IMS Collections
Contemporary Developments in Bayesian Analysis and Statistical Decision Theory: A Festschrift for William E. Strawderman
Vol. 8 (2012) 78–89

DOI: 10.1214/11-IMSCOLL806

Minimax ℓ_q risk in ℓ_p balls

Cun-Hui Zhang[1,*]

Rutgers University

Abstract: This paper provides an extension of earlier results on minimax estimation of a high-dimensional sparse vector to even more sparse vectors. Specifically, an approximation of the minimax ℓ_q risk is obtained and threshold estimators are proved to achieve the minimax risk within an infinitesimal fraction in all small ℓ_p balls.

Contents

1. Introduction

This paper concerns the estimation of a high-dimensional vector with a standard normal error. The multivariate normal distribution is a primary model in many areas in statistics, including compound decision theory and empirical Bayes, admissibility, adaptive nonparametric estimation, variable selection and multiple testing.

The estimation of a vector under the ℓ_q loss can be viewed as a compound decision problem in which the problems of estimating the individual components of the vector are combined. The fundamental idea of the compound decision theory and empirical Bayes (Robbins, 1951, 1956) is that the compound risk of individual problems can be substantially reduced by making individual decisions based on data from all problems involved. This has been proven to hold even in cases where the individual problems are independent.

For the estimation of a vector with standard normal error, the optimal invariant estimator can be improved upon under the ℓ_2 loss in spaces of dimension three or higher (Stein, 1956; James and Stein, 1961). This improvement is achieved by linear shrinkage of the optimal invariant estimator towards zero with an adaptive factor.

*Supported in part by NSF Grants DMS 0804626, DMS 0906420 and NSA Grant H98230-09-1-0006
[1]Department of Statistics and Biostatistics, Rutgers University, Hill Center, Busch Campus, Piscataway, NJ 08854, USA, e-mail: czhang@stat.rutgers.edu
AMS 2000 subject classifications: Primary 62J05, 62J07; secondary 62H12, 62H25
Keywords and phrases: empirical Bayes, compound decision rules, random effects, regression, nonparametric estimation, semiparametric estimation

The linear shrinkage approach was further developed in many directions, including admissibility (Brown, 1966, 1971), minimax Bayes methods with harmonic priors (Strawderman, 1971, 1973), and an linear empirical Bayes interpretation (Efron and Morris, 1972, 1973). We refer to Zhang (2003) for further discussion.

Linear shrinkage has important implications in nonparametric statistics since an infinite-dimensional parameter can be represented by its coefficients in a suitable basis. In nonparametric regression, an application of linear shrinkage estimators to blocks of coefficients of an unknown smooth regression function yields sharp adaptive (simultaneous asymptotic) minimax estimators of a regression function in Sobolev balls of all sizes and degrees of smoothness (Efromovich and Pinsker, 1984, 1986).

Linear shrinkage has its limitations. Rate adaptive minimax estimation (within a constant factor to minimax) in classes of regression functions with inhomogeneous smoothness (e.g. discontinuity) can be attained by thresholding the estimated coefficients in a wavelet basis, but not by linear shrinkage (Donoho and Johnstone, 1995; Johnstone and Silverman, 2005). This phenomenon is due to the fact that at certain crucial resolution levels, the wavelet coefficients of such functions belong to small ℓ_p balls, only for some $p < 2$, and linear estimators in such small ℓ_p balls are not of minimax rate under the ℓ_2 loss (Donoho and Johnstone, 1994; Johnstone, 1994).

The advantages of threshold estimators over the linear ones demonstrate that the possible gain of compounding statistical decision problems may not fully materialize if one confines to a small parametric class of procedures. The objective of the original empirical Bayes approach (Robbins, 1951, 1956), called general empirical Bayes (Robbins, 1983), is to approximate the performance of an oracle Bayes rule. This oracle Bayes rule provides the minimum compound risk among all separable procedures, or equivalently applications of deterministic decision functions to all data points. Since threshold and linear estimators are all separable, the general empirical Bayes aims at a smaller benchmark risk than those of the linear and threshold approaches. Consequently, sharper results can be achieved in principle in the general empirical Bayes approach, provided a sufficiently accurate estimate of the oracle Bayes rule or its prior. In nonparametric regression, sharp adaptive minimax estimation in classes of regression functions with inhomogeneous smoothness are attained by general empirical Bayes, but not by linear or threshold methods (Zhang, 2005).

Nonparametric regression demonstrates the advantages of the linear shrinkage, adaptive threshold and general empirical Bayes approaches, but the core of the problems is still the estimation of a high-dimensional vector. The focus of this paper is the estimation in ℓ_p balls under the ℓ_q loss. In the difficult case of $p < q$, (rate or sharp) adaptive minimax estimation is achieved with threshold (Donoho and Johnstone, 1995; Johnstone and Silverman, 2004; Abramovich et al, 2006) and general empirical Bayes methods (Zhang, 1997; Brown and Greenshtein, 2009; Jiang and Zhang, 2009; Zhang, 2009) when the radius of the ℓ_p balls falls into certain ranges. The limitation on the range of adaptive estimation is not only due to the difficulty of finding accurate adaptive threshold levels or estimates of the oracle prior, but also to an incomplete understanding of the minimax risk in very small ℓ_p balls. This second gap is closed in this paper.

In the rest of the paper, we state and further discuss the main results in Section 2 and provide their proofs in Section 3.

The following notation will be used throughout the paper: For $\boldsymbol{x} = (x_1, \dots, x_n) \in \mathbb{R}^n$, $\|\boldsymbol{x}\|_{p,n} = (n^{-1} \sum_{i=1}^n |x_i|^p)^{1/p}$ is the length-normalized ℓ_p norm, with the usual extension to $0 < p < 1$, and $\|\boldsymbol{x}\|_{0,n} = \#\{i : x_i \neq 0\}/n$; $\|Z\|_p = (E|Z|^p)^{1/p}$ is the L_p norm of random variables Z; $x \wedge y = \min(x,y)$, $x \vee y = \max(x,y)$ and $x_+ = x \vee 0$; $\lfloor x \rfloor$ is the largest integer lower bound of x, and $\lceil x \rceil$ is the smallest integer upper bound of x; $a_n \approx b_n$ means $a_n = (1+o(1))b_n$, and $a_n \ll b_n$ means $a_n/b_n \to 0$; $E[X;A] = \int_A X dP$ for all random variables X and events A; $t(\boldsymbol{x}) = (t(x_1), \ldots, t(x_n))$ for maps $t(\cdot)$ in $\mathbb{R}$.

2. Minimax risk for sparse vectors

Suppose we observe a multivariate normal vector $\boldsymbol{X} \sim N(\boldsymbol{\theta}, \boldsymbol{I}_n)$ under $P_{\boldsymbol{\theta}}$ with an unknown $\boldsymbol{\theta} \in \mathbb{R}^n$. The ℓ_q risk of an estimator $\boldsymbol{\delta} = \boldsymbol{\delta}(\boldsymbol{x}) = (\delta_1(\boldsymbol{x}), \ldots, \delta_n(\boldsymbol{x}))$ is

$$R_{q,n}(\boldsymbol{\theta}, \boldsymbol{\delta}) = E_{\boldsymbol{\theta}} \big\| \boldsymbol{\delta}(\boldsymbol{X}) - \boldsymbol{\theta} \big\|_{q,n}^q = \frac{1}{n} \sum_{i=1}^n E_{\boldsymbol{\theta}} \big| \delta_i(\boldsymbol{X}) - \theta_i \big|^q.$$

For vector classes $\Theta \subset \mathbb{R}^n$, the maximum and minimax ℓ_q risks are respectively

$$R_{q,n}(\Theta, \boldsymbol{\delta}) = \sup_{\boldsymbol{\theta} \in \Theta} R_{q,n}(\boldsymbol{\theta}, \boldsymbol{\delta}), \quad \mathscr{R}_{q,n}(\Theta) = \inf_{\boldsymbol{\delta} \in \mathscr{B}} R_{q,n}(\Theta, \boldsymbol{\delta}),$$

where $\mathscr{B}$ is the collection of all estimators (all Borel maps in $\mathbb{R}^n$). Our objective is to study the minimax ℓ_q risk in ℓ_p balls. Since the simpler case of $p > q$ is well understood (Donoho and Johnstone, 1994; Johnstone, 1994), we confine our investigation to $0 \leq p \leq q$, $q \geq 1$ and $C > 0$. These restrictions on (q, p, C) are in effect throughout the sequel without explicitly repeating the statement. We consider the strong and weak ℓ_p balls in separate subsections.

2.1. *Strong ℓ_p balls*

The strong ℓ_p ball with a length-normalized radius C is

$$\Theta_{p,C,n} = \{\boldsymbol{\theta} : \|\boldsymbol{\theta}\|_{p,n} \leq C\}.$$

Since $\Theta_{p,C,n}$ is commonly referred to as the (regular) ℓ_p ball, we omit the word "strong" in the rest of the paper.

For threshold levels $\lambda > 0$, let $s_\lambda(x) = \mathrm{sgn}(x)(|x| - \lambda)_+$ be the soft-threshold estimator and $h_\lambda(x) = xI\{|x| > \lambda\}$ be the hard threshold estimator. Consider ℓ_p balls with $p > 0$ and small length-normalized radii $C_n^p \to 0$. Let $\lambda_n = \sqrt{2\log(1/C_n^p)}$. Under the additional condition $nC_n^p/\lambda_n^p \to \infty$, Donoho and Johnstone (1994) proved

$$\mathscr{R}_{q,n}(\Theta_{p,C_n,n}) \approx R_{q,n}(\Theta_{p,C_n,n}\,,\, s_{\lambda_n}) \approx C_n^p \lambda_n^{q-p}, \tag{2.1}$$

and analogous results for $p = 0$ and the hard threshold estimator at a slightly larger threshold level. This requires the radii of the ℓ_p ball be small, but not too small. Theorem 1 below removes this additional condition by covering all small ℓ_p balls. Moreover, Theorem 1 is uniform in (q, p, C).

For $0 < C < 1$, define

$$r_{q,n}(\Theta_{p,C,n}) = \max_{\boldsymbol{\theta} \in \Theta_{p,C,n}} \sum_{i=1}^n \frac{|\theta_i|^q \wedge \lambda^q}{n} = \begin{cases} \{(m-1)\lambda^q + \mu^q\}/n, & p > 0 \\ m\lambda^q/n, & p = 0, \end{cases} \tag{2.2}$$

where $\lambda > 0$, $0 \leq \mu \leq \lambda$ and integer $m \geq 1$ are functions of (p, C, n) given by

$$\begin{cases} \lambda = \sqrt{2\log(1/C^p)},\ m = \lceil nC^p/\lambda^p \rceil,\ \mu = nC^p - (m-1)\lambda^p, & \text{if } p > 0 \\ m = \lfloor nC \rfloor,\ \lambda = \sqrt{2\log(n/m)},\ \mu = 0, & \text{if } p = 0. \end{cases} \tag{2.3}$$

Theorem 1. *Let $p' = p$ for $p > 0$, $p' = 1$ for $p = 0$, $1 \le q^* < \infty$ and $\eta_n \to 0+$. Then,*

$$\sup_{q,p,C} \left\{ \Big| \frac{\mathscr{R}_{q,n}(\Theta_{p,C,n})}{r_{q,n}(\Theta_{p,C,n})} - 1 \Big| : q \le q^*,\ C^{p'} \le \eta_n \right\} \to 0, \tag{2.4}$$

and at the threshold level $\lambda = \lambda_{p,C} = \sqrt{2\log(1/C^{p'})}$,

$$\sup_{q,p,C} \left\{ \Big| \frac{\mathscr{R}_{q,n}(\Theta_{p,C,n})}{R_{q,n}(\Theta_{p,C,n}, s_\lambda)} - 1 \Big| : q \le q^*,\ \frac{1}{n} \le C^{p'} \le \eta_n \right\} \to 0. \tag{2.5}$$

Moreover, (2.5) holds for the hard threshold estimator $h_{\lambda'}$ at the threshold level $\lambda' = \lambda + \|Z\|_{2q} + (q+1)(\log\lambda)/\lambda$, where $Z \sim N(0,1)$.

Theorem 1 asserts that the minimax ℓ_q risk in small ℓ_p balls is uniformly approximated by (2.2) and achieved with threshold estimators. The optimal threshold estimator is 0 (i.e. $\lambda = \infty$) for $C^{p'} < 1/n$. The approximation can be viewed as a discrete version of (2.1), since $r_{q,n}(\Theta_{p,C_n,n}) \approx C_n^p \lambda_n^{q-p}$ for $nC_n^p/\lambda_n^p \to \infty$. For $nC_n^p/\lambda_n^p = O(1)$, this discretization is necessary. For $nC_n^p \ll 1$, $C_n^p\lambda_n^{q-p}$ does not approximate the minimax risk since $r_{q,n}(\Theta_{p,C_n,n}) = C_n^q n^{q/p-1}$ is of smaller order.

The approximation (2.1) was proved using the risk of a Bayes estimator with an i.i.d. prior as a lower bound. This requires a weak law of large numbers for Bernoulli$(n, C_n^p/\lambda_n^p)$ variables, or equivalently $nC_n^p/\lambda_n^p \to \infty$, to ensure that the parameter vector is in the target class $\Theta_{p,C_n,n}$ with large prior probability.

The proof of Theorem 1 also uses the risk of a Bayes estimator as a lower bound for the minimax risk, but the prior is the uniform distribution over all permutations of a sparse vector in $\Theta_{p,C,n}$. Thus, the parameter vector is in $\Theta_{p,C,n}$ with prior probability one. However, the Bayes risk becomes more complicated. Our approximation of the Bayes risk turns out to require an upper bound for the posterior probability mass at zero. Since this upper bound is of independent interest, we state it in a proposition under a general distributional assumption in Subsection 2.3.

2.2. Weak ℓ_p balls

In this subsection we consider ℓ_q risk in the (Marcinkiewicz) weak ℓ_p balls

$$\Theta^*_{p,C,n} = \left\{ \boldsymbol{\theta} : \max_k |\theta|_{(k)} (k/n)^{1/p} \le C \right\},\ 0 < p < q,$$

where $|\theta|_{(1)} \ge \cdots \ge |\theta|_{(n)}$ are the ordered values of $|\theta_i|$.

Let $\lambda_n = \sqrt{2\log(1/C_n^p)}$. For $p < q = 2$ Johnstone (1994) proved

$$\mathscr{R}_{2,n}(\Theta^*_{p,C_n,n}) \approx R_{2,n}(\Theta^*_{p,C_n,n}, s_{\lambda_n}) \approx \{2/(2-p)\} C_n^p \lambda_n^{2-p} \tag{2.6}$$

under the condition $C_n^p \to 0$ and $nC^p/\lambda^p \gg (\log n)^3$, and an analogous result for the hard threshold estimator. Theorem 2 below extends his results to general $C^p \to 0$ and q.

For $0 < C < 1$, let $\lambda = \sqrt{2\log(1/C^p)}$ and define

$$r_{q,n}(\Theta^*_{p,C,n}) = \max_{\boldsymbol{\theta}\in\Theta^*_{p,C,n}} \sum_{i=1}^n \frac{|\theta_i|^q \wedge \lambda^q}{n} = \frac{1}{n}\sum_{k=1}^n \min\left\{ C^q (n/k)^{q/p}, \lambda \right\}. \tag{2.7}$$

Theorem 2. *Let $0 < p < q < \infty$ be fixed and $\eta_n \to 0+$. Then,*

$$\sup_{C^p \le \eta_n} \left| \frac{\mathscr{R}_{q,n}(\Theta^*_{p,C,n})}{r_{q,n}(\Theta^*_{p,C,n})} - 1 \right| \to 0, \tag{2.8}$$

and at the threshold level $\lambda = \lambda_{p,C} = \sqrt{2\log(1/C^p)}$,

$$\sup_{n^{-1} \le C^p \le \eta_n} \left| \frac{\mathscr{R}_{q,n}(\Theta^*_{p,C,n})}{R_{q,n}(\Theta^*_{p,C,n}, s_\lambda)} - 1 \right| \to 0. \tag{2.9}$$

Moreover, (2.9) holds for the hard threshold estimator $h_{\lambda'}$ at the threshold level $\lambda' = \lambda + \|Z\|_{2q} + q(\log \lambda)/\lambda$, where $Z \sim N(0,1)$.

We actually proved the uniformity of (2.8) and (2.9) in (q,p,C) for $q-p \ge \epsilon_* > 0$ and $q \le q^* < \infty$. However, for very small ℓ_p balls, an extension of our proof to $q-p \to 0+$ involves the posterior probability for a uniform prior on permutations of a vector with a greater number of small nonzero elements than the case covered by our method.

2.3. Posterior for a uniform prior over permutations of a sparse parameter vector

Let $f(x|\theta)$ be a family of density functions with respect to a certain σ-finite measure $\nu(dx)$. Let E_n be probability measures under which

$$\boldsymbol{X}|\boldsymbol{\Theta} \sim \prod_{i=1}^{n} f(x_i|\Theta_i), \quad P_n\Big\{\boldsymbol{\Theta} = (c_{i_1}, \dots, c_{i_n})\Big\} = \frac{1}{n!}, \tag{2.10}$$

for all permutations $\{i_1, \dots, i_n\} = \{1, \dots, n\}$, where $c_1, \dots, c_n$ are fixed with $c_{m+1} = \cdots = c_n = 0$ for a certain $1 \le m < n$. Define

$$\xi_M = \frac{1}{m} \sum_{i=1}^{m} \int \big\{f(x|c_i) - Mf(x|0)\big\}_+ \nu(dx).$$

Proposition 1. *Let $n > m \ge 1$ and E_n be as in (2.10) with $c_{m+1} = \cdots = c_n = 0$. Suppose $f(x|c_j)$ are absolutely continuous with respect to $f(x|0)$, $j \le m$. Then,*

$$\begin{aligned} & E_n\Big(P_n\{\Theta_1 \ne 0|\boldsymbol{X}\}\Big)^2 \\ \le\ & \frac{m}{n}\Big(\frac{Mm}{n-m+1} + \xi_M\Big)\Big(1 + \frac{Mm}{n-m+1} + \xi_M\Big)^{m-1}, \quad \forall M > 0. \end{aligned} \tag{2.11}$$

Recently, Greenshtein and Ritov (2009) used the uniform prior on the permutations to derive an approximate risk equivalence between the class of separable decision rules and the more general class of all permutation invariant decision rules. Although their result does not require a large proportion of zero components with the unknown vector, it does not apply to the worst case scenario in ℓ_p balls where the maximum magnitude of the components diverges to infinity.

3. Proofs

Proof of Proposition 1. Let $h_j(x) = f(x|c_j)/f(x|0)$. For $0 \le k_1 \le k_2 \le k_3 \le n$ with $k_1 \le m < n$, let $k = (m - k_1) \wedge (k_3 - k_2)$ and define

$$S(\boldsymbol{X}; k_1, k_2, k_3) = \frac{(m-k_1-k)!}{m!} \sum_{(i_1,\ldots,i_{k_1+k}) \in \Lambda_1(k_1,k_2,k_3)} \sum_{(j_1,\ldots,j_k)\in\Lambda_2(k_2,k_3)} h_{i_1}(X_1)\cdots h_{i_{k_1}}(X_{k_1}) h_{i_{k_1+1}}(X_{j_1})\cdots h_{i_{k_1+k}}(X_{j_k}),$$

where $\Lambda_1(k_1,k_2,k_3)$ is the set of all vectors of distinct indices $\{i_1,\ldots,i_{k_1+k}\} \subset \{1,\ldots,m\}$ and $\Lambda_2(k_2,k_3)$ is the set of all vectors of strictly ordered indices $k_2 < j_1 \le \cdots \le j_k \le k_3$. Note that the average is taken over $\Lambda_1(k_1,k_2,k_3)$ but the summation, not the average, is taken over $\Lambda_2(k_2,k_3)$. Let $J = E_n\big(P_n\{\Theta_1 \ne 0\big|\boldsymbol{X}\}\big)^2$ as in (2.11). By (2.10) and the exchangeability of $X_1,\ldots,X_n$,

$$\begin{aligned} J &= \binom{n}{m}^{-1} E_0 \frac{S^2(\boldsymbol{X};1,1,n)}{S(\boldsymbol{X};0,0,n)} \\ &= \binom{n}{m}^{-1}\binom{n-1}{m-1} E_0 S(\boldsymbol{X};m,m,m)\frac{S(\boldsymbol{X};1,1,n)}{S(\boldsymbol{X};0,0,n)} \\ &= \frac{m}{n}\sum_{k=1}^{m}\binom{m-1}{k-1}\binom{n-m}{m-k} J_k, \end{aligned} \tag{3.1}$$

where $J_k = E_0 S(\boldsymbol{X};m,m,m)S(\boldsymbol{X};k,m,2m-k)/S(\boldsymbol{X};0,0,n)$ for products sharing k indices in the numerator. Let $\widetilde{h}_i(x) = (h_i(x) - M)_+$. Since $h_i(x) \le M + \widetilde{h}_i(x)$,

$$\begin{aligned} J_k &= E_0 \frac{S(\boldsymbol{X};k,n-m+k,n)S(\boldsymbol{X};k,m,2m-k)}{S(\boldsymbol{X};0,0,n)} \\ &\le E_0 \frac{S(\boldsymbol{X};k,n-m+k,n)S(\boldsymbol{X};k,m,2m-k)}{S(\boldsymbol{X};0,0,n-m+k)} \\ &= \sum_{\{i_1,\ldots,i_k\}\subset\{1,\ldots,m\}} E_0 \frac{h_{i_1}(X_1)\cdots h_{i_k}(X_k)S(\boldsymbol{X};k,m,2m-k)}{\{m!/(m-k)!\}S(\boldsymbol{X};0,0,n-m+k)} \\ &\le \sum_{j=0}^{k}\binom{k}{j}M^{k-j}E_0\frac{Y_j S(\boldsymbol{X};m,m,m)}{S(\boldsymbol{X};0,0,n-m+k)} \\ &= \sum_{j=0}^{k}\binom{k}{j}M^{k-j}E_0 Y_j\binom{n-m+k-j}{m-j}^{-1}\frac{S(\boldsymbol{X};j,j,n-m+k)}{S(\boldsymbol{X};0,0,n-m+k)}, \end{aligned} \tag{3.2}$$

where $Y_j = \sum_{\{i_1,\ldots,i_j\}\subset\{1,\ldots,m\}} \widetilde{h}_{i_1}(X_1)\cdots\widetilde{h}_{i_j}(X_j)/\{m!/(m-j)!\}$. Let $\{\widetilde{\xi}_1,\ldots,\widetilde{\xi}_j\}$ be a random subset of $\{E_0\widetilde{h}_i(X_1), i \le m\}$. Since the sum for $S(\boldsymbol{X};j,j,n-m+k)$ is taken over a smaller index set that the sum for $S(\boldsymbol{X};0,0,n-m+k)$,

$$E_0 Y_j \frac{S(\boldsymbol{X};j,j,n-m+k)}{S(\boldsymbol{X};0,0,n-m+k)} \le E_0 Y_j = E_0 \prod_{i=1}^{j}\widetilde{\xi}_i \le (E_0\widetilde{\xi}_1)^j = \xi_M^j$$

due to the negative correlation between $\widetilde{\xi}_j$ and $\prod_{i=1}^{j-1}\widetilde{\xi}_i$. Moreover,

$$\binom{n-m}{m-k}\binom{n-m+k-j}{m-j}^{-1} \le m^{k-j}/(n-m+1)^{k-j},\ 0 \le j \le k \le m.$$

Inserting the above upper bounds into (3.2) and then (3.2) to (3.1), we find

$$\begin{aligned} J &\leq \frac{m}{n}\sum_{k=1}^{m}\binom{m-1}{k-1}\binom{n-m}{m-k}\sum_{j=0}^{k}\binom{k}{j}M^{k-j}\binom{n-m+k-j}{m-j}^{-1}\xi_M^j \\ &\leq \frac{m}{n}\sum_{k=1}^{m}\binom{m-1}{k-1}\sum_{j=0}^{k}\binom{k}{j}\Big(\frac{Mm}{n-m+1}\Big)^{k-j}\xi_M^j \\ &= \frac{m}{n}\Big(\frac{Mm}{n-m+1}+\xi_M\Big)\Big(1+\frac{Mm}{n-m+1}+\xi_M\Big)^{m-1}. \end{aligned}$$

This completes the proof. □

The following lemma provides upper bound for the ℓ_q risk of threshold estimators.

Lemma 1. *Let $\theta \in \mathbb{R}$, $Z \sim N(0,1)$ and $\|Z\|_p = (E|Z|^p)^{1/p}$.*
(i) Let $s_\lambda(x) = \mathrm{sgn}(x)(|x|-\lambda)_+$ be the soft threshold estimator. Then,

$$E\big|s_\lambda(Z+\theta)-\theta\big|^q \leq \min\Big\{|\theta|^q, E|Z-\lambda|^q\Big\} + 2\Gamma(q+1)\varphi(\lambda)/\lambda^{q+1}.$$

(ii) Let $h_\lambda(x) = xI\{|x|\leq\lambda\}$ be the hard threshold estimator. Then,

$$\begin{aligned} E\big|h_\lambda(Z+\theta)-\theta\big|^q &\leq \min\big\{(1+e^{-\lambda^2/16})|\theta|^q, (\lambda+q/\lambda)^q + \|Z\|_q^q\big\} \\ &\quad + E\big[|Z|^q; |Z| > \lambda - \|Z\|_{2q}\big]. \end{aligned}$$

Proof. Let $\varphi(x) = e^{-x^2/2}/\sqrt{2\pi}$. Since

$$\frac{\partial}{\partial\theta}\int\Big|s_\lambda(x+\theta)-\theta\Big|^q\varphi(x)dx = q|\theta|^{q-1}\int_{-\lambda-\theta}^{\lambda-\theta}\varphi(x)dx \in [0, q\theta^q]$$

for $\theta > 0$ and $R_q(\theta,\lambda)$ is an even function of θ,

$$E\big|s_\lambda(Z+\theta)-\theta\big|^q \leq \min\Big\{|\theta|^q, E|Z-\lambda|^q\Big\} + E(|Z|-\lambda)_+^q.$$

This proves part (i) since $E(Z-\lambda)_+^q/\varphi(\lambda) = \int_0^\infty x^q e^{-\lambda x - x^2/2}dx \leq \Gamma(q+1)/\lambda^{q+1}$.

For the hard threshold estimator,

$$\begin{aligned} E\big|h_\lambda(Z+\theta)-\theta\big|^q &= |\theta|^q + E\big[|Z|^q - \theta^q; |Z+\theta| > \lambda\big] \\ &\leq |\theta|^q + E\big[|Z|^q; |Z| > \theta \vee (\lambda-\theta)_+\big]. \end{aligned}$$

Since $E\big[|Z|^q; |Z| > \theta\vee(\lambda-\theta)_+\big] \leq |\theta|^q\sqrt{P\{|Z|>\lambda/2\}}$ for $|\theta| > \|Z\|_{2q}$,

$$E\big|h_\lambda(Z+\theta)-\theta\big|^q \leq |\theta|^q + E\big[|Z|^q; |Z| > \lambda - \|Z\|_{2q}\big] + |\theta|^q e^{-\lambda^2/16}. \tag{3.3}$$

For $\theta > \lambda + q/\lambda$, $(\partial/\partial\theta)\log\{(\theta/\lambda)^q e^{-(\theta-\lambda)^2/2}\} = q/\theta - \theta + \lambda < 0$, so that

$$\begin{aligned} E\big|h_\lambda(Z+\theta)-\theta\big|^q &= E|Z|^q + E(|\theta|^q - |Z|^q)I\{|Z+\theta|<\lambda\} \\ &\leq E|Z|^q + |\theta|^q P\{|Z| > \theta-\lambda\} \\ &\leq E|Z|^q + |\theta|^q e^{-(\theta-\lambda)^2/2} \\ &\leq E|Z|^q + (\lambda+q/\lambda)^q. \qquad (3.4) \end{aligned}$$

The combination of (3.3) and (3.4) yields part (ii). □

Proof of Theorem 1. We first treat the case $p > 0$. Let $0 < \eta_n^* < 1$ be the solution of

$$\frac{n\eta_n^*}{\{\log(1/\eta_n^*)\}^{q^*+3}} = 1. \tag{3.5}$$

Since (2.1) holds uniformly in (q, p, C) for $q < q^*$ and $\eta_n^* \le C^p \le \eta_n$, we assume $\eta_n \le \eta_n^*$ without loss of generality.

Taking the average of the ℓ_q loss, we find by Lemma 1 (i) that

$$R_{q,n}(\boldsymbol{\theta}, s_\lambda) \le \frac{1}{n}\sum_{i=1}^{n} \min\Big\{|\theta_i|^q, E|Z-\lambda|^q\Big\} + 2\Gamma(q+1)\varphi(\lambda)/\lambda^{q+1}.$$

Let $\{\lambda, m, \mu\}$ be functions of (p, C, n) as in (2.3). Since $\lambda \ge \sqrt{2\log(1/\eta_n)} \to \infty$ and $\varphi(\lambda) = C^p/\sqrt{2\pi}$, $E|Z-\lambda|^q \approx \lambda^q$ and

$$\mathscr{R}_{q,n}(\Theta_{p,C,n}, s_\lambda) \le (1+o(1)) \max_{\boldsymbol{\theta}\in\Theta_{p,C,n}} \sum_{i=1}^{n} \frac{|\theta_i|^q \wedge \lambda^q}{n} + O(C^p/\lambda) \tag{3.6}$$

uniformly in (q, p, C) for $q \le q^*$ and $C^p \le \eta_n$. Since $|x|^q$ is convex in $|x|^p$, the maximum in (3.6) is attained with as many points $|\theta_i| = \lambda$ as possible. This is the solution with $\boldsymbol{\theta} = \boldsymbol{c}$, where

$$c_1 = \cdots = c_{m-1} = \lambda, \;\; c_m = \mu, \;\; c_{m+1} = \cdots = c_n = 0. \tag{3.7}$$

For this solution, the maximum in (3.6) is $\sum_{i=1}^n c_i^q/n = r_q(\Theta_{p,C,n})$. For $m > 1$, $2r_q(\Theta_{p,C,n}) \ge (m/n)\lambda^q \ge C^p\lambda^{q-p} \gg C^p/\lambda$. For $m = 1$, $1 \le \mu^p = nC^p \le \lambda^p$, so that $r_q(\Theta_{p,C,n}) = \mu^q/n = C^p\mu^{q-p} \gg C^p/\lambda$. Thus, by (3.6)

$$\sup_{0<p\le q\le q^*, C^p\le\eta_n} \Big(\frac{R_{q,n}(\Theta_{p,C,n}, s_{\lambda_{p,C}})}{r_{q,n}(\Theta_{p,C,n})} - 1\Big)_+ \to 0. \tag{3.8}$$

Now we derive a lower bound for $p > 0$. Let $0 < \gamma < 1$ and P_n be a probability under which $\boldsymbol{X} \sim N(\boldsymbol{\theta}, \boldsymbol{I}_n)$ conditionally on $\boldsymbol{\Theta} = \boldsymbol{\theta}$ and $\boldsymbol{\Theta}$ is a uniform random vector over the permutations of $\gamma\boldsymbol{c}$ with the $\boldsymbol{c}$ in (3.7). For vectors $\boldsymbol{w} = \{w_1, w_2, w_3\}$ with $w_j \ge 0$ and $w_1 + w_2 + w_3 = 1$, define

$$f_q(\lambda, \mu, \boldsymbol{w}) = \min_a \Big\{\big|a-\lambda\big|^q w_1 + \big|a-\mu\big|^q w_2 + |a|^q w_3\Big\}.$$

The minimum above is always attained at an a satisfying $|a|^q w_3 \le |\lambda|^q w_1 + |\mu|^q w_2$. Thus, since $0 \le \mu \le \lambda$, for $w_1 + w_2 \le \epsilon_1$

$$f_q(\lambda, \mu, \boldsymbol{w}) \ge \begin{cases} \big|\lambda(1-\epsilon_2)\big|^q w_1 + \big|\mu - \lambda\epsilon_2\big|^q w_2, & w_1 > 0 \\ \big|\mu(1-\epsilon_2)\big|^q w_2, & w_1 = 0, \end{cases} \tag{3.9}$$

where $\epsilon_2 = \{\epsilon_1/(1-\epsilon_1)\}^{1/q}$. Since $(X_i, \Theta_i), i \le n$, are exchangeable random vectors under E_n, the marginal ℓ_q risk of the Bayes estimator $\widehat{\boldsymbol{\Theta}}$ is

$$E_n\big\|\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}\big\|_{q,n}^q = E_n|\widehat{\Theta}_1 - \Theta_1|^q = E_n f_q(\gamma\lambda, \gamma\mu, \widehat{\boldsymbol{w}}), \tag{3.10}$$

where $\widehat{\boldsymbol{w}} = (\widehat{w}_1, \widehat{w}_2, \widehat{w}_3)$ gives the posterior probabilities at $\{\gamma\lambda, \gamma\mu, 0\}$ for Θ_1.

Since $\mathbf{\Theta}$ is a random permutation of $\gamma\boldsymbol{c}$ with the $\boldsymbol{c}$ in (3.7), $E_n\widehat{w}_1 = (m-1)/n$ and $E_n\widehat{w}_2 = 1/n$. Let $\{\epsilon_1,\epsilon_2\}$ be related small constants as in (3.9) and define

$$\epsilon = \epsilon(\gamma\lambda, m, \gamma\mu) = (n/m)E_n(\widehat{w}_1+\widehat{w}_2)^2. \tag{3.11}$$

The Chebyshev inequality gives $E_n\big[\widehat{w}_1+\widehat{w}_2;\widehat{w}_1+\widehat{w}_2>\epsilon_1\big] \le (m/n)\epsilon/\epsilon_1$. It follows from (3.9), (3.10) and the above calculations of the posterior that for $m>1$

$$\begin{aligned} E_n\big\|\widehat{\mathbf{\Theta}}-\mathbf{\Theta}\big\|_{q,n}^q &\ge \gamma^q E_n\Big[\big|\lambda(1-\epsilon_2)\big|^q\widehat{w}_1 + \big|\mu-\lambda\epsilon_2\big|^q\widehat{w}_2;\widehat{w}_1+\widehat{w}_2\le\epsilon_1\Big] \\ &\ge \gamma^q\Big\{\big|\lambda(1-\epsilon_2)\big|^q E_n\widehat{w}_1 + \big|\mu-\lambda\epsilon_2\big|^q E_n\widehat{w}_2 \\ &\qquad -\lambda^q E_n\Big[\widehat{w}_1+\widehat{w}_2;\widehat{w}_1+\widehat{w}_2>\epsilon_1\Big]\Big\} \\ &\ge \gamma^q\Big\{\big|\lambda(1-\epsilon_2)\big|^q(m-1)/n + \big|\mu-\lambda\epsilon_2\big|^q/n - \lambda^q(m/n)\epsilon/\epsilon_1\Big\}. \end{aligned}$$

For $m=1$, $\widehat{w}_1=0$ due to $\Theta_1\in\{0,\gamma\mu\}$, so that the second part of (3.9) gives

$$E_n\big\|\widehat{\mathbf{\Theta}}-\mathbf{\Theta}\big\|_{q,n}^q \ge \gamma^q\Big\{\big|\mu(1-\epsilon_2)\big|^q/n - \mu^q(1/n)\epsilon/\epsilon_1\Big\}.$$

Since $r_{q,n}(\Theta_{p,C,n}) = \lambda^q(m-1)/n+\mu^q/n$, we find that in either cases

$$\epsilon(\gamma\lambda,m,\gamma\mu)=o(1) \ \Rightarrow\ \frac{E_n\big\|\widehat{\mathbf{\Theta}}-\mathbf{\Theta}\big\|_{q,n}^q}{r_{q,n}(\Theta_{p,C,n})} \ge \gamma^q+o(1) \tag{3.12}$$

and the uniformity on the left-hand side implies the uniformity of $o(1)$ on the right.

Since $E_n\big\|\widehat{\mathbf{\Theta}}-\mathbf{\Theta}\big\|_{q,n}^q \le \mathscr{R}_{q,n}(\Theta_{p,C,n}) \le R_{q,n}(\Theta_{p,C,n},s_\lambda)$, it remains to prove that the $\epsilon(\gamma\lambda,m,\gamma\mu)$ in (3.11) is uniformly small for $0<p\le q\le q^*$ and $C^p\le\eta_n$ for each fixed $0<\gamma<1$, where $\{\lambda,m,\mu\}$ are functions of (p,C,n) as in (2.3).

By Proposition 1 and (3.11),

$$\begin{aligned} \epsilon(\gamma\lambda,m,\gamma\mu) &= \frac{n}{m}E_n\Big(P_n\{\Theta_1\neq 0\big|\boldsymbol{X}\}\Big)^2 \\ &\le \Big(\frac{Mm}{n-m+1}+\xi_M\Big)\Big(1+\frac{Mm}{n-m+1}+\xi_M\Big)^{m-1},\ \forall M>0, \end{aligned}$$

where $\xi_M = \int\{\varphi(x-\gamma\lambda)-M\varphi(x)\}_+dx(m-1)/m + \int\{\varphi(x-\gamma\mu)-M\varphi(x)\}_+dx/m$ in view of (3.7). With the $\boldsymbol{c}$ in (3.7) define

$$\lambda_0 = \max_{i\le n}\Big(c_i\vee\sqrt{2\log\log n}\Big) = \begin{cases}\lambda, & m>1\\ \mu\vee\sqrt{2\log\log n}, & m=1.\end{cases} \tag{3.13}$$

We pick $M=\exp(\gamma\lambda_0^2/2)$. Since $\varphi(x-\gamma\lambda_0)/\varphi(x) = e^{x\gamma\lambda_0-\gamma^2\lambda_0^2/2}$, $\varphi(x-\gamma\lambda_0)\ge M\varphi(x)$ iff $x\gamma\lambda_0-\gamma^2\lambda_0^2/2\ge\gamma\lambda_0^2/2$, iff $x\ge(1+\gamma)\lambda_0/2$. It follows that

$$\begin{aligned} \xi_M &\le \int\{\varphi(x-\gamma\lambda_0)-M\varphi(x)\}_+dx \\ &\le P\{N(\gamma\lambda_0,1)>(1+\gamma)\lambda_0/2\} \le e^{-(1-\gamma)^2\lambda_0^2/8}. \end{aligned}$$

For $m>1$, $\lambda=\lambda_0$, $M=C^{-\gamma p}$ and $\xi_M = C^{(1-\gamma)^2p/4}$. Thus, since $m=\lceil nC^p/\lambda^p\rceil$,

$$\begin{aligned} &\epsilon(\gamma\lambda,m,\gamma\mu) \\ &\le \Big(\frac{2C^{(1-\gamma)p}}{\lambda^p-C^p}+C^{(1-\gamma)^2p/4}\Big)\exp\Big\{\frac{nC^p}{\lambda^p}\Big(\frac{2C^{(1-\gamma)p}}{\lambda^p-C^p}+C^{(1-\gamma)^2p/4}\Big)\Big\} \\ &= (1+o(1))\eta_n^{(1-\gamma)^2}\exp\Big\{(1+o(1))n\eta_n^{1+(1-\gamma)^2/4}\Big\} \end{aligned} \tag{3.14}$$

uniformly over the (p, C, n). This provides the left side of (3.12) uniformly, since $\eta_n \le \eta_n^*$ with the η_n^* in (3.5). For $m = 1$ and $\mu > \sqrt{2\log\log n}$, $\lambda_0 = \mu \le \lambda$, $M \le e^{\gamma\lambda^2/2} \le C^{-\gamma p}$ and $\xi_M \le e^{-(1-\gamma)^2(\log\log n)/4}$. It follows that

$$\epsilon(\gamma\lambda, m, \gamma\mu) \le \frac{M}{n} + \xi_M \le \frac{C^{(1-\gamma)p}}{\mu^p} + (\log n)^{-(1-\gamma)^2/4} \to 0. \tag{3.15}$$

For $m = 1$ and $\mu \le \sqrt{2\log\log n}$, $\lambda_0 = \sqrt{2\log\log n}$, $M = (\log n)^\gamma$ and $\xi_M \le e^{-(1-\gamma)^2(\log\log n)/4}$, so that

$$\epsilon(\gamma\lambda, m, \gamma\mu) \le (\log n)^\gamma/n + (\log n)^{-(1-\gamma)^2/4} \to 0. \tag{3.16}$$

Thus, in all three cases, we proved $\epsilon(\gamma\lambda, m, \gamma\mu) = o(1)$ uniformly for the (p, C, n) under consideration. This completes the proof in the case of $p > 0$ in view of (3.8) and (3.12), since $\mathscr{R}_{q,n}(\Theta) \le R_{q,n}(\Theta, s_\lambda)$ for all $\Theta \subset \mathbb{R}^n$.

For $p = 0$, $\boldsymbol{\theta}$ allows at more $\lfloor nC \rfloor$ nonzero entries, so that C is effectively m/n. The same proof follows through since the situation is simpler with $\mu = 0$. This completes the proof of the theorem for the soft threshold estimator.

Finally, we consider the hard threshold estimator. At the threshold level $\lambda' = \lambda + \|Z\|_{2q} + q(\log\lambda)/\lambda$, the upper bound in Lemma 1 (ii) gives

$$E\big|h_{\lambda'}(Z+\mu) - \mu\big|^q \le (1+o(1))(|\mu|^q \wedge \lambda^q) + O(\lambda^{-1}e^{-\lambda^2/2})$$

uniformly for $\lambda \ge \sqrt{2\log(1/\eta_n)} \to \infty$. Thus, (3.6) and then (3.8) follows with s_λ replaced by $h_{\lambda'}$. This completes the proof of the entire theorem. □

Proof of Theorem 2. Since $\Theta^*_{p,C,n} \supset \Theta_{p,C,n}$, the upper bound

$$\begin{aligned} \mathscr{R}_{q,n}(\Theta^*_{p,C,n}) &\le (1+o(1)) \max_{\boldsymbol{\theta}\in\Theta^*_{p,C,n}} \sum_{i=1}^n \frac{|\theta_i|^q \wedge \lambda^q}{n} + O(C^p/\lambda) \\ &= (1+o(1)) r_{q,n}(\Theta^*_{p,C,n}) \end{aligned}$$

is still valid. Thus, it suffices to derive the lower bound. We still assume (3.5) in view of (2.6). The least favorable configuration is attained in (2.7) with

$$r_{q,n}(\Theta^*_{p,C,n}) = \sum_{j=1}^n \frac{c_j^q}{n}, \; c_1 = \cdots = c_{m-1} = \lambda, \; c_j = \frac{\mu}{j^{1/p}}, \; j = m, \ldots, n,$$

where $\mu = n^{1/p}C$ and $m = \lceil nC^p/\lambda^p \rceil$. For fixed $0 < \gamma < 1$ and integer $k > 1$, let E_n be the probability under which $\boldsymbol{X}|\boldsymbol{\Theta} \sim N(\boldsymbol{\Theta}, \boldsymbol{I}_n)$ and $\boldsymbol{\Theta} = (\Theta_1, \ldots, \Theta_n)$ is a random permutation of $\gamma(c_1, \ldots, c_{km}, 0 \ldots, 0)$. The proof of Theorem 1 gives

$$(1+o(1))\gamma^q \sum_{j=1}^{km} \frac{c_j^q}{n} \le \mathscr{R}(\Theta^*_{p,C,n}) \le (1+o(1)) \sum_{j=1}^n \frac{c_j^q}{n}$$

uniformly for all fixed $0 < \gamma < 1$ and integer $k > 1$. Since $c_j = \mu/j^{1/p}$ for $j > km$,

$$\sum_{j=km+1}^n \frac{c_j^q}{j^{q/p}} \le \frac{\mu^q/(km)^{q/p-1}}{q/p-1} = \frac{m\mu^q/m^{q/p}}{k^{q/p-1}(q/p-1)} \le \frac{\sum_{j=1}^m c_j^q}{k^{q/p-1}(q/p-1)}.$$

Thus, $\mathscr{R}(\Theta^*_{p,C,n})$ must be within an infinitesimal fraction of $r_{q,n}(\Theta^*_{p,C,n})$. □

References

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

Brown, L.D. (1966). On the admissibility of invariant estimators of one or more location parameters *Ann. Math. Statist.* **37** 1087–1136.

Brown, L.D. (1971). Admissible estimators, recurrent diffusions, and insoluble boundary value problems. *Ann. Math. Statist.* **42** 855–903.

Brown, L.D. and Greenshtein, E. (2009). Nonparametric empirical Bayes and compound decision approaches to estimation of a high-dimensional vector of normal means. *Ann. Statist.* **37** 1685–1704.

Donoho, D.L. and Johnstone, I.M. (1994). Minimax risk over ℓ_p-balls for ℓ_q-error. *Probab. Theory Related Fields* **99** 277–303.

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

Efromovich, S. and Pinsker, M.S. (1984). An adaptive algorithm for nonparametric filtering. *Automat. Remote Control* **11** 58–65.

Efromovich, S. and Pinsker, M.S. (1986). An adaptive algorithm for minimax nonparametric estimation of a spectral density. *Problems Inform. Transmission* **22** 62–76.

Efron, B. and Morris, C. (1972). Empirical Bayes on vector observations: An extension of Stein's method. *Biometrika* **59** 335–347.

Efron, B. and Morris, C. (1973). Stein's estimation rule and its competitors–an empirical Bayes approach. *J. Amer. Statist. Assoc.* **68** 117–130.

Greenshtein, E. and Ritov, Y. (2009). Asymptotic efficiency of simple decisions for the compound decision problem. In *Optimality: The Third Erich L. Lehmann Symposium*, Javier Rojo, ed. *IMS Lecture NotesMonograph Series* **57** 266–275.

James, W. and Stein, C. (1961). Estimation with quadratic loss. *Proc. Fourth Berkeley Symp. Math. Statist. Probab.* **1** 361–379. Univ. of California Press, Berkeley.

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

Johnstone, I.M. (1994). Minimax Bayes, asymptotic minimax and sparse wavelet priors. In *Statistical Decision Theory and Related Topics V* (S. Gupta and J. Berger, eds.) 303–326. Springer, New York. MR1286310

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

Johnstone, I.M. and Silverman, B.W. (2005). Empirical Bayes selection of wavelet thresholds. *Ann. Statist.* **33** 1700–1752.

Robbins, H. (1951). Asymptotically subminimax solutions of compound statistical decision problems. *Proc. Second Berkeley Symp. Math. Statist. Probab.* **1** 131–148. Univ. of California Press, Berkeley.

Robbins, H. (1956). An empirical Bayes approach to statistics. *Proc. Third Berkeley Symp. Math. Statist. Probab.* **1** 157–163. Univ. of California Press, Berkeley.

Robbins, H. (1983). Some thoughts on empirical Bayes estimation. *Ann. Statist.* **11** 713–723.

Stein, C. (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. *Proc. Third Berkeley Symp. Math. Statist. Probab.* **1** 157–163. Univ. of California Press, Berkeley.

Strawderman, W. E. (1971). Proper Bayes estimators of the multivariate normal mean. *Ann. Math. Statist.* **42** 385–388. MR0397939

Strawderman, W. E. (1973). Proper Bayes minimax estimators of the multivariate normal mean for the case of common unknown variances. *Ann. Math. Statist.* **44** 1189–1194. MR0365806

Zhang, C.-H. (1997). Empirical Bayes and compound estimation of normal means. *Statistica Sinica* **7** 181–193.

Zhang, C.-H. (2003). Compound decision theory and empirical Bayes method. *Ann. Statist.* **33** 379–390.

Zhang, C.-H. (2005). General empirical Bayes wavelet methods and exactly adaptive minimax estimation. *Ann. Statist.* **33** 54–100.

Zhang, C.-H. (2009). Generalized maximum likelihood estimation of normal mixture densities. *Statistica Sinica* **19** 1297–1318.