Strong Consistency of the AIC, BIC, C_p and KOO Methods in High-Dimensional-Response Regression

Jiang Hu*

(Joint work with Zhidong Bai* and Yasunori Fujikoshi†)

*Northeast Normal University, P. R. China

†Hiroshima University, Japan

December, 2019

Outline

1. **Model selection**
   - Linear regression model
   - Classical selection criteria
2. **Asymptotic properties**
   - Low-dimensional
   - Large-dimension and small-model
3. **Main results**
   - Assumptions and notations
   - Strong consistency of AIC, BIC and C_p
   - KOO methods based on the AIC, BIC, and C_p
   - General KOO methods
4. **Proof strategy**
5. **Simulation**

Outline

1. Model selection
   - Linear regression model
   - Classical selection criteria
2. Asymptotic properties
   - Low-dimensional
   - Large-dimension and small-model
3. Main results
   - Assumptions and notations
   - Strong consistency of AIC, BIC and C_p
   - KOO methods based on the AIC, BIC, and C_p
   - General KOO methods
4. Proof strategy
5. Simulation

Linear regression model

Consider the multi-response linear regression model:

$$\underset{1\times p}{\mathbf{y}} = \underset{1\times k}{\mathbf{x}} \cdot \underset{k\times p}{\Theta} + \underset{1\times p}{\mathbf{e}} \cdot \underset{p\times p}{\mathbf{\Sigma}}^{1/2} \quad (1)$$

Aim: find the **TRUE** model if it exits.

References:

[1] Miller ALan. Subset Selection in Regression, Second Edition. Chapman and Hall/CRC, 2002.
[2] Gerda Claeskens, Nils Lid Hjort. Model Selection and Model Averaging. Vol. 330. Cambridge University Press Cambridge, 2008.

Overview of classical model selection criteria

From the point of view of statistical performance of a method, and intended context of its use, there are only two distinct classes of methods: labeled efficient and consistent.

Generally there are two main approaches:

(I) Optimization of some selection criteria;

- **(1)** Criteria based on some form of mean squared error (e.g., Mallows's C_p, Mallows 1973) or mean squared prediction error (e.g., PRESS, Allen 1970);
- **(2)** Criteria that are estimates of Kullback-Leibler (K-L) information or distance (e.g., AIC, AICc, and QAICc);
- **(3)** Criteria that are consistent estimators of the "true model" (e.g., BIC).

(II) Tests of hypotheses.

Notation

Observations: $\mathbf{Y} : n \times p$ and $\mathbf{X}_{\boldsymbol{\omega}} = (\mathbf{x}_1, \dots, \mathbf{x}_k) : n \times k$.

Notations: $\boldsymbol{\omega} = \{1, \dots, k\}$, $\mathbf{j}_* \in \boldsymbol{\omega}$, $\mathbf{j} \in \boldsymbol{\omega}$, $k_{\mathbf{j}} =$ the cardinality of $\mathbf{j}$.

- Full model $\boldsymbol{\omega}$: $\mathbf{Y} = \mathbf{X}_{\boldsymbol{\omega}} \cdot \Theta_{\boldsymbol{\omega}} + \mathbf{E} \cdot \mathbf{\Sigma}^{1/2}$.
- True model $\mathbf{j}_*$: $\mathbf{Y} = \mathbf{X}_{\mathbf{j}_*} \cdot \Theta_{\mathbf{j}_*} + \mathbf{E} \cdot \mathbf{\Sigma}^{1/2}$.
- Candidate model $\mathbf{j}$: $\mathbf{Y} = \mathbf{X}_{\mathbf{j}} \cdot \Theta_{\mathbf{j}} + \mathbf{E} \cdot \mathbf{\Sigma}^{1/2}$.
- $\Theta_{\mathbf{j}} = (\theta_{ji},\ j \in \mathbf{j},\ i = 1, \dots, p)$
- $\mathbf{X}_{\mathbf{j}} = (\mathbf{x}_j, j \in \mathbf{j})$
- $\mathbf{P}_{\mathbf{j}} = \mathbf{X}_{\mathbf{j}}(\mathbf{X}_{\mathbf{j}}'\mathbf{X}_{\mathbf{j}})^{-1}\mathbf{X}_{\mathbf{j}}'$
- $\widehat{\mathbf{\Sigma}}_{\mathbf{j}} = n^{-1}\mathbf{Y}'(\mathbf{I}_n - \mathbf{P}_{\mathbf{j}})\mathbf{Y}$

Classical selection criteria

- Akaike's information criterion (AIC, Akaike (1973,1974)):

$$AIC_{\mathbf{j}} = n \log |\widehat{\boldsymbol{\Sigma}}_{\mathbf{j}}| + 2k_{\mathbf{j}}p \quad \text{and} \quad \hat{\mathbf{j}}_A = \arg\min AIC_{\mathbf{j}}$$

Key: Kullback-Leibler information/distance

Kullback-Leibler Information

Kullback-Leibler information between density functions f and g is defined for continuous functions

$$I(f, g) = \int f(x) \log \left(\frac{f(x)}{g(x)} \right) dx.$$

The notation $I(f, g)$ denotes the "information lost when g is used to approximate f." As a heuristic interpretation, $I(f, g)$ is the distance from g to f.

Classical selection criteria

- Akaike's information criterion (AIC, Akaike (1973,1974)):

$$AIC_{\mathbf{j}} = n \log |\widehat{\mathbf{\Sigma}}_{\mathbf{j}}| + 2k_{\mathbf{j}}p \quad \text{and} \quad \hat{\mathbf{j}}_A = \arg\min AIC_{\mathbf{j}}$$

Key: Kullback-Leibler information/distance

Kullback-Leibler Information

Kullback-Leibler information between density functions f and g is defined for continuous functions

$$I(f,g) = \int f(x) \log\left(\frac{f(x)}{g(x)}\right) dx.$$

The notation $I(f,g)$ denotes the "information lost when g is used to approximate f." As a heuristic interpretation, $I(f,g)$ is the distance from g to f.

Classical selection criteria

- Bayesian information criterion (BIC, Schwarz (1978), Akaike (1977, 1978)) :

$$BIC_{\mathbf{j}} = n \log |\widehat{\boldsymbol{\Sigma}}_{\mathbf{j}}| + \log(n) k_{\mathbf{j}} p \quad \text{and} \quad \hat{\mathbf{j}}_B = \arg\min BIC_{\mathbf{j}}$$

Key: Consistence

Consistence

As $n \to \infty$, under some conditions, $\hat{\mathbf{j}}_B \to \mathbf{j}_*$ almost surely.

Classical selection criteria

- Bayesian information criterion (BIC, Schwarz (1978), Akaike (1977, 1978)) :

$$BIC_{\mathbf{j}} = n \log |\widehat{\mathbf{\Sigma}}_{\mathbf{j}}| + \log(n) k_{\mathbf{j}} p \quad \text{and} \quad \hat{\mathbf{j}}_B = \arg\min BIC_{\mathbf{j}}$$

Key: Consistence

Consistence

As $n \to \infty$, under some conditions, $\hat{\mathbf{j}}_B \to \mathbf{j}_*$ almost surely.

Classical selection criteria

- Mallows's C_p (C_p, Mallows (1973)):

$$C_{p\mathbf{j}} = (n-k)\mathrm{tr}(\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\omega}}^{-1}\widehat{\boldsymbol{\Sigma}}_{\mathbf{j}}) + 2pk_{\mathbf{j}} \quad \text{and} \quad \hat{\mathbf{j}}_C = \arg\min C_{p\mathbf{j}}$$

Key: Mean squared error

Remark 1

Atilgan (1996) provides a relationship between AIC and Mallows's C_p, shows that under some conditions AIC selection behaves like minimum mean squared error selection, and notes that AIC and C_p are somewhat equivalent criteria.

Outline

Low-dimensional

Assume k and p are fixed (Fujikoshi, 1985; Fujikoshi and Veitch, 1979).

- If $\mathbf{j}$ is an over-specified model, i.e., $\mathbf{j}_* \subset \mathbf{j}$,

$$\mathbb{P}(AIC_{\mathbf{j}} - AIC_{\mathbf{j}_*} < 0) \sim \mathbb{P}(\chi^2_{k_{\mathbf{j}} - k_{\mathbf{j}_*}} > 2(k_{\mathbf{j}} - k_{\mathbf{j}_*})) > 0$$
$$\mathbb{P}(BIC_{\mathbf{j}} - BIC_{\mathbf{j}_*} < 0) \sim \mathbb{P}(\chi^2_{k_{\mathbf{j}} - k_{\mathbf{j}_*}} > \log(n)(k_{\mathbf{j}} - k_{\mathbf{j}_*})) \to 0$$
$$\mathbb{P}(C_{p\mathbf{j}} - C_{p\mathbf{j}_*} < 0) \sim \mathbb{P}(\chi^2_{k_{\mathbf{j}} - k_{\mathbf{j}_*}} > 2(k_{\mathbf{j}} - k_{\mathbf{j}_*})) > 0$$

- If $\mathbf{j}$ is an under-specified model, i.e., $\mathbf{j}_* \not\subset \mathbf{j}$,

$$\begin{aligned} AIC_{\mathbf{j}} - AIC_{\mathbf{j}_*} =& O(n) \to +\infty \\ BIC_{\mathbf{j}} - BIC_{\mathbf{j}_*} =& O(n) \to +\infty \\ C_{p\mathbf{j}} - C_{p\mathbf{j}_*} =& O(n) \to +\infty \end{aligned}$$

Large-dimension and small-model

Assume $\mathbf{j}_* \in \boldsymbol{\omega}$ is the true model, k is fixed and $p/n \to c \in (0,1)$.

Theorem 4.1 in (Fujikoshi et al., 2014)

If $c \in (0, c_a \approx 0.797)$ where $\log(1-c_a) + 2c_a = 0$ and for any $\mathbf{j}_* \not\subset \mathbf{j}$ with $k_{\mathbf{j}} - k_{\mathbf{j}_*} \leq 0$,

$$\lim \log(|\mathbf{I} + \Phi_{\mathbf{j}}|) > (k_{\mathbf{j}_*} - k_{\mathbf{j}})[2c + \log(1-c)]$$

where $\Phi_{\mathbf{j}} = \frac{1}{n}\boldsymbol{\Sigma}^{-\frac{1}{2}}\Theta'_{\mathbf{j}_*}\mathbf{X}'_{\mathbf{j}_*}(\mathbf{P}_{\boldsymbol{\omega}} - \mathbf{P}_{\mathbf{j}})\mathbf{X}_{\mathbf{j}_*}\Theta_{\mathbf{j}_*}\boldsymbol{\Sigma}^{-\frac{1}{2}}$. Then,

$$\lim_{p/n \to c} \mathbb{P}(\hat{\mathbf{j}}_A = \mathbf{j}_*) = 1.$$

Otherwise,

$$\lim_{p/n \to c} \mathbb{P}(\hat{\mathbf{j}}_A = \mathbf{j}_*) \neq 1.$$

What about BIC?

Large-dimension and small-model

Assume $\mathbf{j}_* \in \boldsymbol{\omega}$ is the true model, k is fixed and $p/n \to c \in (0,1)$.

Theorem 4.1 in (Fujikoshi et al., 2014)

If $c \in (0, c_a \approx 0.797)$ where $\log(1-c_a) + 2c_a = 0$ and for any $\mathbf{j}_* \not\subset \mathbf{j}$ with $k_{\mathbf{j}} - k_{\mathbf{j}_*} \leq 0$,

$$\lim \log(|\mathbf{I} + \Phi_{\mathbf{j}}|) > (k_{\mathbf{j}_*} - k_{\mathbf{j}})[2c + \log(1-c)]$$

where $\Phi_{\mathbf{j}} = \frac{1}{n}\boldsymbol{\Sigma}^{-\frac{1}{2}}\Theta'_{\mathbf{j}_*}\mathbf{X}'_{\mathbf{j}_*}(\mathbf{P}_{\boldsymbol{\omega}} - \mathbf{P}_{\mathbf{j}})\mathbf{X}_{\mathbf{j}_*}\Theta_{\mathbf{j}_*}\boldsymbol{\Sigma}^{-\frac{1}{2}}$. Then,

$$\lim_{p/n \to c} \mathbb{P}(\hat{\mathbf{j}}_A = \mathbf{j}_*) = 1.$$

Otherwise,

$$\lim_{p/n \to c} \mathbb{P}(\hat{\mathbf{j}}_A = \mathbf{j}_*) \neq 1.$$

What about BIC?

Large-dimension and small-model

Assume $\mathbf{j}_* \in \boldsymbol{\omega}$ is the true model, k is fixed and $p/n \to c \in (0,1)$.

Theorem 4.1 in (Fujikoshi et al., 2014)

If $c \in (0, 1/2)$ and for any $\mathbf{j}_* \not\subset \mathbf{j}$ with $k_{\mathbf{j}} - k_{\mathbf{j}_*} \leq 0$,

$$\mathrm{tr}(\Phi_{\mathbf{j}}) > (k_{\mathbf{j}_*} - k_{\mathbf{j}})c(1-2c)$$

where $\Phi_{\mathbf{j}} = \frac{1}{n}\boldsymbol{\Sigma}^{-\frac{1}{2}}\Theta'_{\mathbf{j}_*}\mathbf{X}'_{\mathbf{j}_*}(\mathbf{P}_{\boldsymbol{\omega}} - \mathbf{P}_{\mathbf{j}})\mathbf{X}_{\mathbf{j}_*}\Theta_{\mathbf{j}_*}\boldsymbol{\Sigma}^{-\frac{1}{2}}$. Then,

$$\lim_{p/n \to c} \mathbb{P}(\hat{\mathbf{j}}_C = \mathbf{j}_*) = 1.$$

Otherwise,

$$\lim_{p/n \to c} \mathbb{P}(\hat{\mathbf{j}}_C = \mathbf{j}_*) \neq 1.$$

Outline

Assumptions and notations

A1: The true model $\mathbf{j}_*$ is a subset of set $\boldsymbol{\omega}$ and $k_* := k_{\mathbf{j}_*}$ is fixed.

A2: $\mathbf{E} = \{e_{ij}\}$ are i.i.d. with zero means, unit variances and finite fourth moments.

A3: $\mathbf{X}'\mathbf{X}$ is (non-random) positive definite uniformly.

A4: As $\{k, p, n\} \to \infty$, $p/n \to c \in (0, 1)$, $k/n \to \alpha \in [0, 1 - c)$.

A5: $\|\Phi\| := \|\frac{1}{n}\boldsymbol{\Sigma}^{-\frac{1}{2}}\Theta'_{\mathbf{j}_*}\mathbf{X}'_{\mathbf{j}_*}\mathbf{X}_{\mathbf{j}_*}\Theta_{\mathbf{j}_*}\boldsymbol{\Sigma}^{-\frac{1}{2}}\|$ is bounded uniformly.

A5': As $\{k, p, n\} \to \infty$,
$\|\Phi_{\mathbf{j}}\| := \|\frac{1}{n}\boldsymbol{\Sigma}^{-\frac{1}{2}}\Theta'_{\mathbf{j}_*}\mathbf{X}'_{\mathbf{j}_*}(\mathbf{P}_{\boldsymbol{\omega}} - \mathbf{P}_{\mathbf{j}})\mathbf{X}_{\mathbf{j}_*}\Theta_{\mathbf{j}_*}\boldsymbol{\Sigma}^{-\frac{1}{2}}\| \to \infty$.

Assumptions and notations

Define two bivariate functions

$$\phi(\alpha, c) = 2c\alpha + \log\left(\frac{(1-c)^{1-c}(1-\alpha)^{1-\alpha}}{(1-c-\alpha)^{1-c-\alpha}}\right)$$
$$\psi(\alpha, c) = \frac{c(\alpha-1)}{1-\alpha-c} + 2c.$$

For under-specified model $\mathbf{j}$ with $k_{\mathbf{j}\cap\mathbf{j}_*^c} = m \geq 0$ and $k_{\mathbf{j}\cap\mathbf{j}_*} = s > 0$, we denote

$$\tau_{n\mathbf{j}} := (1-\alpha_m)^{s-p}|(1-\alpha_m)\mathbf{I}_p + \Phi_{\mathbf{j}}| \to \tau_{\mathbf{j}} \leq \infty$$
$$\kappa_{n\mathbf{j}} := \mathrm{tr}(\Phi_{\mathbf{j}}) \to \kappa_{\mathbf{j}} \leq \infty.$$

Strong consistency of AIC, BIC and C_p

Theorem 1 (Bai, Fujikoshi and H. (2019))

Suppose (A1)-(A5) hold.

- $\phi(\alpha, c) > 0 \Leftrightarrow$ *AIC is almost surely not over-specified;*
- *If* $\phi(\alpha, c) > 0$, *for any under-specified candidate model* $\mathbf{j}$ *with* $\log(\tau_{\mathbf{j}}) > (s - m)(\log(1 - c) + 2c) \Leftrightarrow$ *AIC is almost surely not under-specified;*

Theorem 2 (Bai, Fujikoshi and H. (2019))

Suppose (A1)-(A5) hold, BIC is almost surely under-specified;

Strong consistency of AIC, BIC and C_p

Theorem 1 (Bai, Fujikoshi and H. (2019))

Suppose (A1)-(A5) hold.

- $\phi(\alpha, c) > 0 \Leftrightarrow$ *AIC is almost surely not over-specified;*
- *If* $\phi(\alpha, c) > 0$, *for any under-specified candidate model* $\mathbf{j}$ *with* $\log(\tau_{\mathbf{j}}) > (s - m)(\log(1 - c) + 2c) \Leftrightarrow$ *AIC is almost surely not under-specified;*

Theorem 2 (Bai, Fujikoshi and H. (2019))

Suppose (A1)-(A5) hold, BIC is almost surely under-specified;

Strong consistency of AIC, BIC and C_p

Theorem 1 (Bai, Fujikoshi and H. (2019))

Suppose (A1)-(A5) hold.

- $\phi(\alpha, c) > 0 \Leftrightarrow$ *AIC is almost surely not over-specified;*
- *If* $\phi(\alpha, c) > 0$, *for any under-specified candidate model* $\mathbf{j}$ *with* $\log(\tau_{\mathbf{j}}) > (s - m)(\log(1 - c) + 2c) \Leftrightarrow$ *AIC is almost surely not under-specified;*

Theorem 2 (Bai, Fujikoshi and H. (2019))

Suppose (A1)-(A5) hold, BIC is almost surely under-specified;

Strong consistency of AIC, BIC and C_p

Theorem 1 (Bai, Fujikoshi and H. (2019))

Suppose (A1)-(A5) hold.

- $\phi(\alpha, c) > 0 \Leftrightarrow$ *AIC is almost surely not over-specified;*
- *If* $\phi(\alpha, c) > 0$, *for any under-specified candidate model* $\mathbf{j}$ *with* $\log(\tau_{\mathbf{j}}) > (s - m)(\log(1 - c) + 2c) \Leftrightarrow$ *AIC is almost surely not under-specified;*

Theorem 2 (Bai, Fujikoshi and H. (2019))

Suppose (A1)-(A5) hold, BIC is almost surely under-specified;

Strong consistency of AIC, BIC and C_p

Theorem 3 (Bai, Fujikoshi and H. (2019))

Suppose (A1)-(A5) hold.

- $\psi(\alpha, c) > 0 \Leftrightarrow C_p$ *is almost surely not over-specified;*
- *If* $\psi(\alpha, c) > 0$, *for any under-specified model* $\mathbf{j}$, *satisfying* $\kappa_{\mathbf{j}} > (s-m)\psi(\alpha, c)(1-\alpha-c)/(1-\alpha) \Leftrightarrow C_p$ *is almost surely not under-specified;*

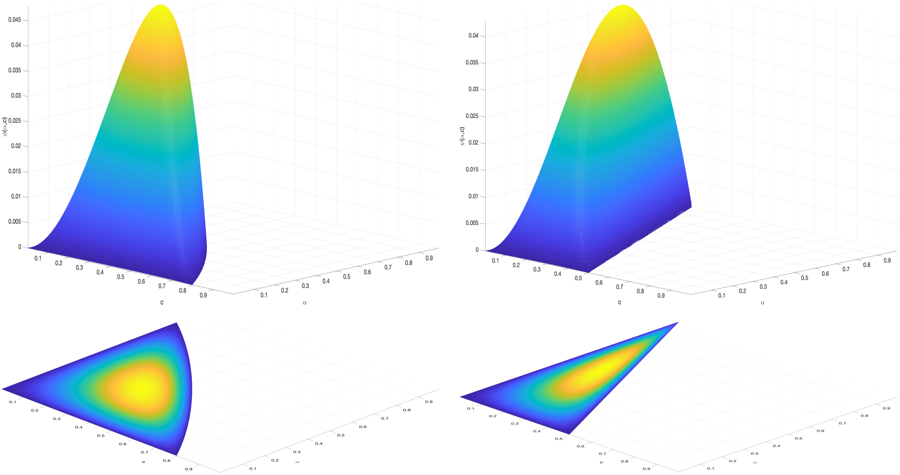


Figure: 3D plots for $\phi(\alpha, c) > 0$ and $\psi(\alpha, c) > 0$.

Strong consistency of AIC, BIC and C_p

Theorem 4 (Bai, Fujikoshi and H. (2019))

Suppose (A1)-(A4) and (A5') hold.

- $\phi(\alpha, c) > 0 \Leftrightarrow$ *AIC is almost surely not over-specified;*
- *AIC is almost surely not under-specified;*

Theorem 5 (Bai, Fujikoshi and H. (2019))

Suppose (A1)-(A4) and (A5') hold.

- *For any under-specified model* $\mathbf{j}$, $\lim_{n,p} \Big(\log(\tau_{n\mathbf{j}}) - c(s-m)\log(n) \Big) > (s-m)\log(1-c) \Leftrightarrow$ *BIC is almost surely not under-specified;*
- *BIC is almost surely not over-specified;*

Strong consistency of AIC, BIC and C_p

Theorem 6 (Bai, Fujikoshi and H. (2019))

Suppose (A1)-(A4) and (A5') hold.

- $\psi(\alpha, c) > 0 \Leftrightarrow C_p$ *is almost surely not over-specified;*
- C_p *is almost surely not under-specified;*

Remark 2

Under the condition $\phi(\alpha, c) > 0$, if the BIC is strongly consistent, then the AIC is strongly consistent but not vice versa.

KOO methods based on the AIC, BIC, and C_p

Knock-one-out (KOO) methods, which is introduced by Nishii et al. (1988), is to avoid the well known computational problem of AIC, BIC and C_p. Denote

$$\tilde{A}_j := \frac{1}{n}(AIC_{\boldsymbol{\omega}\backslash j} - AIC_{\boldsymbol{\omega}}) = \log|\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\omega}\backslash j}| - \log|\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\omega}}| - 2p/n,$$
$$\tilde{B}_j := \frac{1}{n}(BIC_{\boldsymbol{\omega}\backslash j} - BIC_{\boldsymbol{\omega}}) = \log|\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\omega}\backslash j}| - \log|\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\omega}}| - \log(n)p/n,$$
$$\tilde{C}_j := \frac{1}{n}(C_{p\boldsymbol{\omega}\backslash j} - C_{p\boldsymbol{\omega}}) = (1-k/n)\mathrm{tr}\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\omega}}^{-1}\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\omega}\backslash j} - (n-k+2)p/n.$$

Choose the model:

$$\tilde{\mathbf{j}}_A = \{j \in \boldsymbol{\omega} | \tilde{A}_j > 0\}, \quad \tilde{\mathbf{j}}_B = \{j \in \boldsymbol{\omega} | \tilde{B}_j > 0\}$$
$$\tilde{\mathbf{j}}_C = \{j \in \boldsymbol{\omega} | \tilde{C}_j > 0\}.$$

KOO methods based on the AIC, BIC, and C_p

Note that for testing

$$\boldsymbol{\theta}_j = \mathbf{0} \quad \text{v.s.} \quad \boldsymbol{\theta}_\mathbf{j} \neq \mathbf{0}$$

(1) the $-2\log$ likelihood ratio statistic under normality can be expressed as

$$n\left\{\log(|\hat{\Sigma}_{\boldsymbol{\omega}}|) - \log(|\hat{\Sigma}_{\boldsymbol{\omega}/j}|)\right\};$$

(2) the Lawley-Hotelling trace statistic under normality can be expressed as

$$(n-k)tr(\hat{\Sigma}_{\boldsymbol{\omega}}^{-1}\hat{\Sigma}_{\boldsymbol{\omega}\backslash j}).$$

(3) $\tilde{A}_j\ (\tilde{B}_j,\ \tilde{C}_j)$ is regarded as a measure that expresses the degree of contribution of $\mathbf{x_j}$ based on $A_j\ (B_j,\ C_p)$. As such, the KOO methods may also be referred to as test-based methods.

KOO methods based on the AIC, BIC, and C_p

Theorem 7 (Bai, Fujikoshi and H. (2019))

Suppose (A1)-(A5) hold.

- $\log(\frac{1-\alpha}{1-\alpha-c}) < 2c \Leftrightarrow \tilde{\mathbf{j}}_A$ *is almost surely not over-specified.*
- *If* $\log(\frac{1-\alpha}{1-\alpha-c}) < 2c$, *for any* $j \in \mathbf{j}_*$, $\log(\tau_{\boldsymbol{\omega}\backslash j}) > \log(1-\alpha-c) + 2c \Leftrightarrow \tilde{\mathbf{j}}_A$ *is almost surely not under-specified;*

Theorem 8 (Bai, Fujikoshi and H. (2019))

Suppose (A1)-(A5) hold, $\tilde{\mathbf{j}}_B$ *is almost surely under-specified.*

Theorem 9 (Bai, Fujikoshi and H. (2019))

Suppose (A1)-(A5) hold.

- $(1-\alpha) < 2(1-\alpha-c) \Leftrightarrow \tilde{\mathbf{j}}_C$ *is almost surely not over-specified.*
- *If* $(1-\alpha) < 2(1-\alpha-c)$, *for any* $j \in \mathbf{j}_*$, $\kappa_{\boldsymbol{\omega}\backslash j} > \frac{c(1-\alpha-2c)}{1-\alpha} \Leftrightarrow \tilde{\mathbf{j}}_C$ *is almost surely not under-specified;*

KOO methods based on the AIC, BIC, and C_p

Theorem 10 (Bai, Fujikoshi and H. (2019))

Suppose (A1)-(A4) and (A5') hold. $\log(\frac{1-\alpha}{1-\alpha-c}) < 2c \Leftrightarrow \tilde{\mathbf{j}}_A$ *is almost surely consistent.*

Theorem 11 (Bai, Fujikoshi and H. (2019))

Suppose (A1)-(A4) and (A5') hold.

- *For any* $j \in \mathbf{j}_*$, $[\log(\tau_{\boldsymbol{\omega}\backslash j}) - \log(n)c] > \log(1-\alpha-c), \Leftrightarrow \tilde{\mathbf{j}}_B$ *is almost surely not under-specified;*
- $\tilde{\mathbf{j}}_B$ *is almost surely not over-specified.*

Theorem 12 (Bai, Fujikoshi and H. (2019))

Suppose (A1)-(A4) and (A5') hold. $(1-\alpha) < 2(1-\alpha-c) \Leftrightarrow \tilde{\mathbf{j}}_C$ *is almost surely consistent.*

General KOO methods

Recall the KOO AIC: $\log(|\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\omega}\setminus j}|) - \log(|\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\omega}}|) - 2p/n(> 0)$;

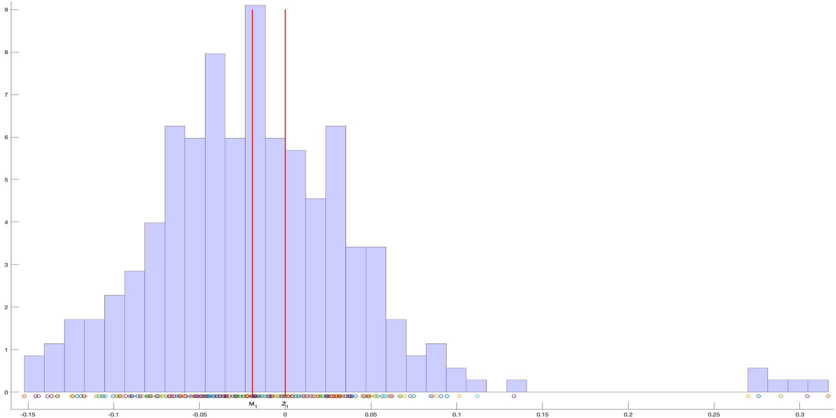


Figure: We chose a Gaussian sample with $p = 750$, $n = 1500$, $k = 450$ and $k_* = 5$. Hence, $c = 0.4$ and $\alpha = 0.3$. The histogram represents the distributions of the k values of $\log(|\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\omega}\setminus j}|) - \log(|\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\omega}}|) - 2p/n$. $M_1 = \log(\frac{1-\alpha}{1-\alpha-c}) - 2c$ and $Z_1 = 0$.

General KOO methods

Denoting

$$\breve{A}_j := \log(|\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\omega}\setminus j}|) - \log(|\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\omega}}|) \text{ and } \breve{C}_j := \mathrm{tr}(\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\omega}\setminus j}\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\omega}}^{-1}),$$

and a fixed value $\vartheta \in (0, \min_{j\in\mathbf{j}_*}\{\kappa_{\boldsymbol{\omega}\setminus j}\})$, choose the model

$$\breve{\mathbf{j}}_A = \{j \in \boldsymbol{\omega} | \breve{A}_j > \log(\frac{1-\alpha+\vartheta}{1-\alpha-c})\}, \quad \breve{\mathbf{j}}_C = \{j \in \boldsymbol{\omega} | \breve{C}_j > \frac{\vartheta+c}{1-\alpha-c} + p\}.$$

Then, we have the following theorem.

Theorem 13

Suppose that assumptions (A1) through (A4) *hold and that for any* $j \in \mathbf{j}_*$, $\kappa_{\boldsymbol{\omega}\setminus j} > 0$. *Then, for any fixed value* $\vartheta \in (0, \min_{j\in\mathbf{j}_*}\{\kappa_{\boldsymbol{\omega}\setminus j}\})$,

$$\lim_{n,p\to\infty} \breve{\mathbf{j}}_A \xrightarrow{a.s.} \mathbf{j}_* \quad \textit{and} \quad \lim_{n,p\to\infty} \breve{\mathbf{j}}_C \xrightarrow{a.s.} \mathbf{j}_*.$$

General KOO methods

Denoting

$$\breve{A}_j := \log(|\widehat{\mathbf{\Sigma}}_{\boldsymbol{\omega}\setminus j}|) - \log(|\widehat{\mathbf{\Sigma}}_{\boldsymbol{\omega}}|) \text{ and } \breve{C}_j := \mathrm{tr}(\widehat{\mathbf{\Sigma}}_{\boldsymbol{\omega}\setminus j}\widehat{\mathbf{\Sigma}}_{\boldsymbol{\omega}}^{-1}),$$

and a fixed value $\vartheta \in (0, \min_{j\in\mathbf{j}_*}\{\kappa_{\boldsymbol{\omega}\setminus j}\})$, choose the model

$$\breve{\mathbf{j}}_A = \{j \in \boldsymbol{\omega} | \breve{A}_j > \log(\frac{1-\alpha+\vartheta}{1-\alpha-c})\}, \quad \breve{\mathbf{j}}_C = \{j \in \boldsymbol{\omega} | \breve{C}_j > \frac{\vartheta + c}{1-\alpha-c} + p\}.$$

Then, we have the following theorem.

Theorem 13

Suppose that assumptions (A1) through (A4) *hold and that for any* $j \in \mathbf{j}_*$, $\kappa_{\boldsymbol{\omega}\setminus j} > 0$. *Then, for any fixed value* $\vartheta \in (0, \min_{j\in\mathbf{j}_*}\{\kappa_{\boldsymbol{\omega}\setminus j}\})$,

$$\lim_{n,p\to\infty} \breve{\mathbf{j}}_A \xrightarrow{a.s.} \mathbf{j}_* \quad \textit{and} \quad \lim_{n,p\to\infty} \breve{\mathbf{j}}_C \xrightarrow{a.s.} \mathbf{j}_*.$$

General KOO methods

Remark 3

- The condition in this theorem is much weaker than that in the AIC, BIC, and C_p and in the KOO methods based on the AIC, BIC, and C_p.
- Although $\kappa_{\boldsymbol{\omega}\setminus j}$ is not estimable for $j \in \mathbf{j}_*$, since the general KOO methods are essentially used to detect the univariate outliers, there are many well-developed methods, such as the standard deviation (SD) method, Z-score method, Tukey's method, and median absolute deviation method, that can be used to determine the value of ϑ for applications.

Outline

Proof strategy

(1) Sylvester's determinant theorem:

$$\begin{aligned}|n\widehat{\mathbf{\Sigma}}_{\mathbf{j}}| =&|\mathbf{Y}'\mathbf{Q}_{\mathbf{j}_{-1}}\mathbf{Y} - \mathbf{Y}'\mathbf{a}_1\mathbf{a}_1'\mathbf{Y}| \\ =&|n\widehat{\mathbf{\Sigma}}_{\mathbf{j}_{-1}}|(1 - \mathbf{a}_1'\mathbf{Y}(\mathbf{Y}'\mathbf{Q}_{\mathbf{j}_{-1}}\mathbf{Y})^{-1}\mathbf{Y}'\mathbf{a}_1).\end{aligned}$$

e.g. $\breve{A}_j := \log(|\widehat{\mathbf{\Sigma}}_{\boldsymbol{\omega}\backslash j}|) - \log(|\widehat{\mathbf{\Sigma}}_{\boldsymbol{\omega}}|)$ and $\breve{C}_j := \mathrm{tr}(\widehat{\mathbf{\Sigma}}_{\boldsymbol{\omega}\backslash j}\widehat{\mathbf{\Sigma}}_{\boldsymbol{\omega}}^{-1})$

(2) Stieltjes transform:

$$\hbar_n(z) := n^{-1}\mathbf{a}_t'\mathbf{Y}(n^{-1}\mathbf{Y}'\mathbf{Q}_{\mathbf{j}_{-t}}\mathbf{Y} - z\mathbf{I})^{-1}\mathbf{Y}'\mathbf{a}_t : \mathbb{C}^+ \longmapsto \mathbb{C}^+.$$

(3) Vitali's convergence theorem: For any fixed $z \in \mathbb{C}^+$, $\hbar_n(z) \stackrel{a.s.}{\to} \hbar(z)$ and then let $z \downarrow 0 + 0i$.

Outline

Simulation

Setting I: Fix $k_* = 5$, $p/n = \{0.2, 0.4, 0.6\}$ and $k/n = \{0.1, 0.2\}$ with several different values of n. Set $\mathbf{X} = (x_{ij})_{n\times k}$, $\Theta_{\mathbf{j}_*} = \sqrt{n}\mathbf{1}_5\boldsymbol{\theta}_*$ and $\Theta = (\Theta_{\mathbf{j}_*}, \mathbf{0})$, where $\{x_{ij}\}$ are i.i.d. generated from the continuous uniform distributions $U(1,5)$, $\mathbf{1}_5$ is a five-dimensional vector of ones and $\boldsymbol{\theta}_* = ((-0.5)^0, \dots, (-0.5)^{p-1})$.

Setting II: This setting is the same as Setting I, except $\Theta_{\mathbf{j}_*} = n\mathbf{1}_5\boldsymbol{\theta}_*$.

Here, we use the 2 SD method to choose the critical points in the general KOO methods:

$$\breve{\mathbf{j}}_A = \{j \in \boldsymbol{\omega} | \breve{A}_j > \log(\frac{1-\alpha}{1-\alpha-c}) + 2sd_A\}$$

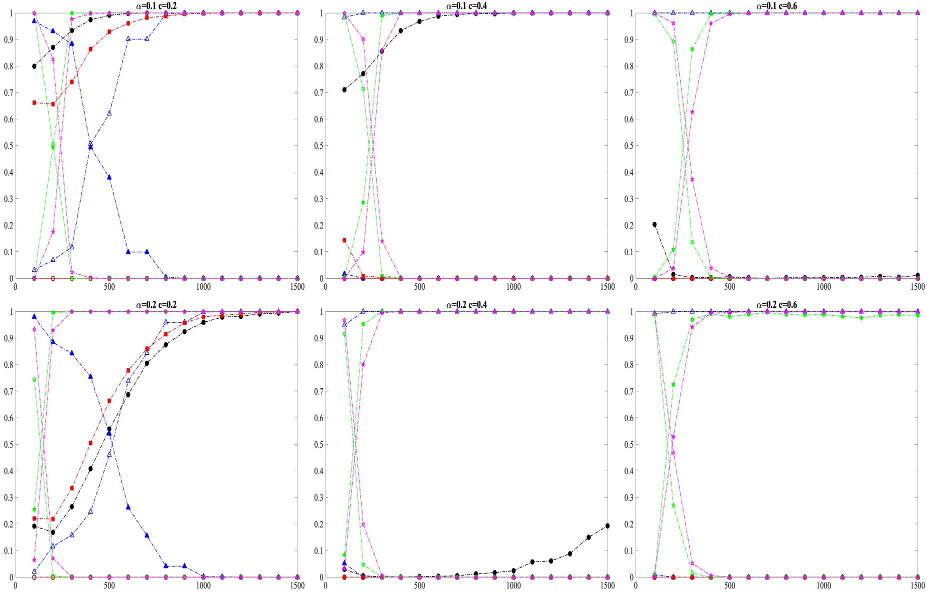
and

$$\breve{\mathbf{j}}_C = \{j \in \boldsymbol{\omega} | \breve{C}_j > \frac{c}{1-\alpha-c} + p + 2sd_C\},$$

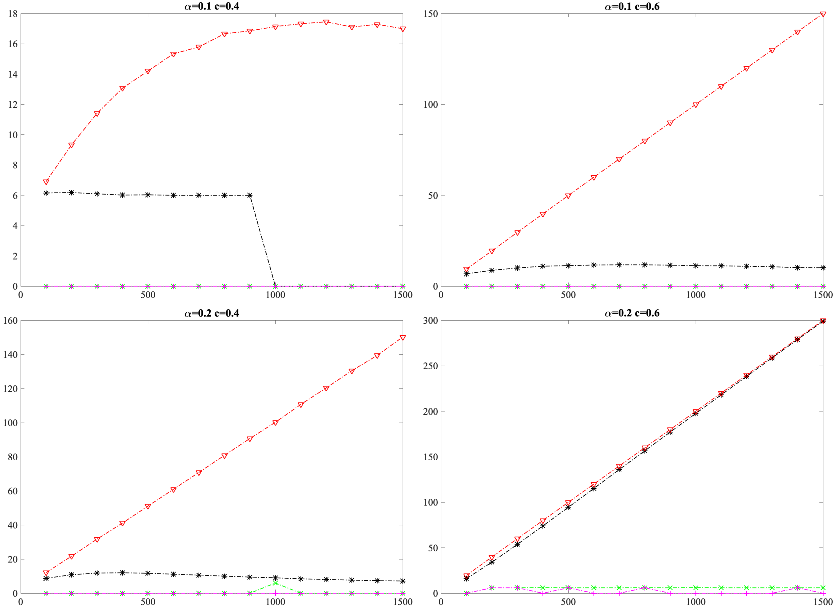
where sd_A and sd_C are the sample standard deviations of $\{\breve{A}_j\}$ and $\{\breve{C}_j\}$, respectively.

	$c=.2$				$c=.4$				$c=.6$			
	V_1	V_2	V_3	V_4	V_1	V_2	V_3	V_4	V_1	V_2	V_3	V_4
$\alpha=.1$	.15	.50	.87	1.49	.21	.10	.81	1.56	.10	-.30	.92	1.80
$\alpha=.2$	.11	.40	.91	1.32	.11	0	.92	1.43	-.19	-.40	1.21	1.72

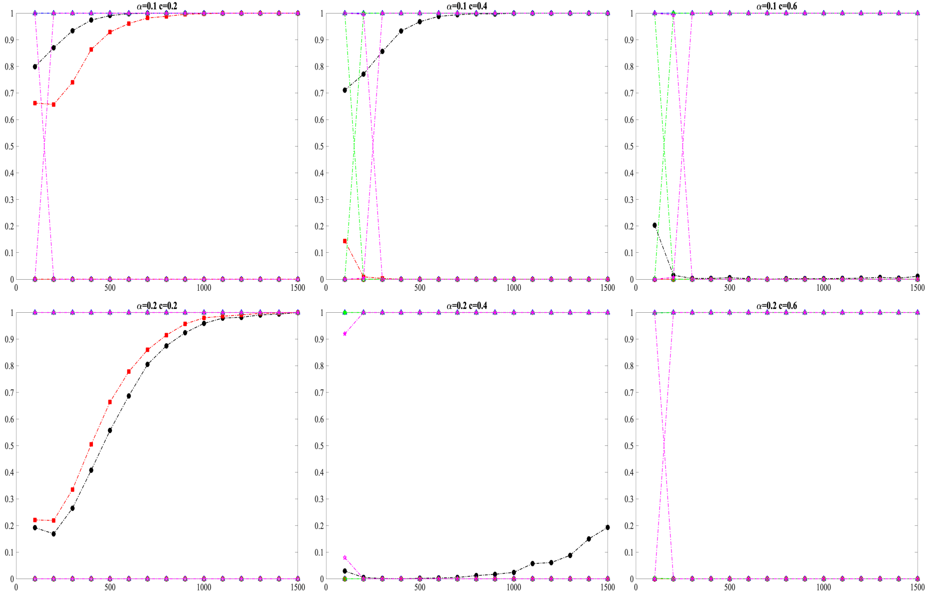
Table: Values of $V_1 := 2c - \log(\frac{1-\alpha}{1-\alpha-c})$, $V_2 := 2(1-\alpha-c) - (1-\alpha)$, $V_3 := \log(\tau_{\boldsymbol{\omega}\setminus\{1\}}) - \log(1-\alpha-c) - 2c$, and $V_4 := \mathrm{tr}(\Phi_{\boldsymbol{\omega}\setminus j}) - \frac{c(1-\alpha-2c)}{1-\alpha}$.


α=0.1 c=0.2
α=0.1 c=0.4
α=0.1 c=0.6
α=0.2 c=0.2
α=0.2 c=0.4
α=0.2 c=0.6
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
500
1000
1500

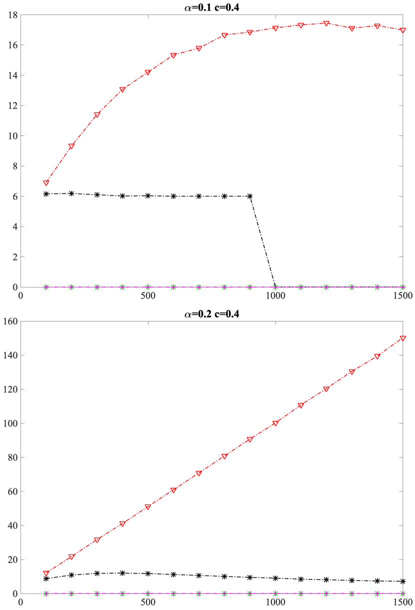
(a) Setting I



(b) Setting I



(c) Setting II


α=0.1 c=0.4
α=0.1 c=0.6
α=0.2 c=0.4
α=0.2 c=0.6
0
500
1000
1500

(d) Setting II

Conclusion

- We show the necessary and sufficient conditions for the strong consistency of variable selection methods based on the AIC, BIC, and C_p in high-dimensional-response regression;
- We examine the strongly consistent properties of the knock-one-out methods based on the AIC, BIC, and C_p;
- On the basis of the KOO methods, we propose two general KOO methods that not only remove the penalty terms but also reduce the conditions for the dimensions and sizes of the predictors.
- Random matrix theory is introduced to high-dimensional high-dimensional-response regression model.

Thank you!