# Model Selection and Sharp Asymptotic Minimaxity

Zheyang Wu[1] and Harrison H. Zhou[2]
Worcester Polytechnic Institute and Yale University

December 13, 2011

**Abstract**

We obtain sharp minimax results for estimation of an $n-$dimensional normal mean under quadratic loss. The estimators are chosen by penalized least squares with a penalty that grows like $ck\log(n/k)$, for $k$ equal to the number of nonzero elements in the estimating vector. For a wide range of sparse parameter spaces, we show that the penalized estimator achieves the exact minimax rate with the correct multiplication constant if and only if $c$ equals 2. Our results unify the theory obtained by many other authors for penalized estimation of normal means. In particular we establish that a conjecture by Abramovich, Benjamini, Donoho and Johnstone (2006) is true.

**Keywords:** FDR; minimax estimation; model selection; multiple comparisons; sharp asymptotic minimaxity; smoothing parameter selection; thresholding; wavelet denoising; wavelets.

**AMS 2000 Subject Classification:** Primary 62G08, Secondary 62G20.

[1] Department of Mathematical Sciences, Worcester Polytechnic Institute, Worcester, MA 01609. The research of Zheyang Wu was supported in part by NIH Grant GM590507

[2] Department of Statistics, Yale University, New Haven, CT 06511. The research of Harrison Zhou was supported in part by NSF Grants DMS-0645676 and DMS- 0854975.

# 1 Introduction

Consider the standard multivariate normal mean problem

$$y_i = \theta_i + \sigma_n z_i,\ i = 1, ..., n, \tag{1}$$

where $\sigma_n$ is the noise level, and $z_i$'s are independent standard normal variables. The goal is to estimate the unknown mean $\{\theta_i\}$ based on sample $\{y_i\}$.

In light of recent advances of high dimensional estimation, we assume that the parameter set $\{\theta_i\}$ has a sparse structure for which its definition varies in literature but the essence can be captured by considering situations where most of the unknown coordinates $\theta_i$ take the value 0 or very close to 0. This normal mean estimation problem is essential to wavelet Gaussian regression (cf. Donoho and Johnstone (1994a, 1994b, 1995, 1998)). It is of independent interest as well such as in microarray data analysis (cf. Efron (2003)). Recent advances of asymptotic equivalence theory showed that nonparametric Gaussian regression captures the essence of many nonparametric estimation problems. See for example Brown and Low (1996), Nussbaum (1996), Golubev, Nussbaum and Zhou (2009), Brown, Cai and Zhou (2008, 2009), and Cai and Zhou (2009). Therefore the problem of normal mean estimation is fundamentally important for general nonparametric estimation too.

For sparse parameter estimation it is natural to consider model selection procedures. Many influential model selection procedures have been proposed in literature such as AIC, BIC and RIC. The AIC model selection procedure was proposed in Akaike (1973, 1974). It had a great influence in statistical practice, but was not accepted for a while in our field. From the hypothesis testing point of view, the AIC procedure rejects $\theta_i = 0$ when $y_i^2 \geq 2$, which is equivalent to testing whether each $\theta_i$ is zero or not at the level of test 16%. When $\{\theta_i\}$ is sparse and $n$ is large, AIC is too aggressive in the sense that it tends to select too many spurious non-zero $\theta_i$. Donoho and Johnstone (1994a) and George and Foster (1994) proposed the RIC procedure independently to avoid such a problem. The RIC procedure is equivalent to Bonferroni correction in multiple comparisons. Usually Bonferroni correction is too conservative in multiple comparisons so that too many unknown coordinates $\theta_i$ are estimated by zero.

Benjamini and Hochberg (1995) proposed a new multiple comparison procedure called FDR which is less aggressive than AIC and less conservative than RIC. Abramovich, Benjamini, Donoho and Johnstone (2006) showed the FDR procedure is asymptotically sharp minimax, in the sense that it achieves optimal rates and constants in minimax

sense adaptively over a wide range of sparse parameter spaces. This celebrated work built an important and possibly productive connection between multiple hypotheses testing and sparse signal estimation. In that paper it was observed that the FDR procedure is closely connected to a penalized estimator with a penalty approximately $2k\log(n/k)$ for an actual model size $k$. This type of penalty has arisen naturally in several areas including information theory, empirical Bayes and model complexity. See for example Foster and Stine (1999), George and Foster (2000), Birgé and Massart (2001) etc. It was then conjectured in Abramovich, Benjamini, Donoho and Johnstone (2006) that the penalized estimation procedure with a simple penalty $2k\log(n/k)$ is asymptotically sharp minimax. However Benjamini and Gavrilov (2009) showed that for some examples the finite sample performance of FDR procedure is similar to CIC by Tibshirani and Knight (1999) for which the penalty is approximately $4k\log(n/k)$. In Abramovich, Grinshtein and Pensky (2007) some Bayesian model selection procedures are proposed with penalties approximately $ck\log(n/k)$ for some $c > 2$.

It is thus desirable to have a unified study of asymptotic risk properties of those penalized estimation which are approximately $ck\log(n/k)$. When a penalty is sufficiently close to $2k\log\frac{n}{k}$ as $n \to \infty$, we show it achieves sharp asymptotic minimaxity adaptively over a wide range of sparse parameter spaces. As a consequence we solve the conjecture 1.2 in Abramovich, Benjamini, Donoho and Johnstone (2006, page 597) which states as follows. Let $\Theta_{n,p}(\eta_n)$ be a sparse (weak) $l_p$ ball defined in (9) or (10) in Section 3, and $R_n(\Theta_{n,p}(\eta_n))$ be the minimax risk under the squared error loss. Define $\mathrm{Pen}(\theta) = 2k\log\frac{n}{k}$ where $k = \|\theta\|_0$. For two sequences $a_n$ and $b_n$ we denote $a_n \sim b_n$ if $a_n = (1 + o(1))\, b_n$. It was conjectured that

$$\sup_{\theta\in\Theta_{n,p}(\eta_n)} \mathbb{E}\left\|\hat{\theta} - \theta\right\|_2^2 \sim R_n\left(\Theta_{n,p}(\eta_n)\right),$$

where

$$\hat{\theta} = \mathrm{argmin}\left\{\|y - \theta\|_2^2 + \sigma_n^2 \mathrm{Pen}(\theta)\right\}.$$

Apparently the search range of model sizes can not cover $n$, otherwise we have to choose model size $k = n$ for which $\mathrm{Pen}(k) = 0$ and consequently the minimum of the objective function is 0. We will restrict the range of model size to be $k \le n/\log n$, which is acceptable when the parameter space is sparse. This restriction is equivalent to define $\mathrm{Pen}(\theta) = 2k\log\frac{n}{k}$ when $\|\theta\|_0 = k \le n/\log n$ and $+\infty$ otherwise. For a penalized procedure approximately $ck\log\frac{n}{k}$ with $c > 2$ we show its risk differs from the risk of minimax estimator with a constant factor. However for $c < 2$, it can be shown that the ratio of the risk of the corresponding penalized estimator with the minimax risk tends to $\infty$.

The paper is organized as follows. In Section 2 we introduce model selection procedures. Main theoretical results are given in Section 3. In Section 4 we consider a new range of penalty functions in more general parameter spaces and discuss the relations to other works. The proofs of theorems and lemmas are given in Section 5.

## 2 Penalized Estimation Procedures

In this section we introduce various penalty functions and the corresponding estimation procedures. Their risk properties are given in Section 3.

There has been an enormous amount of work in statistics to study penalized estimation. Let $\theta = (\theta_1, \theta_2, \ldots, \theta_n)$, $y = (y_1, y_2, \ldots, y_n)$ and $\mathrm{Pen}(\theta)$ be the penalty function. The penalized estimator $\hat{\theta}$ is the minimizer of the following objective function

$$K(\theta, y) = \|y - \theta\|_2^2 + \mathrm{Pen}(\theta). \tag{2}$$

Without loss of generality we assume that $\sigma_n = 1$ (see Remark 1 for general $\sigma_n$). Here is a short list of some classical penalized procedures including AIC with $\mathrm{Pen}(\theta) = 2\|\theta\|_0$, BIC with $\mathrm{Pen}(\theta) = (\log n)\|\theta\|_0$, RIC with $\mathrm{Pen}(\theta) = 2(\log n)\|\theta\|_0$, and Ridge regression with $\mathrm{Pen}(\theta) = \lambda\|\theta\|_2^2$ and LASSO with $\mathrm{Pen}(\theta) = \lambda\|\theta\|_1$ for some $\lambda > 0$ (usually it is challenging to pick a practical $\lambda$ for LASSO when the true signal is sparse). An exhaustive survey of penalty functions is beyond our scope.

In the last decade much progress has been made in the area of model selection to adaptively estimate sparse signals. Several penalized procedures for model selection have been proposed from different aspects. Let $k = \|\theta\|_0$. We use $C$ to denote a generic constant, which may vary from places to places. The following penalties $\mathrm{Pen}(\theta)$ were proposed in the past decade:

**(A).** $2\sum_{i=1}^{k} \log \frac{n}{i}$ in Foster and Stine (1999) from an information-theoretic point of view;

**(B).** Approximately $2\sum_{i=1}^{k} \log\left(\frac{n+1}{i} - 1\right)$ in George and Foster (2000) from an empirical Bayes approach;

**(C).** Approximately $ck \log \frac{n}{k}$ with $c > 2$ ($c = 4$ was recommended) in Birgé and Massart (2001) from model complexity;

**(D).** $4\sum_{i=1}^{k} \log \frac{n}{i}$ in Tibshirani and Knight (1999) from a covariance inflation criterion;

**(E).** $\sum_{i=1}^{k} \left(\bar{\Phi}^{-1}\left(\frac{q_n i}{2n}\right)\right)^2$ with a small $q_n$ in Abramovich, Benjamini, Donoho and Johnstone (2006) from a false discovery rate control approach;

**(F).** $2k \log \frac{n}{k}$ in Abramovich, Benjamini, Donoho and Johnstone (2006) as a conjecture;

**(G).** Approximately $ck \log \frac{n}{k}$ with $c > 2$ in Abramovich, Grinshtein and Pensky (2007) from a Bayes approach.

The advances above motivate us to give a unified theory for all these penalized procedures. For all penalties listed above, they depend only on the model size. From now on we write the penalty as $\text{Pen}\left(\|\theta\|_0\right)$. Let

$$\hat{\theta} = \arg\min_{\theta} K(\theta, y) = \arg\min_{\theta} \left[ \|y - \theta\|_2^2 + \text{Pen}\left(\|\theta\|_0\right) \right]. \tag{3}$$

For a model size $k = \|\theta\|_0$, the penalty is $\text{Pen}\,(k)$, and it is easy to see that the smallest residual sum squares for this model size is $\sum_{i=k+1}^{n} y_{[i]}^2$ where $y_{[1]}^2 \geq ... \geq y_{[n]}^2$. Then we have

$$\begin{aligned} \min_{\theta} K(\theta, y) &= \min_{k} \min_{\left\{\theta : \|\theta\|_0 = k\right\}} \left[ \|y - \theta\|_2^2 + \text{Pen}\,(k) \right] \\ &= \min_{k} \left[ \sum_{i=k+1}^{n} y_{[i]}^2 + \text{Pen}\,(k) \right]. \end{aligned} \tag{4}$$

Define

$$\hat{k} = \arg\min_{k} \left[ \sum_{i=k+1}^{n} y_{[i]}^2 + \text{Pen}\,(k) \right], \tag{5}$$

which attains the minimum of equation (4). The penalized estimators $\hat{\theta}$, i.e., the global minimizer of $K(\cdot, y)$, is just a hard thresholding rule with

$$\hat{\theta}_i = y_i I\left\{ |y_i| \geq |y|_{[\hat{k}]} \right\}.$$

We set $\text{Pen}\,(0) = 0$, and $\hat{\theta}_i = 0$ if $\hat{k} = 0$.

We consider a type of penalties as follows. Define

$$\text{Pen}\,(k) = \sum_{i=1}^{k} u_{ni}^2, \;\; u_{ni}^2 = \begin{cases} c_{ni} \log \frac{n}{i}, & i \leq n/\log n \\ +\infty, & i > n/\log n \end{cases}, \tag{6}$$

where $c_{ni} \to c \geq 2$ uniformly over $i \leq n/\log n$ as $n \to \infty$. This definition is equivalent to restrict the search range of the model size $k$ be within $0 \leq k \leq n/\log(n)$, i.e., $\hat{k} = \arg\min_{0 \leq k \leq n/\log(n)} \left[ \sum_{i=k+1}^{n} y_{[i]}^2 + \text{Pen}\,(k) \right]$. Additionally when $c = 2$ we require that the penalty function satisfies

$$2 \log \frac{n}{i} - (1 - \varepsilon) \log \log \frac{n}{i} \leq c_{ni} \log \frac{n}{i}, \tag{7}$$

for some $0 < \varepsilon < 1$. In the scenario of asymptotics, the $\log\log$ term in (7) is related to the second order approximation to the magnitude of the $i$th order statistics of $n$ i.i.d. $N(0,1)$. When $k$ is not very large, e.g., $0 \le k \le n/\log n$, it is fairly easy to show that for such penalties we have $\mathrm{Pen}\left(\|\theta\|_0\right) \sim \tilde{c}_{nk} k \log \frac{n}{k}$ for some $\tilde{c}_{nk} \to c \ge 2$ uniformly over $k \le n/\log n$ as $n \to \infty$, which is essentially an $l_0$ penalty of the form $\lambda \|\theta\|_0$ with $\lambda$ data driven.

Among those penalties defined in equation (6), a simple class is $\mathrm{Pen}(k) = ck\log(n/k)$ with $c \ge 2$, which can be argued as follows. We may write $\mathrm{Pen}(k) = \sum_{i=1}^{k} (\mathrm{Pen}(i) - \mathrm{Pen}(i-1))$. For $k \le \frac{n}{\log n}$, it is easy to see the penalty term $ck\log(n/k)$ can be written as $ck\log(n/k) = \sum_{i=1}^{k} u_{ni}^2$, where

$$u_{ni}^2 = \begin{cases} c\log n, & \text{for } i = 1 \\ c\log\left(\frac{n}{i}\right) + c(i-1)\log\left(\frac{i-1}{i}\right), & \text{for } k \ge i > 1 \end{cases}. \tag{8}$$

Note that $c(i-1)\log\left(\frac{i-1}{i}\right) = c(i-1)\log\left(1-\frac{1}{i}\right) \to -c$ as $i \to \infty$. The sequence $c(i-1)\log\left(\frac{i-1}{i}\right)$ is then a bounded sequence. Write $c\log\left(\frac{n}{i}\right) + c(i-1)\log\left(\frac{i-1}{i}\right) = c_{ni}\log\frac{n}{i}$, where in this case

$$c_{ni} = c + \frac{c(i-1)\log\left(\frac{i-1}{i}\right)}{\log\left(\frac{n}{i}\right)}.$$

For $i \le \frac{n}{\log n}$, we have $n/i \ge \log n \to \infty$ as $n \to \infty$, then $c_{ni} \to c$ uniformly over $i \le n/\log n$ as $n \to \infty$.

Thus penalties defined in (6) cover all penalties from (A) to (G) when $k \le n/\log n$. In particular, the penalty $2k\log\frac{n}{k}$ in the conjecture of Abramovich, et al. (2006) is a very special case.

## 3 Theoretical Properties

We shall now investigate the asymptotic properties of the procedures proposed in Section 2. Our model selection procedure is based on the assumption that the underlying structure of the unknown true parameter is sparse. We study the theoretical properties of our procedures over the (weak) $l_p$ balls which is by now standard for sparse signals estimation. More specifically, we assume that $\theta$ is in one of the following balls:

- $l_p$ balls:

$$l_p[\eta_n] = \left\{\theta \in \mathbb{R}^n : \frac{1}{n}\sum_{i=1}^{n} |\theta_i|^p \le \eta_n\right\}, \tag{9}$$

where $0 \leq p < 2$. When $p = 0$, we denote $0^0 = 0$ for this specific definition of parameter space that constrains the percentage of nonzero $\theta_i$ be no more than $\eta_n$. When $0 < p < 2$, the parameter space constrains the overall magnitude of $\theta$.

- $m_p$ (weak $l_p$) balls:

$$m_p[\eta_n] = \left\{\theta \in \mathbb{R}^n : |\theta|_{[k]} \leq \left(\eta_n \frac{n}{k}\right)^{1/p}, k = 1, ..., n\right\}, \tag{10}$$

where $0 < p < 2$, constrains the rate of ordered $|\theta|_{[k]}$ with $|\theta|_{[1]} \geq ... \geq |\theta|_{[n]}$. Note that $l_p[\eta_n] \subset m_p[\eta_n]$ because for $\theta \in l_p[\eta_n]$ we have $|\theta|^p_{[k]} \leq \frac{\sum|\theta_i|^p}{k} \leq \eta_n \frac{n}{k}$.

The parameter space consisting of an $l_p$ or $m_p$ ball will be denoted by $\Theta_{n,p}(\eta_n)$ for simplicity. We assume that $\eta_n$ satisfies the following condition

$$\eta_n \in \left[b_1 n^{-1} \log^\gamma n, b_2 n^{-b_3}\right], \tag{11}$$

where $b_1 > 0, b_2 > 0, 1 > b_3 > 0$, and $\gamma > 4.5$ for a technical reason. Under sparsity assumptions for $p < 2$ and $\eta_n \to 0$, the minimax risks over $l_p$ or $m_p$ balls, i.e., $R_n(\Theta_{n,p}(\eta_n)) = \inf_{\hat{\theta}} \sup_{\theta \in \Theta_{n,p}(\eta_n)} E_\theta \left\|\hat{\theta} - \theta\right\|_2^2$ were studied in Donoho et al. (1992), Johnstone (1994), and Donoho and Johnstone (1994b). Under the condition (11), it has been shown that the minimax risk for the $l_p$ ball is

$$R_n(l_p[\eta_n]) \sim k_n \tau^2_{\eta_n} = n\eta_n \left(2 \log \eta_n^{-1}\right)^{(2-p)/2}, \tag{12}$$

where

$$k_n = n\eta_n \tau_{\eta_n}^{-p}, \tau_{\eta_n} = \left(2 \log \eta_n^{-1}\right)^{1/2}, \tag{13}$$

and for the $m_p$ ball

$$R_n(m_p[\eta_n]) \sim \frac{2}{2-p} R_n(l_p[\eta_n]). \tag{14}$$

We first consider the case $c > 2$ for penalties in (6) with $c_{ni} \to c$ uniformly over $i \leq n/\log n$, as $n \to \infty$. The corresponding penalized procedures are shown failing to achieve sharp asymptotic minimaxity by missing the optimal constant.

**Theorem 1** *Consider penalties defined in (6) with $c_{ni} \to c > 2$ uniformly over $i \leq n/\log n$ as $n \to \infty$. Let the parameter space $\Theta_{n,p}(\eta_n)$ be an $l_p$ or $m_p$ ball with $\eta_n$ defined in (11). Then the corresponding penalized estimation procedure $\hat{\theta}$ defined in (3) satisfies*

$$\sup_{\theta \in \Theta_{n,p}(\eta_n)} \mathbb{E} \left\|\hat{\theta} - \theta\right\|_2^2 \sim c^* R_n(\Theta_{n,p}(\eta_n)),$$

*where $c^* = \left(\frac{c}{2}\right)^{1-p/2}$.*

For the case $c = 2$ the following theorem shows that a range of model selection procedures achieve sharp asymptotic minimaxity adaptively over these sparse spaces.

**Theorem 2** *Consider penalties defined in (6) with*

$$2\log\frac{n}{i} - (1-\varepsilon)\log\log\frac{n}{i} \le u_{ni}^2 \le (2+\delta_n)\log\frac{n}{i},$$

*for some constant* $0 < \varepsilon < 1$ *and a sequence* $\delta_n \to 0$. *Let the parameter space* $\Theta_{n,p}(\eta_n)$ *be an* $l_p$ *or* $m_p$ *ball with* $\eta_n$ *defined in (11). Then the corresponding penalized estimation procedure* $\hat{\theta}$ *defined in (3) satisfies*

$$\sup_{\theta\in\Theta_{n,p}(\eta_n)} \mathbb{E}\left\|\hat{\theta}-\theta\right\|_2^2 \sim R_n\left(\Theta_{n,p}(\eta_n)\right).$$

As discussed in Section 2, the penalty term $ck\log(n/k)$ with $c \ge 2$ is one of the procedures considered in Theorems 1 or 2. We then immediately have the following result.

**Corollary 1** *Consider penalties* $\mathrm{Pen}(k) = ck\log\left(\frac{n}{k}\right)$, $c \ge 2$ *when* $k \le n/\log n$, *and* $\mathrm{Pen}(k) = \infty$ *when* $k > n/\log n$. *Let the parameter space* $\Theta_{n,p}(\eta_n)$ *be an* $l_p$ *or* $m_p$ *ball with* $\eta_n$ *defined in (11). Then*

$$\sup_{\theta\in\Theta_{n,p}(\eta_n)} \mathbb{E}\left\|\hat{\theta}-\theta\right\|_2^2 \sim c^* R_n\left(\Theta_{n,p}(\eta_n)\right),$$

*where* $c^* = \left(\frac{c}{2}\right)^{1-p/2}$.

When $c = 2$, we have $c^* = 1$. Thus the conjecture 1.2 in Abramovich, Benjamini, Donoho and Johnstone (2006) is solved as a consequence. For penalties approximately $ck\log\left(\frac{n}{k}\right)$ with $c < 2$ as $n \to \infty$, the convergence rate of the corresponding penalized procedure is no longer optimal as implied by the following theorem.

**Theorem 3** *Consider penalties defined in (6) with* $c_{ni} \to c < 2$ *uniformly over all* $i$. *Let the parameter space* $\Theta_{n,p}(\eta_n)$ *be an* $l_p$ *or* $m_p$ *ball with* $\eta_n$ *defined in (11). Then the corresponding penalized estimation procedure* $\hat{\theta}$ *defined in (3) satisfies*

$$\sup_{\theta\in\Theta_{n,p}(\eta_n)} \mathbb{E}\left\|\hat{\theta}-\theta\right\|_2^2 \ge C\frac{n}{\log n}\log\log n, \tag{15}$$

*for some* $C > 0$.

Note that $\frac{n}{\log n}(\log\log n)/R_n\left(\Theta_{n,p}(\eta_n)\right) \to \infty$. Thus for $c < 2$ the corresponding penalized procedure is not rate optimal in $l_p$ or $m_p$ ball with $\eta_n$ defined in (11).

**Remark 1** *The above results can be extended to general variance $\sigma_n \neq 1$ case. We can rescale the parameter by $\theta/\sigma_n$ in the parameter spaces in (9) and (10). The penalty functions in (6) and (8) need to add a factor $\sigma_n^2$.*

The proofs of results in this section are given in Section 5.

# 4 Discussion

We discuss three topics in this section. First, we consider a new range of penalty functions, for which our main results still hold, but the restriction of the searching range of the model size is now removed. Then, we discuss some results for more general parameter spaces that include very sparse and dense cases. Finally, we comment on some other related works.

## 4.1 Full range of model searching

The penalty function defined in (6) restricts the searching range of the model size be within $\left[0, \frac{n}{\log n}\right]$. In the following we consider a class of penalties which lead the penalized model selection procedure to minimax estimation without restricting the searching range. The penalty function is

$$\mathrm{Pen}\,(k) = ck \log\left(e^d n/k\right) \text{ with } c \geq 2, d \geq 0, \tag{16}$$

It can be rewritten as $\mathrm{Pen}\,(k) = \sum_{i=1}^{k} (\mathrm{Pen}\,(i) - \mathrm{Pen}\,(i-1))$, such that the penalty $ck \log\left(e^d n/k\right) = \sum_{i=1}^{k} u_{ni}'^2$, where

$$u_{ni}'^2 = \begin{cases} c\left[\log n + d\right], & \text{for } i = 1 \\ c\left[\log\left(\frac{n}{i}\right) + d + (i-1)\log\left(\frac{i-1}{i}\right)\right], & \text{for } n \geq k \geq i > 1 \end{cases}.$$

For the penalty class (16), we have the following result.

**Proposition 1** *Consider penalties in (16) with $c \geq 2$ and $d \geq \frac{5.37}{c} + 1$. Let the parameter space $\Theta_{n,p}(\eta_n)$ be an $l_p$ or $m_p$ ball with $\eta_n$ defined in (11). We have*

$$\sup_{\theta \in \Theta_{n,p}(\eta_n)} \mathbb{E}\left\|\hat{\theta} - \theta\right\|_2^2 \sim c^* R_n\left(\Theta_{n,p}(\eta_n)\right),$$

*where $c^* = \left(\frac{c}{2}\right)^{1-p/2}$.*

Note that when $c = 2$, the above proposition indicates that the penalty function $2k \log\left(e^d n/k\right)$ leads the penalized model selection procedure to the sharp asymptotically minimax estimation if $d \geq 3.685$. The proposition can be proved based on Remark 3 in Section 5.4.8. Specifically, we can show that the selected model size is $o\left(\frac{n}{\log n}\right)$ with high probability (see Remark 3). Then the rest of proof is similar to that for Theorems 1 and 2.

Penalty function (16) is closely related to the penalty function $\mathrm{Pen}'(k) = c_1 k + c_2 \log\binom{n}{k}$, which was considered in Yang and Barron (1998) and Yang (1999). Proposition 1 immediately implies the following corollary. See Section 5.4.9 for details.

**Corollary 2** *Let $\hat{\theta}$ be the penalized estimation of procedure (3) with $\mathrm{Pen}'(k) = c_1 k + c_2 \log\binom{n}{k}$ with $c_2 \geq 2$ and $c_1 > 5.37 + c_2$. Under condition (11), we have*

$$\sup_{\theta \in \Theta_{n,p}(\eta_n)} \mathbb{E} \left\| \hat{\theta} - \theta \right\|_2^2 \sim c^* R_n\left(\Theta_{n,p}(\eta_n)\right),$$

*where $c^* = \left(\frac{c}{2}\right)^{1-p/2}$.*

## 4.2 More general parameter spaces

In this section, we consider an $l_p[\eta_n]$ ball with $0 < p < \infty$ and $\eta_n \geq \frac{a}{n}$ for a constant $a > 0$, which is more general than the $l_p$ ball studied in Section 3. For two sequences $a_n$ and $b_n$ we denote $a_n \asymp b_n$ if there exist constants $C_2 \geq C_1 > 0$ such that $C_1 \leq a_n/b_n \leq C_2$. It has been proved that the minimax risk over the general $l_p$ ball is

$$R_n\left(l_p[\eta_n]\right) \asymp \begin{cases} (n\eta_n)^{2/p} & \text{if } \eta_n \leq n^{-1}(1+\log n)^{p/2} \\ n\eta_n\left(1+\log \eta_n^{-1}\right)^{1-p/2} & \text{if } n^{-1}(1+\log n)^{p/2} < \eta_n \leq \tau \\ n & \text{if } \eta_n \geq \tau \end{cases}, \tag{17}$$

for $0 < p < 2$ and any fixed constant $\tau > 0$, and for $p \geq 2$,

$$R_n\left(l_p[\eta_n]\right) \asymp \begin{cases} n\eta_n^{2/p} & \text{if } \eta_n \leq \tau \\ n & \text{if } \eta_n \geq \tau \end{cases},$$

(cf. Johnstone (2011, Theorem 11.7)).

We can show that a family of $ck \log \frac{e^d n}{k}$ type penalty procedures are still rate optimal for these general $l_p$ balls.

**Proposition 2** *Consider a general $l_p[\eta_n]$ ball with $0 < p < \infty$ and $\eta_n \geq \frac{a}{n}$ for any constant $a > 0$. For penalties defined in (16) with $c > 2$ and $d > \frac{c}{c-2} + \frac{1}{2}$, the corresponding penalized estimation procedure $\hat{\theta}$ defined in (3) satisfies*

$$\sup_{\theta \in l_p[\eta_n]} \mathbb{E} \left\| \hat{\theta} - \theta \right\|_2^2 \leq C R_n \left( l_p[\eta_n] \right). \tag{18}$$

**Proposition 3** *Consider a general $l_p[\eta_n]$ ball with $0 < p < \infty$ and $\eta_n = \tau$ for any $\tau > 0$. For penalties defined in (16) with $c \geq 2$ and $d > 1$, the corresponding penalized estimation procedure $\hat{\theta}$ defined in (3) satisfies*

$$\sup_{\theta \in l_p[\eta_n]} \mathbb{E} \left\| \hat{\theta} - \theta \right\|_2^2 \leq C R_n \left( l_p[\eta_n] \right). \tag{19}$$

Both propositions follow easily from Theorems 11.3 and 11.7 of Johnstone (2011), whose proof and formulation can be traced back to Barron, Birgé, and Massart (1999) and Birgé and Massart (2001). Specifically, the penalties in (16) can be written as the penalties considered by Johnstone (2011), i.e., $Pen(k) = \zeta k \left(1 + 2\sqrt{L_{n,k}} + L_{n,k}\right)$, where $\zeta > 1$ and $L_{n,k} = 2(\log \frac{n}{k} + \gamma_{n,k})$ with $\gamma_{n,k} \geq \gamma > 1$. For example, we can set $\zeta = \frac{c+2}{4} > 1$ to obtain the conditions on $c$ and $d$ in Proposition 2.

Propositions 2 and 3 can be similarly extended to the penalties $\mathrm{Pen}'(k) = c_1 k + c_2 \log \binom{n}{k}$ considered in Yang and Barron (1998) and Yang (1999) by connecting $\mathrm{Pen}'(k)$ with the penalties in (16).

**Remark 2** *First, consider a general $l_p[\eta_n]$ ball with $0 < p < \infty$ and $\eta_n \geq \frac{a}{n}$ for any constant $a > 0$. For penalties $\mathrm{Pen}'(k) = c_1 k + c_2 \log \binom{n}{k}$ with $c_2 > 2$ and $c_1 > \frac{c_2^2}{c_2 - 2} + \frac{c_2}{2}$, the corresponding penalized estimation procedure $\hat{\theta}$ defined in (3) satisfies (18). Second, consider a general $l_p[\eta_n]$ ball with $0 < p < \infty$ and $\eta_n = \tau$ for any $\tau > 0$. For penalties $\mathrm{Pen}'(k)$ with $c_1 > 2$ and $c_2 \geq 2$, the corresponding penalized estimation procedure $\hat{\theta}$ defined in (3) satisfies (19).*

We observe that it is hard to extend the general methodology of this paper to very sparse cases, e.g., $\eta_n = n^{-1} (\log n)^{\delta}$ with $\delta < p/2$ and $0 < p < 2$. Our methodology utilizes the bound $\mathbb{E} \left\| \theta - \hat{\theta} \right\|_2^2 \leq K(\theta_0, \theta) + \mathbb{E} 2 \left\langle z, \hat{\theta} - \theta \right\rangle$ and hopes that $K(\theta_0, \theta)$ is the dominant term. However, it can be shown (cf. the proof of Lemma 3) that $\sup_{\theta \in \Theta_{n,p}(\eta_n)} K(\theta_0, \theta) \leq C R_n \left( \Theta_{n,p}(\eta_n) \right)$ does not hold, since $R_n \left( \Theta_{n,p}(\eta_n) \right) \asymp (n\eta_n)^{2/p}$ and

$$\sup_{\theta \in \Theta_{n,p}(\eta_n)} K(\theta_0, \theta) / (n\eta_n)^{2/p} = C (\log n)^{1 - p/2 + (1 - 2/p)\delta} \to \infty.$$

### 4.3 Relations to some other works

Barron, Birgé and Massart (1999) and Birgé and Massart (2001) studied very general models for penalties corresponding to the $c > 2$ case. They obtained rate optimality. It remains an open problem to extend their results to the $c = 2$ case. In this paper, we consider a simpler normal mean problem and give explicit constants for risks when $c \geq 2$, which implies sharp minimaxity for $c = 2$. For the case $c < 2$, Theorem 3 can be extended to penalties defined in (16)). That is, we have

$$\sup_{\theta \in l_p[\eta_n]} \mathbb{E} \left\| \hat{\theta} - \theta \right\|_2^2 \geq C \frac{n}{\log n} \log \log n.$$

Both our result and Birgé and Massart (2007) indicate that penalty functions with $c < 2$ are not optimal in sparse parameter spaces. But they are different in at least two aspects. First, Birgé and Massart (2007) only considered $l_0$ ball with size at most at an order of $n^\alpha$, $0 < \alpha < 1$. So their results do not directly cover the entire $l_p$ ball nor $m_p$ ball. Second, our conclusions are different. Birgé and Massart (2007) established a risk lower bound of order $n^\alpha \log n$, which is smaller than $\frac{n}{\log n} \log \log n$.

## 5 Proofs

In this section we first give technical lemmas in Section 5.1, then a brief outline of the proof of Theorems 1-3 in Section 5.2. Details of the proofs for the theorems and the lemmas are followed in Sections 5.3-5.4.

### 5.1 Technical lemmas

Lemma 1 establishes that for a specific $\theta$ defined in either $l_p$ ball or $m_p$ ball, the $ck \log(n/k)$ type penalty functions with $c > 2$ lead to a small model size $\hat{k}$ with high probability. Lemma 1 is applied to prove the lower bound for Theorem 1.

**Lemma 1** *Let $\theta$ be defined in (24) or (25) for $l_p$ ball or $m_p$ ball respectively, and $\hat{k}$ be defined in equation (5) for penalties in (6) with $c_{ni} \to c > 2$ uniformly over all $i \leq n/\log n$. We have*

$$\mathbb{P}\left(\hat{k} \leq \gamma_n \tilde{k}_n\right) \to 1,$$

*where $\gamma_n = \frac{1}{\log \log n}$ and $\tilde{k}_n = n\eta_n \left(c \log \eta_n^{-1}\right)^{-p/2}$.*

Lemma 2 says that for any $\theta$ in $l_p$ ball or $m_p$ ball, the size $\hat{k}$ of the selected model by penalties defined in (6) is properly upper bounded. Lemma 2 is used in proving both upper and lower bounds in Theorems 1 and 2.

**Lemma 2** *Let $\hat{k}$ be defined in equation (5) for penalties considered in Theorems 1 and 2. Then*

$$\mathbb{P}\left(A_n^c\right) \leq C_D n^{-D}, \text{ for all } D > 0,$$

*where $A_n = \left\{\hat{k} \leq k_+\left(\theta, q_n\right)\right\}$, $k_+\left(\theta, q_n\right)$ is defined in (34) with $1 > q_n \geq \frac{1}{\sqrt{\log\log\log n}}$.*

Lemmas 3 and 4 show the upper bounds of the bias part and the error part of the risk, respectively. They are used to prove the upper bounds for Theorems 1 and 2.

**Lemma 3** *Let $\Theta_{n,p}\left(\eta_n\right)$ be one of those spaces defined in (9) and (10). Assume that the penalty function satisfies equation (6). Then the objective function defined in (2) with $\theta_0$ given in (28) satisfies*

$$\sup_{\theta \in \Theta_{n,p}(\eta_n)} K\left(\theta_0, \theta\right) \leq c^* \left(1 + o\left(1\right)\right) R_n\left(\Theta_{n,p}\left(\eta_n\right)\right), \tag{20}$$

*where $c^* = \left(\frac{c}{2}\right)^{1-p/2}$.*

**Lemma 4** *Let $\Theta_{n,p}\left(\eta_n\right)$ be one of those spaces defined in (9) and (10). Assume that the penalty function satisfies equation (6). Then the procedure $\hat{\theta}$ defined in (3) satisfies*

$$\sup_{\theta \in \Theta_{n,p}(\eta_n)} \mathbb{E}\left\langle z, \hat{\theta} - \theta \right\rangle = o\left(1\right) R_n\left(\Theta_{n,p}\left(\eta_n\right)\right). \tag{21}$$

Lemma 5 says that the selected model size $\hat{k}$ for $\theta = 0$ is large when the penalty corresponds to $c < 2$.Lemma 5 is applied to prove Theorem 3.

**Lemma 5** *Let $\theta = 0$, and $\hat{k}$ be defined in equation (5) for penalties defined in (6) with $c_{ni} \to c < 2$ uniformly over all $i \leq \frac{n}{\log n}$. Then there exists an $\varepsilon \in (0, 1)$ such that*

$$\mathbb{P}\left(\hat{k} > \varepsilon \frac{n}{\log n}\right) \to 1.$$

Lemma 6 gives the boundaries of sum of extremes from a standard normal vector, and is used to prove Lemmas 1 and 5.

**Lemma 6** *Let* $z_i \overset{i.i.d.}{\sim} N(0,1)$, $i = 1, ..., n$, $z_{[1]}^2 \geq z_{[2]}^2 \geq ... \geq z_{[n]}^2$. *For any* $k_n = o(n)$, *any* $\delta > 0$, $D > 0$, *we have*

$$\begin{aligned}
\mathbb{P}\left(\sum_{i=1}^{k_n} z_{[i]}^2 > (2+\delta) k_n \log\frac{n}{k_n} + 2(\log n)^3\right) &< C_D n^{-D}, \\
\mathbb{P}\left(\sum_{i=1}^{k_n} z_{[i]}^2 < (2-\delta) k_n \log\frac{n}{k_n} - 2(\log n)^3\right) &< C_D n^{-D}.
\end{aligned}$$

Lemma 7 is applied to prove Lemma 1.

**Lemma 7** *Let* $\theta_1$, $c_n$, $\varepsilon_n$ *and* $\tilde{k}_n$ *be defined in (24). For any* $\delta \in (0, c-2)$, $k_1$, $k_2 > 0$ *with* $k_1 + k_2 = k$, $k \in \left(\gamma_n \tilde{k}_n, C\tilde{k}_n\right)$, $\gamma_n \to \infty$, *we have*

$$k_1\theta_1^2 + (2+\delta) k_2 \log\frac{n}{k_2} \leq c_n k (1-\varepsilon_n) \log\frac{n}{\tilde{k}_n} + C\tilde{k}_n.$$

Lemma 8 shows that penalties of the type $ck\log\left(e^d n/k\right)$, $c \geq 2$, $d \geq 0$, are larger than the FDR penalty by Abramovich et al. 2006 for proper $q_n$. Lemma 8 is used for proving Lemmas 3 and 4 and Proposition 1.

**Lemma 8** *Define* $z(x) = \bar{\Phi}^{-1}(x)$ *be the upper* $(1-x)^{th}$ *percentile of standard normal distribution. Let* $q_n \geq \frac{1}{\sqrt{\log\log\log n}}$. *If* $u_{ni}^2$ *be defined in (6), we have*

$$z^2\left(\frac{iq_n}{2n}\right) \leq u_{ni}^2,$$

*and* $u_{ni}^2 = \frac{c}{2}(1+\delta_{ni}) z^2\left(\frac{iq_n}{2n}\right)$, *for some* $\delta_{ni} \to 0$ *uniformly over* $i \leq \frac{n}{\log n}$. *If* $u_{ni}'^2$ *is defined in (16) with* $d \geq \frac{5.37}{c} + 1$, *we have*

$$z^2\left(\frac{i}{2n}\right) < u_{ni}'^2,$$

*for each* $i \leq n$.

### 5.2 An outline of proofs of Theorems 1-3

Let $\hat{\theta}$ be the penalized procedure considered in Theorems 1-3. We will first prove the lower bound for Theorem 1 in Section 5.3.1. We will show that there is a specific $\theta$ in $l_p[\eta_n]$ or $m_p[\eta_n]$ such that

$$\mathbb{E}\left\|\hat{\theta} - \theta\right\|_2^2 \geq c^*(1+o(1)) R_n(\Theta_{n,p}(\eta_n)), \tag{22}$$

where $c^* = \left(\frac{c}{2}\right)^{1-\frac{p}{2}}$, $\Theta_{n,p}(\eta_n)$ is either $l_p[\eta_n]$ or $m_p[\eta_n]$ ball, $R_n(\Theta_{n,p}(\eta_n))$ is the minimax risk among all estimators given in (12) or (14) with $\eta_n$ defined in (11). Then we will prove the upper bounds in Theorems 1 and 2 in Section 5.3.2,

$$\sup_{\theta \in \Theta_{n,p}(\eta_n)} \mathbb{E} \left\| \hat{\theta} - \theta \right\|_2^2 \leq c^* (1 + o(1)) R_n (\Theta_{n,p}(\eta_n)), \tag{23}$$

with penalties considered in both theorems, where $c^* = \left(\frac{c}{2}\right)^{1-p/2}$, in particular $c^* = 1$ when $c = 2$. The lower bound (22) and the upper bound (23) imply Theorem 1. Theorem 2 for sharp asymptotic minimaxity follows immediately from the upper bound (23). For Theorem 3, we obtain the lower bound in Section 5.3.3 by considering the risk of the penalized procedure at $\theta = 0$.

## 5.3 Proofs of Theorems

### 5.3.1 Lower bounds for Theorem 1

In this section we define a specific $\theta \in \Theta_{n,p}(\eta_n)$ for proving (22). Theorem 1 follows immediately from this lower bound and the upper bound to be proved in Section 5.3.2.

Let $\varepsilon_n = 1/\log\log n$, $\tilde{k}_n = \left\lfloor n\eta_n \left(c \log \eta_n^{-1}\right)^{-p/2} \right\rfloor$ and $c_n = \min_{i \leq n/\log n} c_{ni}$. For the $l_p[\eta_n]$ ball we define

$$\theta_i = \begin{cases} \sqrt{c_n (1 - \varepsilon_n) \log \frac{n}{\tilde{k}_n}}, & i \leq \tilde{k}_n - 1 \\ 0, & i \geq \tilde{k}_n \end{cases}, \tag{24}$$

while for the $m_p[\eta_n]$ ball we set

$$\theta_i = \begin{cases} \sqrt{c_n (1 - \varepsilon_n) \log \frac{n}{\tilde{k}_n}}, & i \leq \tilde{k}_n - 1 \\ \left(\eta_n^* \frac{n}{i}\right)^{1/p}, & \tilde{k}_n \leq i < r_n \tilde{k}_n \\ 0, & i \geq r_n \tilde{k}_n \end{cases}, \tag{25}$$

where $r_n = \log\log n$ and $\eta_n^* = \min\left\{ \left[c_n (1 - \varepsilon_n) \log \frac{n}{\tilde{k}_n}\right]^{p/2} \tilde{k}_n/n, \eta_n \right\} \sim \eta_n$. It is easy to see that $\theta$ defined in (24) and (25) are truly in $l_p[\eta_n]$ ball defined in (9) and $m_p[\eta_n]$ ball defined in (10) respectively.

From the definition of $\theta$ in (24) and (25), the actual model size is more than or equal to $\tilde{k}_n$, but estimated model size is $o\left(\tilde{k}_n\right)$ by Lemma 1. The resulting loss in estimation is then expected to be $(1 - o(1)) \sum \theta_i^2 \sim c^* R_n (\Theta_{n,p}(\eta_n))$, which can be rigorously proved

as follows. Write

$$\begin{aligned}\mathbb{E}\sum_{i=1}^{n}\left(\hat{\theta}_i-\theta_i\right)^2 &= \mathbb{E}\sum z_i^2\left\{i:|y_i|\geq|y|_{[\hat{k}]}\right\}+\mathbb{E}\sum\theta_i^2\left\{i:|y_i|<|y|_{[\hat{k}]}\right\}\\ &= R_1+R_2.\end{aligned}$$

To establish the result $\mathbb{E}\left\|\hat{\theta}-\theta\right\|^2\sim c^*R_n\left(\Theta_{n,p}\left(\eta_n\right)\right)$, it is enough to show

$$R_1=o\left(1\right)\cdot R_n\left(\Theta_{n,p}\left(\eta_n\right)\right)=o\left(1\right)\cdot\tilde{k}_n\log\frac{n}{\tilde{k}_n},\tag{26}$$

where $\tilde{k}_n\sim n\eta_n\left(c\log\eta_n^{-1}\right)^{-p/2}$ and $\tilde{k}_n\log\frac{n}{\tilde{k}_n}\asymp\tilde{k}_n\log n$, and show

$$R_2\sim c^*R_n\left(\Theta_{n,p}\left(\eta_n\right)\right)\sim\begin{cases}\left(\frac{c}{2}\right)^{1-p/2}n\eta_n\left(2\log\eta_n^{-1}\right)^{1-p/2}, & \text{for } l_p \text{ ball}\\ \left(\frac{c}{2}\right)^{1-p/2}\frac{2}{2-p}n\eta_n\left(2\log\eta_n^{-1}\right)^{1-p/2}, & \text{for } m_p \text{ ball}\end{cases},\tag{27}$$

where $c^*=\left(\frac{c}{2}\right)^{1-\frac{p}{2}}$. The dominating term is then $R_2$. In the following we establish equations (26) and (27).

**Negligibility of $R_1$** Since $R_1$ is a sum of $\hat{k}$ number of $z_i^2$, an upper bound for $R_1$ is then $\mathbb{E}\sum z_{[i]}^2\left\{i\leq\hat{k}\right\}$ which can be written as

$$\begin{aligned}&\mathbb{E}\sum z_{[i]}^2\left\{i\leq\hat{k}\right\}\left[\left\{\hat{k}\leq\gamma_n\tilde{k}_n\right\}+\left\{\gamma_n\tilde{k}_n<\hat{k}\leq k_+\left(q_n\right)\right\}+\left\{\hat{k}>k_+\left(q_n\right)\right\}\right]\\ \leq\ &\mathbb{E}z_{[1]}^2\sum_i\left\{i\leq\hat{k}\right\}\left[\left\{\hat{k}\leq\gamma_n\tilde{k}_n\right\}+\left\{\gamma_n\tilde{k}_n<\hat{k}\leq k_+\left(q_n\right)\right\}+\left\{\hat{k}>k_+\left(q_n\right)\right\}\right]\\ \equiv\ &R_{11}+R_{12}+R_{13},\end{aligned}$$

where $k_+\left(q_n\right)$ is defined in (34) with $q_n=\frac{1}{\sqrt{\log\log\log n}}$. We will show $R_{1i}=o\left(1\right)\cdot\tilde{k}_n\log n$ for $i=1,2$ and 3 separately. For $R_{11}$ term we have

$$R_{11}\leq\mathbb{E}z_{[1]}^2\sum_i\left\{i\leq\gamma_n\tilde{k}_n\right\}\leq\gamma_n\tilde{k}_n\mathbb{E}z_{[1]}^2\leq C\gamma_n\tilde{k}_n\log n=o\left(\tilde{k}_n\log n\right).$$

Note that

$$R_{12}\leq\mathbb{E}z_{[1]}^2\sum_i\left\{i\leq k_+\left(q_n\right)\right\}\left\{\gamma_n\tilde{k}_n<\hat{k}\right\}=k_+\left(q_n\right)\mathbb{E}\left[z_{[1]}^2\left\{\gamma_n\tilde{k}_n<\hat{k}\right\}\right],$$

which can be further bounded by

$$k_+\left(q_n\right)\sqrt{\mathbb{E}z_{[1]}^4}\mathbb{P}^{1/2}\left(\gamma_n\tilde{k}_n<\hat{k}\right)\leq Ck_+\left(q_n\right)\log n\cdot\mathbb{P}^{1/2}\left(\gamma_n\tilde{k}_n<\hat{k}\right),$$

by applying the Cauchy-Schwarz inequality and $\mathbb{E}z_{[1]}^4=O\left(\log^2 n\right)$. Since $k_+\left(q_n\right)=O\left(\tilde{k}_n\right)$ and $\mathbb{P}\left\{\gamma_n\tilde{k}_n<\hat{k}\right\}=o\left(1\right)$ by Lemma 1, we have $R_{12}=o\left(\tilde{k}_n\log n\right)$.

The Negligibility of $R_{13}$ is mainly due to Lemma 2 whose proof is given in Section 5.4.2. From Lemma 2 we know $\mathbb{P}^{1/2}\left(\hat{k} > k_+(q_n)\right) \leq C_D n^{-D}$ for all $D > 0$. We apply the Cauchy-Schwarz to $R_{13}$, then

$$\begin{aligned} R_{13} &\leq \mathbb{E}z_{[1]}^2 \sum_i \left\{\hat{k} > k_+(q_n)\right\} = n\mathbb{E}z_{[1]}^2 \left\{\hat{k} > k_+(q_n)\right\} \\ &\leq n\sqrt{\mathbb{E}z_{[1]}^4}\mathbb{P}^{1/2}\left(\hat{k} > k_+(q_n)\right) = o\left(\tilde{k}_n \log n\right). \end{aligned}$$

**Upper and Lower bounds for $R_2$** From the definition of $\theta$, it is easy to see

$$R_2 \leq \sum_{i=1}^n \theta_i^2 \sim \begin{cases} c\tilde{k}_n \log \frac{n}{\tilde{k}_n}, & \text{for } l_p \text{ ball} \\ c\tilde{k}_n \log \frac{n}{\tilde{k}_n} + \sum_{i=\tilde{k}_n+1}^{r_n\tilde{k}_n} \left(\eta_n \frac{n}{i}\right)^{2/p}, & \text{for } m_p \text{ ball} \end{cases} \sim c^* R_n\left(\Theta_{n,p}(\eta_n)\right).$$

In the following we show that the above upper bound is sharp, i.e., $R_2 \geq c^*(1+o(1))R_n(\Theta_{n,p}(\eta_n))$. Since $(\theta_i)$ is a deceasing sequence, we have

$$R_2 = \mathbb{E}\sum \theta_i^2 \left\{i : |y_i| < |y|_{[\hat{k}]}\right\} \geq \mathbb{E}\sum \theta_i^2 \left\{i : i > \hat{k}\right\}.$$

Note that$\left\{i : i > \hat{k}\right\}\left\{\hat{k} \leq \gamma_n\tilde{k}_n\right\} \geq \left\{i : i > \gamma_n\tilde{k}_n\right\}\left\{\hat{k} \leq \gamma_n\tilde{k}_n\right\}$. We then have

$$\begin{aligned} R_2 &\geq \mathbb{E}\sum \theta_i^2 \left\{i : i > \hat{k}\right\}\left\{\hat{k} \leq \gamma_n\tilde{k}_n\right\} \geq \mathbb{E}\sum \theta_i^2 \left\{i : i > \gamma_n\tilde{k}_n\right\}\left\{\hat{k} \leq \gamma_n\tilde{k}_n\right\} \\ &= \sum \theta_i^2 \left\{i : i > \gamma_n\tilde{k}_n\right\} \cdot \mathbb{P}\left(\hat{k} \leq \gamma_n\tilde{k}_n\right), \end{aligned}$$

where $\mathbb{P}\left(\hat{k} \leq \gamma_n\tilde{k}_n\right) = 1 + o(1)$. For the $l_p$ ball case,

$$\sum \theta_i^2 \left\{i : i > \gamma_n\tilde{k}_n\right\} = \left(\tilde{k}_n - \gamma_n\tilde{k}_n\right) c_n(1-\varepsilon_n) \log\frac{n}{\tilde{k}_n} \sim c\tilde{k}_n \log\frac{n}{\tilde{k}_n} \sim c^* R_n\left(\Theta_{n,p}(\eta_n)\right).$$

For the $m_p$ ball case,

$$\begin{aligned} \sum \theta_i^2 \left\{i : i > \gamma_n\tilde{k}_n\right\} &= \sum \theta_{[i]}^2 \left\{i : \tilde{k}_n \geq i > \gamma_n\tilde{k}_n\right\} + \sum \theta_{[i]}^2 \left\{i : r_n\tilde{k}_n \geq i > \tilde{k}_n\right\} \\ &= \left(\tilde{k}_n - \gamma_n\tilde{k}_n\right) c_n(1-\varepsilon_n) \log\frac{n}{\tilde{k}_n} + \sum_{i=\tilde{k}_n+1}^{r_n\tilde{k}_n} \left(\eta_n^* \frac{n}{i}\right)^{2/p} \\ &\sim c\tilde{k}_n \log\frac{n}{\tilde{k}_n} + \sum_{i=\tilde{k}_n+1}^{r_n\tilde{k}_n} \left(\eta_n^* \frac{n}{i}\right)^{2/p} \sim c^* R_n\left(\Theta_{n,p}(\eta_n)\right). \end{aligned}$$

### 5.3.2 Upper bounds for Theorems 1 and 2

Let the penalty term be defined as in (6). Define the minimizer of the theoretical complexity over $\mu$ as

$$\theta_0 = \arg\min_{\mu} K(\mu, \theta) = \arg\min_{\mu} \left[ \|\theta - \mu\|_2^2 + \mathrm{Pen}\left(\|\mu\|_0\right) \right], \tag{28}$$

which can be understood as the parameter estimator in the noiseless case. By definitions of $K(\theta, y)$ in (2) and $\hat{\theta}$ in (3), we have $K\left(\hat{\theta}, y\right) \leq K(\theta_0, y)$. Also note that $\mathbb{E}\left\langle z, \hat{\theta} - \theta_0 \right\rangle = \mathbb{E}\left\langle z, \hat{\theta} - \theta \right\rangle$. It is straightforward to derive (see for example Abramovich et al. 2006 page 632) that

$$\mathbb{E}\left\|\theta - \hat{\theta}\right\|_2^2 \leq K(\theta_0, \theta) + \mathbb{E}2\left\langle z, \hat{\theta} - \theta \right\rangle. \tag{29}$$

The upper bound (23) follows from (29) and Lemmas 3 and 4 whose proofs are given in Sections 5.4.3 and 5.4.4. Specifically, we have

$$\sup_{\theta \in \Theta_{n,p}(\eta_n)} \mathbb{E}\left\|\hat{\theta} - \theta\right\|_2^2 \leq \sup_{\theta \in \Theta_{n,p}(\eta_n)} K(\theta_0, \theta) + \sup_{\theta \in \Theta_{n,p}(\eta_n)} \mathbb{E}\left\langle z, \hat{\theta} - \theta \right\rangle \leq c^* (1 + o(1)) R_n\left(\Theta_{n,p}(\eta_n)\right).$$

### 5.3.3 Proof of Theorem 3

We obtain the lower bound $\frac{n}{\log n} \log\log n$ in equation (15) by considering the risk of the penalized procedure at $\theta = 0$. From Lemma 5 we have

$$\begin{aligned}
\mathbb{E}\left\|\hat{\theta}\right\|_2^2 &= \mathbb{E}\sum z_{[i]}^2 \left\{ i : i \leq \hat{k} \right\} \geq \mathbb{E}\sum z_{[i]}^2 \left\{ i : i \leq \hat{k} \right\} \left\{ \hat{k} > \varepsilon \frac{n}{\log n} \right\} \\
&\geq \mathbb{E}\sum z_{[i]}^2 \left\{ i : i \leq \varepsilon \frac{n}{\log n} \right\} \left\{ \hat{k} > \varepsilon \frac{n}{\log n} \right\} = \mathbb{E} \sum_{i=1}^{\varepsilon n / \log n} z_{[i]}^2 \left( 1 - \left\{ \hat{k} \leq \varepsilon \frac{n}{\log n} \right\} \right).
\end{aligned}$$

Then the Cauchy–Schwarz inequality implies

$$\mathbb{E}\left\|\hat{\theta}\right\|_2^2 \geq \mathbb{E} \sum_{i=1}^{\varepsilon n / \log n} z_{[i]}^2 - \left[ \mathbb{E}\left( \sum_{i=1}^{\varepsilon n / \log n} z_{[i]}^2 \right)^2 \right]^{1/2} \mathbb{P}^{1/2}\left( \hat{k} \leq \varepsilon \frac{n}{\log n} \right).$$

By Theorem 1.5 in Csörgő and Mason (1985) it is easy to show that

$$\begin{aligned}
\mathbb{E} \sum_{i=1}^{\varepsilon n / \log n} z_{[i]}^2 &\sim \left[ \mathbb{E}\left( \sum_{i=1}^{\varepsilon n / \log n} z_{[i]}^2 \right)^2 \right]^{1/2} \sim 2(\varepsilon n / \log n) \left( \log \frac{n}{\varepsilon n / \log n} \right) \\
&\sim 2\varepsilon \frac{n}{\log n} \log\log n.
\end{aligned}$$

Thus $\mathbb{E}\left\|\hat{\theta} - \theta\right\|_2^2 = \mathbb{E}\left\|\hat{\theta}\right\|_2^2 \geq C \frac{n}{\log n} \log\log n$ for some $C > 0$.

## 5.4 Proofs of lemmas

### 5.4.1 Proof of Lemma 1

By (34) and (38), we have $k_+(q_n) < C_{c,p}\tilde{k}_n$ for some positive constant $C_{c,p}$ depending on $c$, $p$. By Lemma *2,* we have $\mathbb{P}\left(\hat{k} < C\tilde{k}_n\right) \to 1$. Let

$$S(k) = \sum_{i=k+1}^{n} y_{[i]}^2 + \mathrm{Pen}(k), \tag{30}$$

where $\mathrm{Pen}(k)$ is defined in (6). It is enough for us to show that

$$\mathbb{P}\left(\cap_{\gamma_n\tilde{k}_n<k<C\tilde{k}_n}\{S(k) > S(0)\}\right) \to 1,$$

which immediately implies $\mathbb{P}\left(\hat{k} < \gamma_n\tilde{k}_n\right) \to 1$. In the following we show $\mathbb{P}(S(k) > S(0)) = 1 - o\left(\frac{1}{n}\right)$ for each $k \in \left(\gamma_n\tilde{k}_n, C\tilde{k}_n\right)$.

**For $l_p$ ball:** We divide the indices $i$'s for $\theta_i$ into two sets: $S_1 \equiv \left\{1, ..., \tilde{k}_n - 1\right\}$, $S_2 \equiv \left\{\tilde{k}_n, ..., n\right\}$. Let $y_{1[i]}^2$ denote the decreasing order statistics from $\left\{y_i^2 = (\theta_i + z_i)^2, i \in S_1\right\}$, $y_{2[i]}^2 = z_{2[i]}^2$ denote the decreasing order statistics from $\left\{y_i^2 = z_i^2, i \in S_2\right\}$. We have

$$\sum_{i=1}^{k} y_{[i]}^2 = \sup_{k_1+k_2=k}\left\{\sum_{i=1}^{k_1} y_{1[i]}^2 + \sum_{i=1}^{k_2} y_{2[i]}^2\right\} \le \sup_{k_1+k_2=k}\left\{k_1\theta_1^2 + 2\theta_1\sum_{i=1}^{\tilde{k}_n-1}|z_i| + \sum_{i=1}^{\tilde{k}_n-1} z_i^2 + \sum_{i=1}^{k_2} z_{2[i]}^2\right\}.$$

The Chernoff inequality implies that for any $D > 0$,

$$(A) \quad : \quad \mathbb{P}\left(\sum_{i=1}^{\tilde{k}_n}|z_i| > 2\tilde{k}_n\right) < C_D n^{-D},$$

$$(B) \quad : \quad \mathbb{P}\left(\sum_{i=1}^{\tilde{k}_n} z_i^2 > 2\tilde{k}_n\right) < C_D n^{-D},$$

and by Lemmas 6 and 7, we have that for $\delta = \frac{c-2}{2} > 0$ and any $D > 0$,

$$(C) \quad : \quad \mathbb{P}\left(\sum_{i=1}^{k_2} z_{2[i]}^2 > (2+\delta)k_2\log\frac{n}{k_2} + 2(\log n)^3\right) < C_D n^{-D},$$

$$(D) \quad : \quad k_1\theta_1^2 + (2+\delta)k_2\log\frac{n}{k_2} \le c_n k(1-\varepsilon_n)\log\frac{n}{\tilde{k}_n} + C\tilde{k}_n.$$

Since $\varepsilon_n = \gamma_n = \frac{1}{\log\log n}$, under condition (11), it is easy to see that $\varepsilon_n k\log\frac{n}{\tilde{k}_n} \ge \varepsilon_n\gamma_n\tilde{k}_n\log\frac{n}{\tilde{k}_n} \asymp \frac{\log n}{(\log\log n)^2}\tilde{k}_n$ and $\theta_1\tilde{k}_n \asymp \sqrt{\log n}\tilde{k}_n = o\left(\varepsilon_n k\log\frac{n}{\tilde{k}_n}\right)$. From $(A)$, $(B)$,

$(C)$ and $(D)$, we have, with probability $1-o\left(\frac{1}{n}\right)$, that

$$\begin{aligned}\sum_{i=1}^{k} y_{[i]}^2 &\leq 2\theta_1 \cdot 2\tilde{k}_n + 2\tilde{k}_n + c_n k\left(1-\varepsilon_n\right)\log\frac{n}{\tilde{k}_n} + C\tilde{k}_n + 2\left(\log n\right)^3 \\ &= c_n k\left(1-\varepsilon_n\right)\log\frac{n}{\tilde{k}_n} + o\left(c_n\varepsilon_n k\log\frac{n}{\tilde{k}_n}\right).\end{aligned}$$

Since $c_{ni} \to c > 2$ uniformly over $1 \leq i \leq \frac{n}{\log n}$, $c_n = \min_{i\leq\frac{n}{\log n}} c_{ni} \to c > 2$, as $n \to \infty$. Further note that $\sum_{i=1}^{k} \log i \sim \int_1^k \log x dx = k\log k - k$, so $\sum_{i=1}^{k} u_{ni}^2 \geq c_n\left(k\log\frac{n}{k} + k\right) \geq c_n k\log\frac{n}{\tilde{k}_n} - C'k$. Thus, with probability $1-o\left(\frac{1}{n}\right)$, we have

$$\begin{aligned}S\left(k\right) - S\left(0\right) &= \sum_{i=1}^{k} u_{ni}^2 - \sum_{i=1}^{k} y_{[i]}^2 \qquad (31)\\ &\geq c_n k\log\frac{n}{\tilde{k}_n} - c_n k\left(1-\varepsilon_n\right)\log\frac{n}{\tilde{k}_n} - o\left(c_n\varepsilon_n k\log\frac{n}{\tilde{k}_n}\right) \\ &= c_n\varepsilon_n k\log\frac{n}{\tilde{k}_n} - o\left(c_n\varepsilon_n k\log\frac{n}{\tilde{k}_n}\right) > 0.\end{aligned}$$

**For $m_p$ ball:** The proof is similar as above for $l_p$ ball, except we define $S_1 \equiv \left\{1, ..., r_n\tilde{k}_n - 1\right\}$, $S_2 \equiv \left\{r_n\tilde{k}_n, ..., n\right\}$, $r_n = \log\log n$. From the statements $(A)$ and $(B)$ we have

$$\mathbb{P}\left(\sum_{i=1}^{r_n\tilde{k}_n} |z_i| > 2r_n\tilde{k}_n\right) < C_D n^{-D} \text{ and } \mathbb{P}\left(\sum_{i=1}^{r_n\tilde{k}_n} z_i^2 > 2r_n\tilde{k}_n\right) < C_D n^{-D},$$

which yields

$$\begin{aligned}\sum_{i=1}^{k} y_{[i]}^2 &\leq 2\theta_1 \cdot 2r_n\tilde{k}_n + 2r_n\tilde{k}_n + c_n k\left(1-\varepsilon_n\right)\log\frac{n}{\tilde{k}_n} + C\tilde{k}_n + 2\left(\log n\right)^3 \\ &= c_n k\left(1-\varepsilon_n\right)\log\frac{n}{\tilde{k}_n} + o\left(c_n\varepsilon_n k\log\frac{n}{\tilde{k}_n}\right)\end{aligned}$$

with probability $1-o\left(\frac{1}{n}\right)$. Then by similar arguments as in (31) we have $S\left(k\right) - S\left(0\right) \geq 0$ with probability $1-o\left(\frac{1}{n}\right)$.

### 5.4.2 Proof of Lemma 2

It is shown in Lemma 8 that $u_{ni}^2 \geq z^2\left(\frac{iq_n}{2n}\right)$ uniformly over $i \leq \frac{n}{\log n}$. Let $\hat{k}^F$ defined in (32) be the selected model size based on the FDR penalty (Abramovich et al. (2006)), and let $S\left(k\right)$ defined in (30) correspond to the penalties in (6). Then for all $\hat{k}^F < k \leq \frac{n}{\log n}$

we have

$$S\left(\hat{k}^F\right) - S\left(k\right) = \sum_{i=\hat{k}^F+1}^{k} y_{[i]}^2 - \sum_{i=\hat{k}^F+1}^{k} u_{ni}^2 \leq \sum_{i=\hat{k}^F+1}^{k} y_{[i]}^2 - \sum_{i=\hat{k}^F+1}^{k} z^2\left(\frac{iq}{2n}\right),$$

which is non-positive, since the minimum of the objective function $K\left(\theta, y\right)$ with penalty $\sum_{i=1}^{k} z^2\left(\frac{iq}{2n}\right)$ is attained at $\left\|\hat{\theta}^F\right\|_0 = \hat{k}^F$. The inequality $S\left(\hat{k}^F\right) \leq S\left(k\right)$, uniformly over $\hat{k}^F < k \leq \frac{n}{\log n}$, implies $\hat{k} \leq \hat{k}^F$. Therefore we have $\mathbb{P}\left(\hat{k} \geq k_+\left(q_n\right)\right) \leq \mathbb{P}\left(\hat{k}^F \geq k_+\left(q_n\right)\right) \leq C_D n^{-D}$ from Property (P2) in (36).

Note that Abramovich, et al.(2006) require $\gamma > 5$ for condition (11) in order to prove the aforementioned inequality. By careful calculations it can be shown that a looser condition of $\gamma > 4.5$ is enough.

### 5.4.3 Proof of Lemma 3

Let $k = \left\|\mu\right\|_0$ and $u_{ni}^2$ be defined in (6). We have

$$\begin{aligned}
\sup_{\theta\in\Theta_{n,p}(\eta_n)} K\left(\theta_0, \theta\right) &= \sup_{\theta\in\Theta_{n,p}(\eta_n)} \min_{k\leq\frac{n}{\log n}} \min_{\left\{\mu:\left\|\mu\right\|_0=k\right\}} \left[\left\|\theta-\mu\right\|_2^2 + \sum_{i=1}^{k} u_{ni}^2\right] \\
&= \sup_{\theta\in\Theta_{n,p}(\eta_n)} \min_{k\leq\frac{n}{\log n}} \left[\sum_{i=k+1}^{n} \theta_{[i]}^2 + \sum_{i=1}^{k} u_{ni}^2\right].
\end{aligned}$$

By Lemma 8 we have $u_{ni}^2 = \frac{c}{2}\left(1+\delta_{ni}\right) z^2\left(\frac{iq_n}{2n}\right)$, for $q_n \geq \frac{1}{\sqrt{\log\log\log n}}$, $c \geq 2$, and some $\delta_{ni} \to 0$ uniformly over $i \leq \frac{n}{\log n}$ as $n \to \infty$. Define $t_{ni}^2 = \left(1+\delta_{ni}\right) z^2\left(\frac{iq_n}{2n}\right)$. Let $\theta_i' = \left(\frac{c}{2}\right)^{-1/2}\theta_i$ and $\eta_n' = \eta_n\left(\frac{c}{2}\right)^{-p/2}$. It is clear that $\theta' \in \Theta_{n,p}\left(\eta_n'\right)$ which is either $l_p\left[\eta_n'\right]$ or $m_p\left[\eta_n'\right]$ as defined in (9) and (10). Note that $\eta_n'$ satisfies the condition in (11). Write

$$\begin{aligned}
\sup_{\theta\in\Theta_{n,p}(\eta_n)} K\left(\theta_0, \theta\right) &= \frac{c}{2} \sup_{\theta'\in\Theta_{n,p}(\eta_n')} \min_{k\leq\frac{n}{\log n}} \left[\sum_{i=k+1}^{n} \theta_{[i]}'^2 + \sum_{i=1}^{k} t_{ni}^2\right] \\
&\sim \frac{c}{2} \sup_{\theta'\in\Theta_{n,p}(\eta_n')} \min_{k\leq\frac{n}{\log n}} \left[\sum_{i=k+1}^{n} \theta_{[i]}'^2 + \sum_{i=1}^{k} z^2\left(\frac{iq_n}{2n}\right)\right].
\end{aligned}$$

In Abramovich, et al. (2006, pages 633-634) it has been shown that

$$\sup_{\theta'\in\Theta_{n,p}(\eta_n')} \min_{k'} \left[\sum_{i=k'+1}^{n} \theta_{[i]}'^2 + \sum_{i=1}^{k'} z^2\left(\frac{iq_n}{2n}\right)\right] \sim R\left(\Theta_{n,p}\left(\eta_n'\right)\right)$$

with the minimizer attained at $k' \sim n\eta_n'\tau_{\eta_n'}^{-p} \leq \frac{n}{\log n}$. We then immediately have

$$\sup_{\theta\in\Theta_{n,p}(\eta_n)} K\left(\theta_0, \theta\right) \sim \frac{c}{2} R\left(\Theta_{n,p}\left(\eta_n'\right)\right) \sim \left(\frac{c}{2}\right)^{1-p/2} R\left(\Theta_{n,p}\left(\eta_n\right)\right),$$

where the last step follows from equations (12) and (14).

### 5.4.4 Proof of Lemma 4

Recall that for both $l_p$ and $m_p$ balls, $R_n\left(\Theta_{n,p}\left(\eta_n\right)\right) \sim Ck_n\tau_{\eta_n}^2$, where $C = 1$ for $l_p$ ball and $\frac{2}{2-p}$ for $m_p$ ball. So we need only to show $\sup_\theta \mathbb{E}\left\langle z, \hat{\theta} - \theta \right\rangle = o\left(k_n\tau_{\eta_n}^2\right)$.

**Some auxiliary results** Our proof is closely related to previous results in Abramovich, et al. (2006). In this section we review some notations and results on the estimation based on the FDR-penalty procedure there. Specifically, the FDR-thresholding estimators are

$$\begin{aligned} \hat{k}^F &= \arg\min_{k\leq n}\left[\sum_{i=k+1}^{n} y_{[i]}^2 + \sum_{i=1}^{k} z^2\left(\frac{iq_n}{2n}\right)\right], \\ \hat{\theta}_i^F &= y_i I\left\{|y_i| \geq |y|_{\left[\hat{k}^F\right]}\right\}, \end{aligned} \tag{32}$$

where $q_n$ is the FDR control level. Based on the mean discovery number

$$k\left(\theta, q_n\right) = \inf\left\{k : \mathbb{E}_\theta \sum_{i=1}^{n}\left\{y_i^2 \geq z^2\left(\frac{kq_n}{2n}\right)\right\} = k\right\}, \tag{33}$$

there are two important bounding points for the location of $\hat{k}^F$:

$$\begin{aligned} k_-\left(\theta, q_n\right) &= \begin{cases} k\left(\theta, q_n\right) - \alpha_n k_n, & \text{for } k\left(\theta, q_n\right) \geq 2\alpha_n k_n \\ 0, & \text{otherwise} \end{cases}, \\ k_+\left(\theta, q_n\right) &= k\left(\theta, q_n\right) \vee \alpha_n k_n + \alpha_n k_n, \end{aligned} \tag{34}$$

where $k_n$ is defined in (13) and $\alpha_n \equiv \left(b_4\tau_{\eta_n}\right)^{-1}$ for some constant $b_4 > 0$. It has been shown in Abramovich, et al. (2006, pages 637-639) that for $q_n = o(1)$,

**(P1)**

$$\sup_\theta \mathbb{E}\left\langle z, \hat{\theta}^F - \theta_0 \right\rangle = o\left(1\right) R_n\left(\Theta_{n,p}\left(\eta_n\right)\right); \tag{35}$$

Furthermore, for any $q_n$ such that $q_n \geq \frac{C}{\log n}$ and $q_n \to q \in [0, \frac{1}{2})$, Abramovich, et al. (2006, pages 608, 640, 607 and 624, respectively) showed that

**(P2)**

$$\mathbb{P}\left(\hat{k}^F \geq k_+\left(q_n\right)\right) \leq C_D n^{-D}, \text{ for any constant } D; \tag{36}$$

**(P3)**

$$\text{Card}\left\{i : |\theta_i| > z\left(\frac{k_+\left(\theta, q_n\right) q_n}{2n}\right)\right\} \leq Ck_n; \tag{37}$$

**(P4)**

$$k\left(\theta, q_n\right) \leq \left(1 + o\left(1\right)\right) k_n, \text{ for all } \theta \in \Theta_{n,p}\left(\eta_n\right). \tag{38}$$

**Proof of Lemma 4** From Property (P1) in equation (35) it is then enough to show that

$$\mathbb{E}\left\langle z, \hat{\theta} - \theta \right\rangle - \mathbb{E}\left\langle z, \hat{\theta}^F - \theta \right\rangle = o(1) R_n\left(\Theta_{n,p}(\eta_n)\right).$$

Let $q_n = \frac{1}{\sqrt{\log\log\log n}}$. Define

$$\begin{aligned} \eta(y_i, t) &= y_i \{|y_i| \geq t\}, \\ A_n &= \left\{\hat{k}^F \leq k_+(\theta, q_n)\right\} \\ S_n &= \left\{i : |\theta_i| \leq t_-^*\right\}, \end{aligned}$$

where the simplified notations $t_{i,q_n}^2 = z^2\left(\frac{iq_n}{2n}\right)$ and $t_-^* = t_{k_+(\theta,q_n),q_n}$. Note that $\hat{\theta}_i = \eta\left(y_i, u_{\hat{k}}\right) = y_i\left\{|y_i| \geq |y|_{[\hat{k}]}\right\}$ and $\hat{\theta}_i^F = \eta\left(y_i, t_{\hat{k}^F,q_n}\right)$. Write $\hat{\Delta}_i = z_i\left(\eta\left(y_i, u_{\hat{k}}\right) - \eta\left(y_i, t_{\hat{k}^F,q_n}\right)\right)$. So we have

$$\begin{aligned} \mathbb{E}\left\langle z, \hat{\theta} - \theta \right\rangle - \mathbb{E}\left\langle z, \hat{\theta}^F - \theta \right\rangle &= \mathbb{E}\sum z_i\left(\hat{\theta}_i - \hat{\theta}_i^F\right) \\ &= \mathbb{E}\sum z_i\left(\eta\left(y_i, u_{\hat{k}}\right) - \eta\left(y_i, t_{\hat{k}^F,q_n}\right)\right) \\ &= \mathbb{E}\left[\sum_{S_n} \hat{\Delta}_i A_n\right] + \mathbb{E}\left[\sum \hat{\Delta}_i A_n^c\right] + \mathbb{E}\left[\sum_{S_n^c} \hat{\Delta}_i A_n\right] \\ &\equiv T_{1n} + T_{2n} + T_{3n}. \end{aligned}$$

We first study the term $T_{1n}$. It can be shown that $z_i(\eta(y_i, t) - \theta_i)$ is deceasing over $t \geq |\theta_i|$, because for any $t_1$, $t_2$ such that $t_2 \geq t_1 \geq |\theta_i|$, $\Delta \equiv z_i(\eta(y_i, t_1) - \theta_i) - z_i(\eta(y_i, t_2) - \theta_i) \geq 0$. To show this is true, note that in the case $|y_i| \leq t_1 \leq t_2$ or $t_1 \leq t_2 \leq |y_i|$, we have $\Delta = 0$; in the case $t_1 \leq |y_i| \leq t_2$, the inequality $|\theta_i| \leq t_1 \leq |y_i| = |\theta_i + z_i|$ indicates $\theta_i$ and $z_i$ have the same sign, so $\Delta = z_i y_i \geq 0$. By Lemma 8, we have $t_{i,q_n} \leq u_{ni}$ for each $i$, which further indicates $\hat{k} \leq \hat{k}^F$ (refer to the proof of Lemma 2 for details). So we have $u_{\hat{k}} \geq t_{\hat{k},q_n} \geq t_{\hat{k}^F,q_n}$. Moreover, under the event $A_n$ and within the set $S_n$, we have $t_{\hat{k}^F,q_n} \geq t_-^* \geq |\theta_i|$. Thus $u_{\hat{k}} \geq t_{\hat{k}^F,q_n} \geq |\theta_i|$, which implies $\hat{\Delta}_i \leq 0$ and $T_{1n} \leq 0$.

The Negligibility of $T_{2n}$ is basically due to the fact that $\mathbb{P}(A_n^c) \leq C_D n^{-D}$. Since $u_{\hat{k}} \geq t_{\hat{k},q_n} \geq t_{\hat{k}^F,q_n}$, it is easy to see that

$$\left|\eta\left(y_i, u_{\hat{k}}\right) - \eta\left(y_i, t_{\hat{k}^F,q_n}\right)\right| \leq u_{\hat{k}} \leq C u_1, \tag{39}$$

then

$$T_{2n} = \sum \mathbb{E}\left(\hat{\Delta}_i A_n^c\right) \leq C u_1 \sum \mathbb{E}|z_i| \{A_n^c\} \leq C u_1 n \left[\mathbb{P}(A_n^c)\right]^{1/2} = o\left(k_n \tau_{\eta_n}^2\right),$$

where the second inequality is due to the Cauchy–Schwarz inequality, i.e., $\mathbb{E}\left|z_i\right|\left\{A_n^c\right\} \leq \sqrt{\mathbb{E}z_i^2}\left[\mathbb{P}\left(A_n^c\right)\right]^{1/2} = \left[\mathbb{P}\left(A_n^c\right)\right]^{1/2}$.

We now show that $T_{3n}$ is $o\left(k_n\tau_{\eta_n}^2\right)$. From equation (39), we have

$$T_{3n} = \mathbb{E}\left[\sum_{S_n^c}\hat{\Delta}_i A_n\right] \leq \sum_{S_n^c}\mathbb{E}\left|z_i\left[\eta\left(y_i, u_{\hat{k}}\right) - \eta\left(y_i, t_{\hat{k}^F, q_n}\right)\right]\right| \leq Cu_1\sum_{S_n^c}\mathbb{E}\left|z_i\right| \leq Cu_1\left|S_n^c\right|.$$

Since $\left|S_n^c\right| \leq Ck_n$ from Property (P3) in equation (37) we immediately have

$$T_{3n} \leq Cu_1\left|S_n^c\right| = o\left(\tau_{\eta_n}^2 k_n\right).$$

### 5.4.5 Proof of Lemma 5

Let $S(k) = \sum_{i=k+1}^{n} z_{[i]}^2 + \sum_{i=1}^{k} u_{ni}^2$, where $u_{ni}^2 = c_{ni}\log\frac{n}{i}$, $c_{ni} \to c < 2$ uniformly over $1 \leq i \leq \frac{n}{\log n}$. We show that there exists a constant $0 < \varepsilon < 1$ such that $\mathbb{P}\left(\cap_{k\leq\varepsilon n/\log n}\left\{S(k) > S\left(\frac{n}{\log n}\right)\right\}\right) \to 1$, which immediately implies $\mathbb{P}\left(\hat{k} > \varepsilon\frac{n}{\log n}\right) \to 1$. It is enough to show that $\mathbb{P}\left(S(k) > S\left(\frac{n}{\log n}\right)\right) = 1 - o\left(\frac{1}{n}\right)$ for each $0 \leq k \leq \varepsilon\frac{n}{\log n}$ to complete the proof.

Since $c_{ni} \to c < 2$ uniformly over $1 \leq i \leq \frac{n}{\log n}$, then $c_n \equiv \max_{1\leq i\leq n/\log n} c_{ni} \to c < 2$. Let $\delta_1 = \frac{2-c}{2}$. The identity $\left(x\log\frac{n}{x} + x\right)' = \log\frac{n}{x}$ yields

$$\sum_{i=k+1}^{n/\log n} u_{ni}^2 \leq c_n\int_k^{n/\log n}\log\frac{n}{x}dx \leq (2-\delta_1)\left[\frac{n}{\log n}\left(\log\log n + 1\right) - k\log\frac{n}{k} - k\right],$$

for $n$ sufficiently large. Let $\delta_2 = \frac{2-c}{4}$. Following Lemma 6 we have

$$\sum_{i=k+1}^{n/\log n} z_{[i]}^2 \geq (2-\delta_2)\frac{n}{\log n}\log\log n - (2+\delta_2)\,k\log\frac{n}{k} - 4\left(\log n\right)^3$$

with probability $1 - o\left(\frac{1}{n}\right)$. Note that $k\log\frac{n}{k} \leq \frac{n}{\log n}\log\log n$ for $k \leq \frac{n}{\log n}$. Thus for $n$ sufficiently large we have

$$\begin{aligned} S(k) - S\left(\frac{n}{\log n}\right) &= \sum_{i=k+1}^{n/\log n} z_{[i]}^2 - \sum_{i=k+1}^{n/\log n} u_{ni}^2 \\ &\geq \left(\delta_1 - \delta_2 - \varepsilon\left(\delta_1+\delta_2\right)\right)\frac{n}{\log n}\log\log n - 4\left(\log n\right)^3 - (2-\delta_1)\frac{n}{\log n} > 0, \end{aligned}$$

where $0 < \varepsilon < \frac{\delta_1-\delta_2}{\delta_1+\delta_2}$, with probability $1 - o\left(\frac{1}{n}\right)$.

### 5.4.6 Proof of Lemma 6

First we consider the upper bound. Let $m_n = (2+\delta)\, k_n \log \frac{n}{k_n} + 2\,(\log n)^3$. We have

$$\mathbb{P}\left(\sum_{i=1}^{k_n} z_{[i]}^2 > m_n\right) \leq \binom{n}{k_n} \mathbb{P}\left(\sum_{i=1}^{k_n} z_i^2 > m_n\right) = \binom{n}{k_n} \frac{\Gamma\left(\frac{k_n}{2}, \frac{m_n}{2}\right)}{\Gamma\left(\frac{k_n}{2}\right)},$$

where

$$\Gamma(s,x) = \int_x^\infty t^{s-1} e^{-t} dt \leq x^{s+1} e^{-x} \int_x^\infty t^{-2} dt \leq x^s e^{-x},$$

when $x \geq s-1$. By Stirling's approximation $n! \sim \sqrt{2\pi n}\left(\frac{n}{e}\right)^n$, we then have

$$\begin{aligned}
\mathbb{P}\left(\sum_{i=1}^{k_n} z_{[i]}^2 > m_n\right) &\leq \binom{n}{k_n} \frac{\left(\frac{m_n}{2}\right)^{k_n/2} \exp\left(-\frac{m_n}{2}\right)}{\Gamma\left(\frac{k_n}{2}\right)} \\
&\leq C \frac{n^n}{k_n^{k_n}\,(n-k_n)^{n-k_n}} \left(\frac{e m_n}{k_n}\right)^{k_n/2} \exp\left(-(1+\delta/2)\, k_n \log \frac{n}{k_n}\right) \exp\left(-(\log n)^3\right) \\
&\leq \exp\left(-\frac{\delta}{4} k_n \log \frac{n}{k_n}\right) \exp\left(-(\log n)^3\right) < C_D n^{-D}.
\end{aligned}$$

Now we consider the lower bound. When $k_n \leq (\log n)^2$, the result is obvious as the bound is 0. When $k_n > (\log n)^2$, for any $t > 0$, we have

$$\begin{aligned}
\mathbb{P}\left(\sum_{i=1}^{k_n} z_{[i]}^2 < (2-\delta)\, k_n \log \frac{n}{k_n}\right) &\leq \mathbb{P}\left(k_n z_{[k_n]}^2 < (2-\delta)\, k_n \log \frac{n}{k_n}\right) \\
&\leq \mathbb{E}\left(e^{-t z_{[k_n]}^2 + t(2-\delta)\log \frac{n}{k_n}}\right) = \mathbb{E}\left(e^{-t z_{[k_n]}^2}\right) \left(\frac{n}{k_n}\right)^{t(2-\delta)},
\end{aligned}$$

where the last inequality follows from the Markov inequality. Let $F(x)$ be the c.d.f. of $z_i^2 \sim \chi_1^2$, i.e., $F^{-1}(p) = \left(\bar{\Phi}^{-1}\left(\frac{1-p}{2}\right)\right)^2$. When $\frac{1-p}{2} \leq 0.01$, equation (12.7) in Abramovich, et al. (2006) gives $F^{-1}(p) = 2\log\frac{2}{1-p} - \log\log\frac{2}{1-p} - r(p)$, where $r(p) \in (1.8, 3)$, which implies $F^{-1}(p) \geq \left(2 - \frac{\delta}{2}\right)\log\frac{2}{1-p} - C$ for some constant $C > 0$ for $0 < p < 1$. Thus by Stirling's approximation, for $t = o(k_n)$, we have

$$\begin{aligned}
\mathbb{E}\left(e^{-t z_{[k_n]}^2}\right) &= \frac{1}{B(n-k_n+1, k_n)} \int_0^1 e^{-t F^{-1}(p)} p^{n-k_n} (1-p)^{k_n-1}\, dp \\
&\leq \frac{1}{B(n-k_n+1, k_n)} \int_0^1 e^{-t\left(\left(2-\frac{\delta}{2}\right)\log\frac{2}{1-p} - C\right)} p^{n-k_n} (1-p)^{k_n-1}\, dp \\
&= \frac{e^{-t\left(\left(2-\frac{\delta}{2}\right)\log 2 - C\right)}}{B(n-k_n+1, k_n)} \int_0^1 (1-p)^{t\left(2-\frac{\delta}{2}\right)} \cdot p^{n-k_n} (1-p)^{k_n-1}\, dp \\
&\leq e^{C't} \frac{B\left(n-k_n+1, k_n + t\left(2-\frac{\delta}{2}\right)\right)}{B(n-k_n+1, k_n)} \leq e^{C't} \left(\frac{k_n + t\left(2-\frac{\delta}{2}\right)}{n + t\left(2-\frac{\delta}{2}\right)}\right)^{t\left(2-\frac{\delta}{2}\right)}.
\end{aligned}$$

Let $t = \log n$. We then have

$$\mathbb{E}\left(e^{-tz^2_{[k_n]}}\right)\left(\frac{n}{k_n}\right)^{t(2-\delta)} \leq e^{C't}\left(\frac{k_n + t\left(2-\frac{\delta}{2}\right)}{n + t\left(2-\frac{\delta}{2}\right)}\right)^{t\frac{\delta}{2}} < C_D n^{-D}.$$

### 5.4.7 Proof of Lemma 7

Since $\theta_1^2 = c_n(1-\varepsilon_n)\log\frac{n}{\tilde{k}_n}$, it is enough to show

$$(2+\delta)k_2\log\frac{n}{k_2} \leq c_n k_2(1-\varepsilon_n)\log\frac{n}{\tilde{k}_n} + C\tilde{k}_n,$$

which follows immediately from the following inequalities

$$(2+\delta)k_2\log\frac{n}{k_2} \leq \begin{cases} \tilde{k}_n, & \text{for } k_2 \leq \tilde{k}_n/\log^2 n \\ c_n k_2(1-\varepsilon_n)\log\frac{n}{\tilde{k}_n}, & \text{for } k_2 > \tilde{k}_n/\log^2 n \end{cases}$$

under condition (11), as $n$ sufficiently large.

### 5.4.8 Proof of Lemma 8

For $u_{ni}^2$ defined in (6), we have $u_{ni}^2 = c_{ni}\log\frac{n}{i} \geq 2\log\frac{n}{i} - (1-\varepsilon)\log\log\frac{n}{i}$, for each $i \leq \frac{n}{\log n}$, some $\varepsilon \in (0,1)$ and $n$ sufficiently large. By equation (12.7) in Abramovich, et al. (2006), if $\eta \leq 0.01$, we have $z^2(\eta) = 2\log\eta^{-1} - \log\log\eta^{-1} - r(\eta)$, where $r(\eta) \in [1.8, 3]$. Thus for $q_n^* = \frac{1}{\sqrt{\log\log\log n}}$ and $i \leq \frac{n}{\log n}$,

$$\begin{aligned} z^2\left(\frac{iq_n}{2n}\right) &\leq z^2\left(\frac{iq_n^*}{2n}\right) \leq 2\log\frac{2n}{iq_n^*} - \log\log\frac{2n}{iq_n^*} \\ &\leq 2\log\frac{n}{i} + 2\log\frac{2}{q_n^*} - \log\log\frac{n}{i} - \log\log\frac{2}{q_n^*} + C \qquad (40) \\ &\leq 2\log\frac{n}{i} - (1-\varepsilon)\log\log\frac{n}{i} - \varepsilon\log\log\frac{n}{i} + 2\log\frac{2}{q_n^*} + C \\ &\leq 2\log\frac{n}{i} - (1-\varepsilon)\log\log\frac{n}{i} - \varepsilon\log\log\log n + 2\log\left(2\sqrt{\log\log\log n}\right) + C \\ &\leq 2\log\frac{n}{i} - (1-\varepsilon)\log\log\frac{n}{i}, \end{aligned}$$

which immediately implies $u_{ni}^2 \geq z^2\left(\frac{iq_n}{2n}\right)$. Note that $u_{ni}^2 = c_{ni}\log\frac{n}{i} = \frac{c}{2}\cdot\frac{c_{ni}}{c}\cdot\frac{2\log\frac{n}{i}}{z^2\left(\frac{iq_n}{2n}\right)}\cdot z^2\left(\frac{iq_n}{2n}\right)$. Since $\frac{c_{ni}}{c} \to 1$ and $\frac{2\log\frac{n}{i}}{z^2\left(\frac{iq_n}{2n}\right)} \to 1$ uniformly over $i \leq \frac{n}{\log n}$, we then have $u_{ni}^2 = \frac{c}{2}(1+\delta_{ni})z^2\left(\frac{iq_n}{2n}\right)$ where $\delta_{ni} \to 1$ over $i \leq \frac{n}{\log n}$.

For $u'^2_{ni}$ defined in (16) and $q = 1$, with each $i \leq n$, we have

$$\begin{aligned} z^2\left(\frac{iq}{2n}\right) &< 2\log\frac{2n}{i} - \log\log\frac{2n}{i} - 1.8 + z^2(0.01) \\ &= 2\log\frac{n}{i} - \log\log\frac{2n}{i} + 2\log 2 - 1.8 + z^2(0.01) \\ &< 2\log\frac{n}{i} - \log\log\frac{2n}{i} + 5. \end{aligned}$$

Since $u'^2_{ni} = c\log\frac{n}{i} + c\left(d + (i-1)\log\frac{i-1}{i}\right) \geq c\log\frac{n}{i} + c(d-1)$ and $\log\log\frac{2n}{i} \geq \log\log 2$, we have $z^2\left(\frac{i}{2n}\right) < u'^2_{ni}$ for each $i \in [1, n]$ if $-\log\log 2 + 5 \leq c(d-1)$, i.e., $d \geq \frac{5.37}{c} + 1$.

**Remark 3** *For every $D > 0$ and $d \geq \frac{5.37}{c} + 1$, by a similar argument as in the proof of Lemma 2, Lemma 8 together with Equition (36) imply that there is a $q \in (0,1)$ such that $\mathbb{P}\left(\hat{k} \geq k_+(q)\right) \leq C_D n^{-D}$ for penalties in (16), where $k_+(q) = o\left(\frac{n}{\log n}\right)$.*

#### 5.4.9 Proof of Corollary 2

Let $S'(k)$ be the objective function defined in (30) with penalty function $\mathrm{Pen}'(k)$. Let $\hat{k}$ and $\hat{k}'$ be the minimizers of $S(k)$ and $S'(k)$, respectively. By definition, we have $S\left(\hat{k}\right) + S'\left(\hat{k}'\right) \leq S\left(\hat{k}'\right) + S'\left(\hat{k}\right)$, and thus $\mathrm{Pen}'\left(\hat{k}'\right) - \mathrm{Pen}\left(\hat{k}'\right) \leq \mathrm{Pen}'\left(\hat{k}\right) - \mathrm{Pen}\left(\hat{k}\right)$. Since $\left(\frac{n}{k}\right)^k \leq \binom{n}{k} \leq \left(\frac{en}{k}\right)^k$, with $c_2 = c$, we have $(c_1 - cd)k \leq \mathrm{Pen}'(k) - \mathrm{Pen}(k) \leq (c_1 - cd + c)k$. So $(c_1 - cd)\hat{k}' \leq (c_1 - cd + c)\hat{k}$. According to Remark 3, the selected model size by $\mathrm{Pen}'(k)$ is $\hat{k}' \leq C\hat{k} = o\left(\frac{n}{\log n}\right)$ with high probability. Then the corollary for $\mathrm{Pen}'(k)$ follows Proposition 1 for $\mathrm{Pen}(k)$ in (16) with $c \geq 2$ *and* $d \geq \frac{5.37}{c} + 1$.


## Acknowledgement

We appreciate the kind review and valuable suggestions from the editor and four anonymous referees.


## References


[1] 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.

[2] ABRAMOVICH, F., GRINSHTEIN, V. and PENSKY, M. (2007). On optimality of Bayesian testimation in the normal means problem. *Ann. Stat.* **35**, 2261-2286.

[3] AKAIKE, H. (1973), Information theory and an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory* (B.N. Petrov and F. Czáki, eds). 267-281. Akadémiai Kiadó, Budapest.

[4] AKAIKE, H. (1974). A new look at the statistical model identification. *IEEE T. Automat. Contr.* **19** (6): 716–723.

[5] BARRON, A. R. BIRGÉ L. and MASSART, P. (1999). Risk bounds for model selection via penalization. *Probab. Theory. Rel. Fields*, **113**, 301–413.

[6] BENJAMINI, Y. and HOCHBERG, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. Roy. Stat. Soc. B.* **57** 289–300.

[7] BENJAMINI, Y. and GAVRILOV, Y. (2009) A simple forward selection procedure based on false discovery rate control. *Ann. Appl. Stat.* Volume **3, 1,** 179-198.

[8] BIRGÉ L. and MASSART, P. (2001). Gaussian model selection. *J. Eur. Math. Soc.* **3** 203–268.

[9] BIRGÉ L. and MASSART, P. (2007). Minimal penalties for Gaussian model selection. *Probab. Theory. Rel. Fields* **138** 33–73.

[10] BROWN, L. D. and LOW, M. G. (1996). Asymptotic equivalence of nonparametric regression and white noise. *Ann. Stat.* **24** 2384-2398.

[11] BROWN, L.D. CAI, T. T. and ZHOU, H. H. (2008), Robust nonparametric estimation via wavelet median regression, *Ann. Stat.* **36**, 2055-2084.

[12] BROWN, L.D. CAI, T. and ZHOU, H. H. (2009), Nonparametric regression in exponential families, *Ann. Stat.* To appear.

[13] CAI, T. T. and ZHOU, H. H. (2009), Asymptotic equivalence and adaptive estimation for robust nonparametric regression, *Ann. Stat.* To appear.

[14] CHENG, S. (1992). Large deviation theorem for Hill's estimator. *Acta Math. Sin.* **8** (3): 243-254.

[15] CSÖRGŐ, S. and MASON, D. (1985). Central limit theorems for sums of extreme values. *Math. Proc. Cambridge* **98** 547-558.

[16] DONOHO, D.L. and JOHNSTONE, I.M. and HOCH, J.C. and STERN, A.S. (1992). Maximum entropy and the nearly black object. *J. Roy. Stat. Soc. B.* **54**(1) 41-81.

[17] DONOHO, D. L. and JOHNSTONE, I. M. (1994a). Ideal spatial adaptation by wavelet shrinkage. *Biometrika* **81**(3) 425-455.

[18] DONOHO, D. L. and JOHNSTONE, I. M. (1994b). Minimax Risk Over $l_p$-Balls for $l_q$-Error. *Probab. Theory. Rel Fields.* **99**, 277–303.

[19] DONOHO, D. L. and JOHNSTONE, I. M. (1995), Adapting to unknown smoothness via wavelet Shrinkage. *J. Amer. Statistical Assoc.* **90**, 1200–1224.

[20] DONOHO, D. L. and JOHNSTONE, I. M. (1998), Minimax estimation via wavelet shrinkage. *Ann. Stat.* **26**, 879–921.

[21] DAVID, H. A. and NAGARAJA, H. N. (2003). *Order Statistics, Third Edition.* Wiley & Sons, New York.

[22] EFRON, B. (2003). Robbins, empirical Bayes and microarrays. *Ann. Stat.* **31** 366–378.

[23] FELLER, W. (1968). *An introduction to probability theory and its applications. Vol. I*, Third edition, John Wiley & Sons Inc., New York.

[24] FOSTER, D. P. and GEORGE, E. I. (1994). The risk inflation criterion for multiple regression. *Ann. Stat.* **22** 1947–1975.

[25] FOSTER, D. P. and STINE, R. A. (1999). Local asymptotic coding and the minimum description length. *IEEE T. Inform. Theory* **45** 1289–1293.

[26] GEORGE, E. I. and FOSTER, D. P. (2000). Calibration and empirical Bayes variable selection. *Biometrika* **87** 731–747.

[27] GOLUBEV, G. K., NUSSBAUM, M. and ZHOU, H. H. (2009). Asymptotic equivalence of spectral density estimation and Gaussian white noise. *Ann. Stat.* To appear.

[28] JOHNSTONE, I.M. (1994). Minimax Bayes, asymptotic minimax and sparse wavelet priors. in S. Gupta and J. Berger, eds, *Statistical Decision Theory and Related Topics V*, Springer-Verlag, 303–326.

[29] JOHNSTONE, I.M. (2011). Gaussian Estimation: Sequence and Multiresolution Models. Unpublished manuscript. http://www-stat.stanford.edu/~imj/.

[30] NUSSBAUM, M. (1996). Asymptotic equivalence of density estimation and Gaussian white noise. *Ann. Stat.* **24** 2399–2430.

[31] PINSKER, M. S. (1980). Optimal filtering of square-integrable signals in Gaussian white noise. *Problems Inform. Transmission* 120-133

[32] TIBSHIRANI, R. and KNIGHT, K. (1999). The covariance inflation criterion for adaptive model selection. *J. Roy. Stat. Soc. B.* **61** 529–546.

[33] YANG,Y. and BARRON, A. R. (1998). An asymptotic property of model selection criteria. *IEEE T. Inform. Theory*, **44**, pp.117-133.

[34] YANG, Y. (1999). Model selection for nonparametric regression. *Stat. Sinica.* **9** 475-499.