Can the Strengths of AIC and BIC Be Shared?

Eichel, P. J., & Donghui Yan Yang, Y. (2005)

Sparse Estimators and the Oracle Property, or the Return of Hodges' Estimator

Leeb. H., & Pötshcer, B. M. (2008)

Presenter: Sarah Kim

2017.01.10

Outline

Can consistency and minimax rate optimality be shared?

Sparse estimators and the oracle property

- Bad risk behavior of sparse estimator
- Numerical results

Setup

- Consider the regression model

$$y_i = f(\mathbf{x}_i) + \varepsilon_i, \quad i = 1, 2, \ldots, n,$$

where $\mathbf{x}_i = (x_{i1}, \ldots, x_{id})$, f is the true regression function, and $\varepsilon_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$.

- To estimate f, we consider linear models as follows:

$$y = f_k(\mathbf{x}, \theta_k) + \varepsilon,$$

where for each k, $\mathcal{F}_k = \{f_k(\mathbf{x}, \theta_k), \theta_k \in \Theta_k\}$ is a linear familiy of regression functions with θ_k being the parameter of a finite dimension m_k.

Minimax property of AIC

For a model selection criterion δ that select model $\hat{k}$, let

$$ASE(f_{\hat{k}}) = \frac{1}{n}\sum_{i=1}^{n}\left(f(\mathbf{x}_i) - f_{\hat{k}}(\mathbf{x}_i, \hat{\theta}_{\hat{k}})\right)^2,$$

where $\hat{\theta}_{\hat{k}}$ is the LSE. The corresponding risk is

$$R(f;\delta;n) = \frac{1}{n}\sum_{i}\mathbb{E}\left(f(\mathbf{x}_i) - f_{\hat{k}}(\mathbf{x}_i, \hat{\theta}_{\hat{k}})\right)^2.$$

Definition (Minimax-rate optimal)

δ is said to be minimax-rate optimal over a class of regression functions $\mathcal{F}$ if $\sup_{f\in\mathcal{F}} R(f;\delta;n)$ converges at the same rate as $\inf_{\hat{f}}\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left(f(\mathbf{x}_i) - \hat{f}(\mathbf{x}_i)\right)^2$, where $\hat{f}$ is over all estimators based on the observations of $y_1,\ldots,y_n$.

Minimax property of AIC

Notations

- Γ: the collection of all the models being considered
- N_m: the number of models that have the same dimension m in Γ
  (Here we assume that $\exists c > 0$ such that $N_m \leq e^{cm}$.)
- M_k: the projection matrix of model k, r_k: the rank of M_k

Proposition (Yang, 1999)

There exists $C > 0$ depending only on c such that for every f, we have

$$R(f; \delta_{AIC}; n) \leq C \inf_{k \in \Gamma} \left(\frac{\|f - M_k f\|_n^2}{n} + \frac{r_k}{n} \right).$$

Corollary

Suppose that model $k^* \in \Gamma$ is the true model. Then

$$\sup_{f \in \mathcal{F}_{k^*}} R(f; \delta_{AIC}; n) \leq \frac{C m_{k^*}}{n}.$$

Can consistency and minimax rate optimality be shared?

Assumptions (A1)

There exist two models $k_1, k_2 \in \Gamma$ such that

1. $\mathcal{F}_{k_1} = \{f_{k_1}(\mathbf{x}, \theta_{k_1}) : \theta_{k_1} \in \Theta_{k_1}\}$ is a sub-linear space of $\mathcal{F}_{k_2} = \{f_{k_2}(\mathbf{x}, \theta_{k_2}) : \theta_{k_2} \in \Theta_{k_2}\}$;
2. $\exists \phi(\mathbf{x}) \in \mathcal{F}_{k_2}$ orthogonal to $\mathcal{F}_{k_1}$ with $\frac{1}{n}\sum_{i=1}^{n} \phi^2(\mathbf{x}_i)$ being bounded between two positive constants;
3. $\exists f_0 \in \mathcal{F}_{k_1}$ such that f_0 is not in any family $\mathcal{F}_k, k \in \Gamma$ that does not contain $\mathcal{F}_{k_1}$.

Theorem 1

Under A1, if any model selection method δ is consistent in selection, then we must have

$$n \sup_{f \in \mathcal{F}_{k_2}} R(f; \delta; n) \to \infty.$$

Theorem 1 proof (1)

In a simple case, suppose that we have two models, model 0: $y_i = \alpha + \varepsilon_i$ and model 1: $y_i = \alpha + \beta x_i + \varepsilon_i$, $i = 1, 2, \ldots, n$.

- Assume that $\bar{x}_n = \frac{1}{n}\sum_i x_i = 0$ and $\frac{1}{n}\sum_i x_i^2$ is bounded between two positive constants for all n.
- Let δ denote a consistent model selection criterion and let A_n be the event that model 1 is selected.
- Then under the squared error loss,

$$\mathbb{E}(f(x_i) - \hat{f}(x_i))^2 = \frac{\sigma^2}{n} + x_i^2 \mathbb{E}(\hat{\beta} I_{A_n} - \beta)^2 + 2x_i \mathbb{E}(\hat{\alpha} - \alpha)(\hat{\beta} I_{A_n} - \beta)$$

and we have

$$R(f; \delta; n) = \frac{\sigma^2}{n} + \left(\frac{1}{n}\sum_{i=1}^{n} x_i^2\right)\mathbb{E}(\hat{\beta} I_{A_n} - \beta)^2,$$

where $(\hat{\alpha}, \hat{\beta})$ is the LSE.

Theorem 1 proof (2)

- Note that $\sup_{|\beta|\leq c} \mathbb{E}_\beta(\sqrt{n}\hat{\beta} I_{A_n} - \sqrt{n}\beta)^2 = \sup_{|\beta|\leq c} [\mathbb{E}_\beta n(\hat{\beta} - \beta)^2 I_{A_n} + n\beta^2 \mathbb{P}_\beta(A_n^c)]$.
- To show $n \sup_f R(f; \delta; n) \to \infty$, it suffices to show that for $\forall c > 0$,

$$\sup_{|\beta|\leq c} n\beta^2 \mathbb{P}_\beta(A_n^c) \to \infty. \tag{1}$$

- Since δ is consistent, we have $\mathbb{P}_{\beta=0}(A_n) \to 0$ as $n \to \infty$.
- Consider a testing problem with $H_0 : \beta = 0$ vs $H_1 : \beta > 0$. If we take A_n as the rejection region, δ becomes a testing rule with probability of type 1 error approaching 0.

Theorem 1 proof (3)

- Here we assume $\varepsilon_i \overset{i.i.d}{\sim} \mathcal{N}(0,1)$ and $\alpha = 0$.
- Let $f(y_1, \ldots, y_n; \beta)$ denote the joint probability density function. Then $\{f_\beta\}$ has a monotone likelihood ratio in $T(\mathbf{x}) = \sum_{i=1}^{n} x_i y_i$.
- By the Karlin-Rubin Theorem, a UMP test exists which is to reject H_0 when $\sum x_i y_i$ is larger than some constant . Choose h_n so that $\mathbb{P}_{\beta=0}(\sum x_i y_i \geq h_n) = \mathbb{P}_{\beta=0}(A_n)$.
- Let $A_{n,*}$ denote the event $\{\sum x_i y_i \geq h_n\}$. Then we have for all $\beta > 0$, $\mathbb{P}_\beta(A_{n,*}) \geq \mathbb{P}_\beta(A_n)$. Hence $\sup_{|\beta| \leq c} n\beta^2 \mathbb{P}_\beta(A_n^c) \geq \sup_{0 \leq \beta \leq c} n\beta^2 \mathbb{P}_\beta(A_{n,*}^c)$.
- Let $\beta_n = \min\left(\frac{h_n}{2\sum x_i^2}, c\right)$. Then we have $\sup_{0 \leq \beta \leq c} n\beta^2 \mathbb{P}_\beta(A_{n,*}^c) \geq n\beta_n^2 \mathbb{P}_{\beta_n}(A_{n,*}^c)$. Since $\sum x_i y_i$ is normally distributed, we can show that $\mathbb{P}_{\beta_n}(A_{n,*}^c) \to 1$.

Theorem 1 proof (4)

In general case, under A1

1. $\mathcal{F}_{k_1}$ is a sub-linear space of $\mathcal{F}_{k_2}$;
2. Let $\phi(x) \in \mathcal{F}_{k_2}$ that is orthogoanl to $\mathcal{F}_{k_1}$;
3. Let $f_0 \in \mathcal{F}_{k_1}$ such that f_0 doe not belong to any other $\mathcal{F}_k$ that does not contain $\mathcal{F}_{k_1}$.

Let B_n be the event that model k_1 is not selected for δ. If δ is consistent, then $P_{f_0}(B_n) \to 0$ as $n \to \infty$. Consider a simplified model $y_i = f_0(\mathbf{x}_i) + \beta\phi(\mathbf{x}_i) + \varepsilon_i$ and the testing problem $H_0 : \beta = 0$ vs $H_1 : \beta > 0$. Denote

- $\mathbf{f} = (f(\mathbf{x}_1), \ldots, f(\mathbf{x}_n))'$, $\mathbf{y} = (y_1, \ldots, y_n)'$, $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)'$, $\boldsymbol{\phi} = (\phi(\mathbf{x}_1), \ldots, \phi(\mathbf{x}_n))'$
- M_{k_1}: the projection matrix of model k_1.

Then we have

$$\begin{aligned}
\|\mathbf{f} - M_{k_1}\mathbf{y}\|_n^2 &= \|\mathbf{f} - M_{k_1}\mathbf{f}\|_n^2 + \boldsymbol{\varepsilon}' M_{k_1} \boldsymbol{\varepsilon} \\
&= \|\beta\boldsymbol{\phi} - \beta M_{k_1}\boldsymbol{\phi}\|_n^2 + \boldsymbol{\varepsilon}' M_{k_1} \boldsymbol{\varepsilon} \\
&= \beta^2\|\boldsymbol{\phi}\|_n^2 + \boldsymbol{\varepsilon}' M_{k_1} \boldsymbol{\varepsilon}.
\end{aligned}$$

Theorem 1 proof (5)

Then the risk of the estimator associated with δ is

$$\begin{aligned} R(f;\delta;n) &= \frac{1}{n}\sum_{k\in\Gamma}\mathbb{E}_\beta\|\mathbf{f}-M_k\mathbf{y}\|_n^2 I_{\{\hat{k}=k\}} \\ &\geq \frac{1}{n}\mathbb{E}_\beta\|\mathbf{f}-M_{k_1}\mathbf{y}\|_n^2 I_{\{\hat{k}=k_1\}} \\ &\geq \frac{\beta^2}{n}\mathbb{E}_\beta\|\phi\|_n^2 I_{\{\hat{k}=k_1\}} \\ &= \frac{\sum_i \phi^2(\mathbf{x}_i)}{n}\beta^2\mathbb{P}_\beta(\hat{k}=k_1). \end{aligned} \tag{2}$$

Hence it is enough to show $\sup_{|\beta|\leq c} n\beta^2\mathbb{P}_\beta(B_n^c)\to\infty$.

- Let $z_i = y_i - f_0(\mathbf{x}_i)$. Then $z_1,\ldots,z_n$ are indep. with $z_i\sim\mathcal{N}(\beta\phi(\mathbf{x}_i),\sigma^2)$.
- $\{f(z_1,\ldots,z_n;\beta)\}$ has a MLR in $T(\mathbf{x})=\sum_i z_i\phi(\mathbf{x}_i)$.
- By the Karlin-Rubin Thm, and UMP test property we can show

$$\sup_{|\beta|\leq c} n\beta^2\mathbb{P}_\beta(B_n^c)\to\infty$$

Theorem 1 proof (6)

- Let h_n so that $\mathbb{P}_{\beta=0}(\sum_{i=1}^n z_i\phi(\mathbf{x}_i) > h_n) = \mathbb{P}_{\beta=0}(B_n) \to 0$. Let $B_{n,*}$ be denote the event $\{\sum_{i=1}^n z_i\phi(\mathbf{x}_i) \geq d_n\}$. Then a UMP test exists which is to reject H_0 when $\sum_i z_i\phi(\mathbf{x}_i)$ is larger than h_n.
- By the UMP property of $B_{n,*}$, for any $\beta > 0$, $\mathbb{P}_\beta(B_{n,*}) \geq \mathbb{P}_\beta(B_n)$, hence

$$\sup_{|\beta|\leq c} n\beta^2\mathbb{P}_\beta(B_n^c) \geq \sup_{0\leq\beta\leq c} n\beta^2\mathbb{P}_\beta(B_{n,*}^c)$$

- Let $\beta_n = \min\left(\frac{h_n}{2\sum\phi^2(\mathbf{x}_i)}, c\right)$. Then

$$\sup_{0\leq\beta\leq c} n\beta^2\mathbb{P}_\beta(B_{n,*}^c) \geq n\beta_n^2\mathbb{P}_{\beta_n}(B_{n,*}^c)$$

  and (note that $\mathbf{h}_n/\sqrt{n} \to \infty$ as $n \to \infty$.)

$$\mathbb{P}_{\beta_n}(B_{n,*}^c) = \mathbb{P}_{\beta_n}(\sum_i z_i\phi(\mathbf{x}_i) < h_n) = \mathbb{P}\Big(\mathcal{N}(0,1) < \frac{h_n - \beta_n\sum_i\phi^2(\mathbf{x}_i)}{\sqrt{\sigma^2\sum_i\phi^2(\mathbf{x}_i)}}\Big)$$
$$\geq \mathbb{P}\Big(\mathcal{N}(0,1) < \frac{h_n}{2\sqrt{\sigma^2\sum_i\phi^2(\mathbf{x}_i)}}\Big) \to 1, \quad n \to \infty.$$

Outline

Can consistency and minimax rate optimality be shared?

Sparse estimators and the oracle property

- Bad risk behavior of sparse estimator
- Numerical results

Bad risk behavior of sparse estimator

- Consider

$$y_i = \mathbf{x}_i'\theta + \varepsilon_i, \quad i = 1, \ldots, n,$$

  where $\mathbf{x}_i \in \mathbb{R}^d$ satisfy $\frac{1}{n}\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i' \to Q > 0$ as $n \to \infty$.
- $\varepsilon_i \overset{i.i.d.}{\sim} f$ with mean 0 and variance 1 where a density f possesses an absolutely continuous derivative df/dx satisfying

$$0 < \int_{-\infty}^{\infty} \left\{ (df(x)/dx)/f(x) \right\}^2 f(x)dx < \infty.$$

- Let $\mathbb{P}_{n,\theta}$ denote the distribution of the sample $(y_1, \ldots, y_n)'$ and let $\mathbb{E}_{n,\theta}$ denote the corresponding expectation operator.
- For $\theta \in \mathbb{R}^d$, let $r(\theta)$ denote a $d \times 1$ vector where $r_i(\theta) = 0$ if $\theta_i = 0$ and $r_i(\theta) = 1$ if $\theta_i \neq 0$.

Bad risk behavior of sparse estimator

Sparsity-type condition

An estimator $\hat{\theta}$ for θ based on the sample $(y_1, \ldots, y_n)'$ is said to satisfy the sparsity-type condition if for every $\theta \in \mathbb{R}^d$

$$\mathbb{P}_{n,\theta}\left(r(\hat{\theta}) \leq r(\theta)\right) \to 1 \quad \text{as } n \to \infty, \tag{3}$$

where the inequality sign is to be interpreted componentwise.

Theorem 2

Let $\hat{\theta}$ be an aribitrary estimator for θ that satisties the sparsity-type condition (3). Then

$$\sup_{\theta \in \mathbb{R}^d} \mathbb{E}_{n,\theta}\left[n(\hat{\theta} - \theta)'(\hat{\theta} - \theta)\right] \to \infty \tag{4}$$

for $n \to \infty$. More generally, let $\ell : \mathbb{R}^d \to \mathbb{R}$ be a nonnegative loss functon. Then

$$\sup_{\theta \in \mathbb{R}^d} \mathbb{E}_{n,\theta} \ell(\sqrt{n}(\hat{\theta} - \theta)) \to \sup_{s \in \mathbb{R}^d} \ell(s) \tag{5}$$

for $n \to \infty$.

Theorem 2 proof

Let $\theta_n = -n^{-1/2}s$, $s \in \mathbb{R}^d$ arbitrary. And note that

$$\sup_{u\in\mathbb{R}^d} \ell(u) = \sup_{\theta\in\mathbb{R}^d} \ell(\sqrt{n}(\hat{\theta}-\theta)) = \mathbb{E}_{n,\theta} \sup_{\theta\in\mathbb{R}^d} \ell(\sqrt{n}(\hat{\theta}-\theta)) \\ \geq \mathbb{E}_{n,\theta}\ell(\sqrt{n}(\hat{\theta}-\theta)), \quad \forall\theta\in\mathbb{R}^d \tag{6}$$

Then we have

$$\sup_{u\in\mathbb{R}^d} \ell(u) \geq \sup_{\theta\in\mathbb{R}^d} \mathbb{E}_{n,\theta}\ell(\sqrt{n}(\hat{\theta}-\theta)) \geq \mathbb{E}_{n,\theta_n}\ell(\sqrt{n}(\hat{\theta}-\theta_n)) \\ \geq \mathbb{E}_{n,\theta_n}\left[\ell(\sqrt{n}(\hat{\theta}-\theta_n))I(\hat{\theta}=0)\right] = \ell(-\sqrt{n}\theta_n)\mathbb{P}_{n,\theta_n}(r(\hat{\theta})=0) \\ = \ell(s)\mathbb{P}_{n,\theta_n}(r(\hat{\theta})=0). \tag{7}$$

- By the sparsity-type condition, $\mathbb{P}_{n,0}(r(\hat{\theta})=0) \to 1$ as $n\to\infty$.
- Under our assumption, the model is locally asymptotical normal. Hence we can show that the sequence of probability measures $\mathbb{P}_{n,\theta_n}$ is contiguous w.r.t. $\mathbb{P}_{n,0}$.

So, $\sup_{\theta\in\mathbb{R}^d} \mathbb{E}_{n,\theta}\ell(\sqrt{n}(\hat{\theta}-\theta)) \to \sup_{s\in\mathbb{R}^d} \ell(s)$ as $n\to\infty$. ■

Let $(\Omega_n, \mathcal{A}_n)$ be measurable spaces, each equipped with a pair of probability measures P_n and Q_n.

Definition (Contiguity)

The sequence Q_n is contiguous with respect to the sequence P_n if $P_n(A_n) \to 0$ implies $Q_n(A_n) \to 0$ for every sequence of measurable sets A_n. This is denoted $Q_n \triangleleft P_n$.

Lemma 1 (Le Cam's first lemma)

$$Q_n \triangleleft P_n$$

$\Leftrightarrow$ If dP_n/dQ_n converges in distribution under Q_n to a random variable U, then $P(U > 0) = 1$.

Lemma 2

Consider the linear model $y_i = \mathbf{x}_i'\theta + \varepsilon_i$, $i = 1, \ldots, n$, where $\mathbf{x}_i \in \mathbb{R}^d$ and $\varepsilon_i \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$. If $\theta_n - \vartheta_n = O(n^{-1/2})$ and $n^{-1}\sum_{i=1}^n \mathbf{x}_i'\mathbf{x}_i \to Q > 0$ as $n \to \infty$, then $\mathbb{P}_{n,\vartheta_n}$ is contiguous w.r.t. $\mathbb{P}_{n,\theta_n}$.

Lemma 2 proof

Suppose that $\theta_n - \vartheta_n = O(n^{-1/2})$ and $n^{-1}\sum_{i=1}^n \mathbf{x}_i'\mathbf{x}_i \to Q > 0$ as $n \to \infty$. First, we will show that $d\mathbb{P}_{n,\theta_n}/d\mathbb{P}_{n,\vartheta_n} \xrightarrow{d} U$ under $\mathbb{P}_{n,\vartheta_n}$ where U is almost surely positive. Then by Le Cam's first lemma, $\mathbb{P}_{n,\vartheta_n} \triangleleft \mathbb{P}_{n,\theta_n}$.

- Under $\mathbb{P}_{n,\vartheta_n}$, $y_i \overset{indep.}{\sim} \mathcal{N}(\mathbf{x}_i'\vartheta_n, 1)$, for $i = 1, \ldots, n$. Hence

$$\log \frac{d\mathbb{P}_{n,\theta_n}}{d\mathbb{P}_{n,\vartheta_n}} = \sum_{i=1}^n \mathbf{x}_i'(\theta_n - \vartheta_n)y_i + \sum_{i=1}^n \frac{(\mathbf{x}_i'\vartheta_n)^2 - (\mathbf{x}_i'\theta_n)^2}{2} \sim \mathcal{N}(-\frac{1}{2}A_n, A_n),$$

  where $A_n = (\theta_n - \vartheta_n)' \sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i'(\theta_n - \vartheta_n)$.

- Since $\theta_n - \vartheta_n = O(n^{-1/2})$, any subsequence contain a further subsequence s.t. along this subsequence $\sqrt{n}(\theta_n - \vartheta_n) \to \alpha$ for some $\alpha \in \mathbb{R}^d$.
- $\log\left(d\mathbb{P}_{n,\theta_n}/d\mathbb{P}_{n,\vartheta_n}\right)$ converges in distribution under $\mathbb{P}_{n,\vartheta_n}$ to $Z \sim \mathcal{N}(-A/2, A)$ where $A = \alpha' Q\alpha$. Hence $d\mathbb{P}_{n,\theta_n}/d\mathbb{P}_{n,\vartheta_n} \xrightarrow{d} \exp(Z)$ which is always positive. ■

Numerical results on the SCAD estimator

Consider $y_i = \mathbf{x}_i'\theta + \varepsilon_i$, $i = 1, \ldots, n$

where

- $\theta \in \mathbb{R}^d$ and $d = 8$, $\varepsilon_i \overset{i.i.d.}{\sim} \mathcal{N}(0, 1)$, $n = 60, 120, 240, 480, 960$;
- $\mathbf{x}_i$ are $\mathcal{N}_d(0, \Sigma)$, $\Sigma_{ij} = \rho^{|i-j|}$ with $\rho = 0.5$;
- True parameter: $\theta_n = \theta_0 + (\gamma/\sqrt{n}) \times \eta$, $\theta_0 = (3, 1.5, 0, 0, 2, 0, 0, 0)'$, $\eta = (0, 0, 1, 0, 1, 0, 1, 1, 1)'$, γ is the sequence with length 101 from 0 to 8.

Tuning parameter of SCAD estimator:

- $a = 3.7$ (Fan and Li, 2001);
- the range of λ's : $\{\delta \frac{\hat{\sigma}}{\sqrt{n}} \frac{\log n}{\log 60} : \delta = 0.9, 1.1, 1.3, \ldots, 2\}$, $\hat{\sigma}^2$ denotes $SSE/(n-d)$ from a least-squares fit.

✓ Then $\lambda \to 0$ and $\sqrt{n}\lambda \to \infty$, it guarantees that the resulting SCAD estimator possesses the Oracle property.

Numerical results on the SCAD estimator

Two types of performance measures are considered:

- Median relative model error, $ME(\hat{\theta}) = (\hat{\theta} - \theta)'\Sigma(\hat{\theta} - \theta)$;
- Relative mean squared error, $RE(\hat{\theta}) = ME(\hat{\theta})/ME(\hat{\theta}_{LS})$.

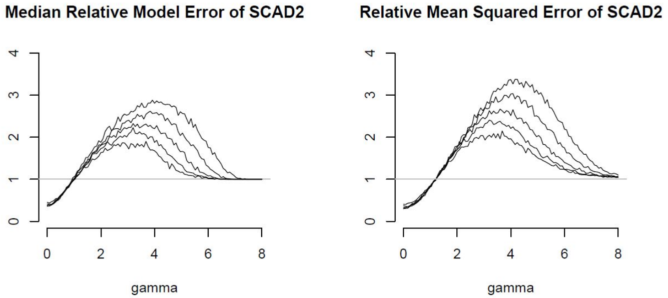

Figure 1 : Monte Carlo performance estimates for ME, RE, under the trur parameter $\theta_n = \theta_0 + (\gamma/\sqrt{n})(0, 0, 1, 1, 0, 1, 1, 1)'$ each based on 500 Monte Carlo replications. Larger sample size correspond to larger maximal errors.