

Model Selection and Sharp Asymptotic Minimavity

Harrison Zhou

Yale University

Jointly with Zheyang Wu

Overview

1. Introduction
2. Main Results
3. Discussions
4. Summary

1. Introduction

Observe that

$$Y_i = \theta_i + Z_i, \quad Z_i \stackrel{iid}{\sim} N(0, 1), \quad i = 1, 2, \dots, n.$$

Setting:

$$\Theta_{n,p}(c_n) = \left\{ \theta \in \mathbb{R}^n : \frac{1}{n} \sum_{i=1}^n |\theta_i|^p \leq c_n \right\}$$

Goal: to estimate $\theta = (\theta_1, \dots, \theta_n)$.

Remark: For $p < 2$, linear estimators fail for the squared error loss.

AIC, BIC and RIC

Let

$$\hat{\theta}^P = \arg \min_{\theta} \{ \|\mathbf{Y} - \theta\|^2 + \text{Pen}(\theta) \}$$

where

$$\text{Pen}(\theta) = \lambda \|\theta\|_0,$$

i.e.,

$$\hat{\theta}_i^P = \begin{cases} Y_i, & Y_i^2 \geq \lambda \\ 0, & \text{otherwise} \end{cases} .$$

- **AIC:** $\lambda = 2$. Akaike (1973,1974): “IC stands for information criterion and A is added so that similar statistics BIC, DIC etc., may follow.”
- **RIC:** $\lambda = 2 \log n$. Foster and George (1994), Donoho and Johnstone (1994).

False Discovery Rate (FDR) Procedure

Abramovich, Benjamini, Donoho, and Johnstone (2006, AOS): Let P_1, P_2, \dots, P_n be the corresponding P-values, and let $\hat{k} = \max \{k : P_{(k)} \leq k\alpha/n\}$.

FDR procedure:

$$\hat{\theta}_i^{FDR} = \begin{cases} Y_i, & P_i \leq \hat{k}\alpha/n, \text{ or } Y_i^2 \geq z^2 \left(\frac{\hat{k}\alpha}{2n} \right) \\ 0, & \text{otherwise} \end{cases} .$$

Remarks:

(i) Benjamini and Hochberg (1995) showed that

$$\mathbb{E} \frac{\# \text{ of false rejections}}{\hat{k}} \left\{ \hat{k} > 0 \right\} \leq \alpha.$$

(ii) Seeger, P. (1968, Technometrics)

$$\frac{\# \text{ of false rejections}}{\text{rejections}} \preceq \frac{nP_{(\hat{k})}}{\hat{k}} \leq \alpha.$$

(iii) FDR procedure is connected to a penalized estimator with

$$\text{Pen}(\theta) = \sum_{i=1}^k z^2 \left(\frac{i\alpha}{2n} \right), \quad k = \|\theta\|_0.$$

Proposition (ABDJ, 2006)

For a range of false discovery levels α ,

$$\sup_{\theta \in \Theta_{n,p}(c_n)} \mathbb{E} \left\| \hat{\theta}^{FDR} - \theta \right\|^2 = (1 + o(1)) R(\Theta_{n,p}(c_n))$$

where

$$c_n \in [n^{-1} \log^5 n, n^{-\delta}], \quad 0 < \delta < 1.$$

Remark: $R(\Theta_{n,p}(c_n))$ is the minimax risk

$$R(\Theta_{n,p}(c_n)) = \inf_{\hat{\theta}} \sup_{\theta \in \Theta_{n,p}(c_n)} \mathbb{E} \left\| \hat{\theta} - \theta \right\|^2.$$

Conjecture 1.2. in ABDJ (2006).

Let

$$\text{Pen}(\theta) = 2k \log \frac{n}{k}, k = \|\theta\|_0$$

and define

$$\hat{\theta}^P = \arg \min_{\theta} \{ \|\mathbf{Y} - \theta\|^2 + \text{Pen}(\theta) \}.$$

Then

$$\sup_{\theta \in \Theta_{n,p}(c_n)} \mathbb{E} \left\| \hat{\theta}^P - \theta \right\|^2 = (1 + o(1)) R(\Theta_{n,p}(c_n)).$$

Remark: (i) It can be shown that for fixed $\alpha \in (0, 1)$,

$$\sum_{i=1}^k z^2 \left(\frac{i\alpha}{2n} \right) \sim 2k \log \frac{n}{k}, \text{ as } \frac{n}{k} \rightarrow \infty.$$

(ii) Apparently we need to have a certain restriction for k , otherwise the minimum is at attained at $k = n$. We restrict $k \leq n / \log n$.

Does the conjecture make sense?

A Puzzle: Benjamini and Gavrilov (2007).

Example: $n = 160$, (true) $k = 20$.

The performance of FDR procedure is very close to the penalized estimation with

$$\text{Pen}(\theta) = 4k \log \frac{n}{k}, k = \|\theta\|_0$$

CIC in Tibshirani and Knight (1999).

2. Main Results

Theorem 1.

Conjecture 1.2. in ABDJ (2006) is true.

Proof: A quote from ABDJ (2006), “We suspect that the methods developed in this paper may be extended to yield a proof of this conjecture”.

Key idea: We may bound the penalty by two FDR type penalties.

Extension of the conjecture

Let

$$\text{Pen}(\theta) = \sum_{i=1}^k t_i, \text{ where } k = \|\theta\|_0$$

where

$$2 \log \frac{n}{i} - l_n \leq t_i \leq 2 \log \frac{n}{i} + u_n,$$

where $l_n = (1 - \epsilon) \log \log \log n$ and $u_n = (1 - \epsilon) \log \log n$ for some $\epsilon > 0$.

It can be shown that $\hat{\theta}^P$ is asymptotically minimax.

Why such an "ugly" extension of the conjecture?

The following procedures are asymptotically sharp minimax.

- Foster and Stine (1999). See also Hansen and Yu (1998, 2002), Barron, Birgé and Massart (1999).

$$\text{Pen}(\theta) = \sum_{i=1}^k 2 \log \frac{n}{i}, k = \|\theta\|_0.$$

- George and Foster (2000)

$$\text{Pen}(\theta) = \sum_{i=1}^k 2 \log \left(\frac{n+1}{i} - 1 \right), k = \|\theta\|_0.$$

- Birgé and Massart (2001)

$$\text{Pen}(\theta) = 2k \log \frac{n}{k}, k = \|\theta\|_0.$$

Theorem 2.

If

$$\text{Pen}(\theta) = ck \log \frac{n}{k}, \text{ for some } c \geq 2,$$

then

$$\frac{\sup_{\theta \in \Theta_{n,p}(c_n)} \mathbb{E} \left\| \hat{\theta}^P - \theta \right\|^2}{R(\Theta_{n,p}(c_n))} \rightarrow \left(\frac{c}{2} \right)^{1-p/2}.$$

Remark:

- (i) CIC in Tibshirani and Knight (1999) is not asymptotically sharp minimax.
- (ii) Bayes procedures in Abramovich, Grinshtein and Pensky (2007, AOS) are not asymptotically sharp minimax.
- (iii) For $c < 2$, the ratio goes to ∞ .

Proof of Theorem 2. Let

$$K(\theta) = \min_{\mu} [\|\theta - \mu\|_2^2 + \text{Pen}(\mu)]$$

then

$$\mathbb{E}_{y|\theta} \left\| \hat{\theta}^P - \theta \right\|^2 \leq K(\theta) + 2\mathbb{E}_{y|\theta} \left\langle \hat{\theta}^P - \theta, z \right\rangle.$$

It can be shown

$$\begin{aligned} \sup_{\Theta_{n,p}} K(\theta) &= (1 + o(1)) \left(\frac{c}{2}\right)^{1-p/2} R_n(\Theta_{n,p}) \\ \sup_{\Theta_{n,p}} \mathbb{E}_{y|\theta} \left\langle \hat{\theta}^P - \theta, z \right\rangle &= o(1) R_n(\Theta_{n,p}) \end{aligned}$$

for $c \geq 2$.

Discussions

How to solve the puzzle?

$$z^2 \left(\frac{k\alpha}{2n} \right) = 7.48, 2 \log \frac{n}{k} = 4.16$$

when $n = 160$, $k = 20$ and $\alpha = 0.05$, although

$$\lim_{n/k \rightarrow \infty} \frac{z^2 \left(\frac{k\alpha}{2n} \right)}{2 \log \frac{n}{k}} = 1.$$

Some Questions

Possible extension of the result along the following directions:

- Non-Gaussian noise.

(i) Transformation. Brown, Cai and Zhou (2007, AOS), Brown, Cai, Zhang, Zhao and Zhou (2007), Brown, Cai and Zhou (2008).

(ii) Direct approach. For the double exponential noise

$$\text{Pen}(\theta) \approx k \left(\log \frac{n}{k} \right)^2$$

- Dependence.

Example

$$Y_i = \theta_i + Z + Z_i, \quad Z_i \stackrel{iid}{\sim} N(0, 1), \quad Z \sim N(0, \tau^2) \quad i = 1, 2, \dots, n.$$

- Estimating proportion of true nulls.
- Large Covariance Matrices Estimation.

3. Summary

- Conjecture 1.2. in ABDJ (2006) is true.
- An extension of the conjecture implies a range of model selection procedures are asymptotically sharp minimax.
- Penalties $(2 \pm \epsilon) k \log \frac{n}{k}$ don't work for some $\epsilon > 0$.
- Connections to Bayesian model selection, MDL theory, wavelet Gaussian estimation and general nonparametric estimation.