-norm. It seems unlikely that similar lower bounds are available for other sparse ℓ^p-norms. This complements our previous observation in Section 2.2 that lower bounds for soft-thresholding like (3) do not appear to exist for other ℓ^p-norms ($0 \leq p < 2$) and further highlights interesting features of the truncated ℓ^2-norm.

5.2 Examples

5.2.1 The Laplace prior

The Laplace prior

$$\pi_{\text{L}}(\theta|b) = (2b)^{-1}e^{-|\theta|/b}, \quad \theta \in \mathbb{R},\ b > 0.$$

has frequently been associated with soft-thresholding and lasso procedures (Tibshirani, 1996). Park and Casella (2008) provide a Bayesian analysis of π_{L}, focused mostly on computational issues in regression settings. Here, we study the Bayes risk of π_{L}, using tools from the previous subsection.

It is elementary to check that

$$\nu_t\{\pi_{\text{L}}(\cdot|b)\} = 1 - \left(1 + \frac{1}{b}\right)e^{-1/b}.$$

Since

$$\inf_{0<b\leq B} \nu_t\{\pi_{\text{L}}(\cdot|b)\} = 1 - \left(1 + \frac{1}{B}\right)e^{-1/B} > 0$$

for any positive real number B, Theorem 3 implies that $r\{\pi_{\text{L}}(\cdot|b)\}/\eta_2\{\pi_{\text{L}}(\cdot|b)\}$ is bounded away from 0 for any bounded collection of hyperparameters b. On the other hand, $\eta_2\{\pi(\cdot|b)\} \to \infty$ as $b \to \infty$, which implies that the bound in Theorem 3 is trivial for unbounded b. Still, by direct calculation or by appealing to more general results, e.g. Theorem 5.3 of (Mukopadhyay and DasGupta, 1993), one can check that $\inf_{B<b} r\{\pi_{\text{L}}(\cdot|b)\} > 0$. We conclude that

$$\inf_{0<b<\infty} \frac{r\{\pi_{\text{L}}(\cdot|b)\}}{\eta_2\{\pi_{\text{L}}(\cdot|b)\} \wedge 1} > 0.$$

This may be viewed as a kind of inconsistency result for Laplace priors π_{L} when used in certain high-dimensional problems.

5.2.2 The horseshoe prior

Carvalho et al. (2010) proposed the horseshoe prior for high-dimensional Bayesian modeling. Under the horseshoe prior $\pi_{\mathrm{H}}(\cdot|\tau)$,

$$\theta|\psi \sim N(0, \psi^2),$$

where ψ follows a Cauchy(τ) distribution with density

$$f(\psi|\tau) = \frac{1}{\pi}\left(\frac{\tau}{\tau^2+\psi^2}\right), \quad \psi \in \mathbb{R},\ \tau > 0.$$

Clearly, if $\tau > 0$, then $\eta_2\{\pi_{\mathrm{H}}(\cdot|\tau)\} = \infty$ and Theorem 3 does not apply. However, it may still be informative to examine the Bayes risk $r\{\pi_{\mathrm{H}}(\cdot|\tau)\}$.

Supressing the dependence of π_{H} on τ in our notation, basic properties of conditional expectation imply that

$$r(\pi_{\mathrm{H}}) \geq E_{\pi_{\mathrm{H}}}\left[\{E_{\pi_{\mathrm{H}}}(\theta|x,\psi) - \theta\}^2\right], \tag{19}$$

where the subscript π_{H} in the expectation above indicates that $x|\theta \sim N(\theta, 1)$ and $\theta \sim \pi_{\mathrm{H}}$. Conditional on ψ, (x, θ) are jointly Gaussian. It follows that

$$E_{\pi_{\mathrm{H}}}(\theta|x,\psi) = \frac{\psi^2}{1+\psi^2}x$$

and

$$E_{\pi_{\mathrm{H}}}\left[\{E_{\pi_{\mathrm{H}}}(\theta|x,\psi) - \theta\}^2 \middle| \psi\right] = \frac{\psi^2}{1+\psi^2}.$$

Combining this with (19), we obtain

$$r(\pi_{\mathrm{H}}) \geq \frac{1}{\pi}\int_{\mathbb{R}} \frac{\psi^2}{1+\psi^2}\left(\frac{\tau}{\tau^2+\psi^2}\right)\,d\psi = \frac{\tau}{\tau+1}.$$

It follows that if τ is bounded away from 0, then so is the Bayes risk $r(\pi_{\mathrm{H}})$. Furthermore, if $\tau \to 0$, then $r(\pi_{\mathrm{H}})$ can converge to 0 at a rate no faster than τ.

The results in the previous paragraph suggest that only small values of τ could potentially yield priors $\pi_{\mathrm{H}}(\cdot|\tau)$ with arbitrarily small Bayes risk. This, in turn, suggests that the horseshoe may be most effective when τ is small or if τ follows some hyperprior that is concentrated near 0.

6 Conclusions

In this paper, we derived exact asymptotic minimax results for truncated ℓ^2-balls in the Gaussian location model. We proposed simple adaptive thresholding estimators, which are inspired by the truncated ℓ^2-norm and are adaptive asymptotic minimax over ℓ^p-balls for all $p \in [0, 2) \cup \{t\}$. Additionally, we used the truncated ℓ^2-norm to derive lower bounds on the Bayes risk in estimation problems that may have implications for high-dimensional Bayesian modeling; in particular, our lower bounds provide necessary conditions for effective Bayesian inference in certain high-dimensional problems. One limitation of our lower bounds is that they only apply to prior distributions with finite variance. Relaxing this requirement, or identifying other analytcal methods for understanding the behavior of infinite variance priors in high-dimensional Bayesian models is an area of interest for future research. Extending this work to other statistical settings, e.g., regression models, is also of interest.

Acknowledgements

This work was supported by U.S. National Science Foundation Grant DMS-1208785.

References

Abramovich, F., Benjamini, Y., Donoho, D. L. and Johnstone, I. M. (2006). Adapting to unknown sparsity by controlling the false discovery rate. *Ann. Stat.* **34** 584–653.

Breiman, L. (1996). Heuristics of instability and stabilization in model selection. *Ann. Stat.* **24** 2350–2383.

Candès, E. and Tao, T. (2007). The Dantzig selector: Statistical estimation when p is much larger than n. *Ann. Stat.* **35** 2313–2351.

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

Donoho, D. L. (1995). De-noising by soft-thresholding. *IEEE T. Inform. Theory* **41** 613–627.

Donoho, D. L. (2006). Compressed sensing. *IEEE T. Inform. Theory* **52** 1289–1306.

Donoho, D. L. and Jin, J. (2004). Higher criticism for detecting sparse heterogeneous mixtures. *Ann. Stat.* **32** 962–994.

Donoho, D. L. and Johnstone, I. M. (1994a). Ideal spatial adaptation by wavelet shrinkage. *Biometrika* **81** 425–455.

Donoho, D. L. and Johnstone, I. M. (1994b). Minimax risk over ℓ^p-balls for ℓ^q-error. *Probab. Theory Rel.* **99** 277–303.

Donoho, D. L. and Johnstone, I. M. (1995). Adapting to unknown smoothness via wavelet shrinkage. *J. Am. Stat. Assoc.* **90** 1200–1224.

Donoho, D. L., Johnstone, I. M., Hoch, J. C. and Stern, A. S. (1992). Maximum entropy and the nearly black object. *J. Roy. Stat. Soc. B* 41–81.

Erlich, Y., Gordon, A., Brand, M., Hannon, G. J. and Mitra, P. P. (2010). Compressed genotyping. *IEEE T. Inform. Theory* **56** 706–723.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Am. Stat. Assoc.* **96** 1348–1360.

Fan, J. and Lv, J. (2011). Nonconcave penalized likelihood with NP-dimensionality. *IEEE T. Inform. Theory* **57** 5467–5484.

Foster, D. P. and George, E. I. (1994). The risk inflation criterion for multiple regression. *Ann. Stat.* 1947–1975.

Jiang, W. and Zhang, C.-H. (2009). General maximum likelihood empirical bayes estimation of normal means. *Ann. Stat.* **37** 1647–1684.

Johnstone, I. M. (2013). *Gaussian Estimation: Sequence and Wavelet Models*. Monograph. Available at http://www-stat.stanford.edu/∼imj/GE06-11-13.pdf.

Johnstone, I. M. and Silverman, B. W. (2004). Needles and straw in haystacks: Empirical bayes estimates of possibly sparse sequences. *Ann. Stat.* **32** 1594–1649.

Jolliffe, I. T. (2002). *Principal Component Analysis*. 2nd ed. Springer.

Mukopadhyay, S. and DasGupta, A. (1993). Uniform approximation of Bayes solutions and posteriors: Frequentistly valid Bayes inference. Tech. Rep. #93-12C, Purdue University.

Park, T. and Casella, G. (2008). The Bayesian lasso. *J. Am. Stat. Assoc.* **103** 681–686.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Stat. Soc. B* **58** 267–288.

Tropp, J. A. (2006). Just relax: Convex programming methods for identifying sparse signals in noise. *IEEE T. Inform. Theory* **52** 1030–1051.

Wright, J., Yang, A. Y., Ganesh, A., Sastry, S. S. and Ma, Y. (2009). Robust face recognition via sparse representation. *IEEE T. Pattern Anal.* **31** 210–227.

Zhang, C.-H. (2005). General empirical Bayes wavelet methods and exactly adaptive minimax estimation. *Ann. Stat.* 54–100.

Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *Ann. Stat.* **38** 894–942.