Asymptotic Theory for Model Selection

Jun Shao

Department of Statistics
University of Wisconsin
Madison, WI 53706, USA

Reference: Shao (1997, Statistica Sinica, pp. 221-264)

Introduction

Responses and covariates

$\mathbf{y}_n = (y_1, ..., y_n)$: independent responses
$\mathbf{X}_n = (\mathbf{x}_1', ..., \mathbf{x}_n')'$: an $n \times p_n$ matrix whose ith row $\mathbf{x}_i$ is the value of a p_n-dimensional covariate associated with y_i

We are interested in the relationship between $\mathbf{y}_n$ and $\mathbf{X}_n$ through

$$\mu_n = E(\mathbf{y}_n | \mathbf{X}_n)$$

We may be interested in inference on μ_n

Model/Variable selection

A class of models, indexed by $\alpha \in \mathscr{A}_n$, is proposed for $E(\mathbf{y}_n | \mathbf{X}_n)$

If $\mathscr{A}_n$ contains more than one model, then we need to select a model from $\mathscr{A}_n$ using the observed $\mathbf{y}_n$ and $\mathbf{X}_n$

If each α corresponds to an $n \times p_n(\alpha)$ sub-matrix of $\mathbf{X}_n$, then model selection is also called variable selection

Introduction

Responses and covariates

$\mathbf{y}_n = (y_1, ..., y_n)$: independent responses
$\mathbf{X}_n = (\mathbf{x}_1', ..., \mathbf{x}_n')'$: an $n \times p_n$ matrix whose ith row $\mathbf{x}_i$ is the value of a p_n-dimensional covariate associated with y_i

We are interested in the relationship between $\mathbf{y}_n$ and $\mathbf{X}_n$ through

$$\mu_n = E(\mathbf{y}_n | \mathbf{X}_n)$$

We may be interested in inference on μ_n

Model/Variable selection

A class of models, indexed by $\alpha \in \mathscr{A}_n$, is proposed for $E(\mathbf{y}_n | \mathbf{X}_n)$

If $\mathscr{A}_n$ contains more than one model, then we need to select a model from $\mathscr{A}_n$ using the observed $\mathbf{y}_n$ and $\mathbf{X}_n$

If each α corresponds to an $n \times p_n(\alpha)$ sub-matrix of $\mathbf{X}_n$, then model selection is also called variable selection

Example 1. Linear regression

- $p_n = p$ for all n
- $\boldsymbol{\mu}_n = \mathbf{X}_n\boldsymbol{\beta}$
- $\boldsymbol{\beta} = (\boldsymbol{\beta}_1', \boldsymbol{\beta}_2')'$, $\mathbf{X}_n = (\mathbf{X}_{n1}, \mathbf{X}_{n2})$
  - It is suspected that $\boldsymbol{\beta}_2 = 0$ ($\mathbf{X}_{n2}$ is unrelated to $\mathbf{y}_n$)
  - Model 1: $\boldsymbol{\mu}_n = \mathbf{X}_{n1}\boldsymbol{\beta}_1$
  - Model 2: $\boldsymbol{\mu}_n = \mathbf{X}_n\boldsymbol{\beta}$
  - $\mathscr{A}_n = \{1,2\}$
  - Model 1 is better if $\boldsymbol{\beta}_2 = 0$
- In general, $\mathscr{A}_n =$ all subsets of $\{1,...,p\}$
  - Model α: $\boldsymbol{\mu}_n = \mathbf{X}_n(\alpha)\boldsymbol{\beta}(\alpha)$
  - $\boldsymbol{\beta}(\alpha)$: sub-vector of $\boldsymbol{\beta}$ with indices in α
  - $\mathbf{X}_n(\alpha)$: the corresponding sub-matrix of $\mathbf{X}_n$
  - The number of models in $\mathscr{A}_n$ is 2^p
- Approximation to a response surface
  - The ith row of $\mathbf{X}_n(\alpha_h) = (1, t_i, t_i^2, ..., t_i^h)$, $t_i \in \mathscr{R}$
  - $\alpha_h = \{1,...,h\}$: a polynomial of order h
  - $\mathscr{A}_n = \{\alpha_h : h = 0, 1, ..., p_n\}$

Example 2. 1-mean vs p-mean

- $n = pr$, $p = p_n$, $r = r_n$
- There are p groups, each has r identically distributed observations
- Select one model from two models
  - 1-mean model: all groups have the same mean μ_1
  - p-mean model: p groups have different means $\mu_1, ..., \mu_p$
- $\mathscr{A}_n = \{\alpha_1, \alpha_p\}$

$$\mathbf{X}_n = \begin{pmatrix} \mathbf{1}_r & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{1}_r & \mathbf{1}_r & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{1}_r & \mathbf{0} & \mathbf{1}_r & \cdots & \mathbf{0} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \mathbf{1}_r & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{1}_r \end{pmatrix} \qquad \beta = \begin{pmatrix} \mu_1 \\ \mu_2 - \mu_1 \\ \mu_3 - \mu_1 \\ \cdots \\ \mu_p - \mu_1 \end{pmatrix}$$

$$\mathbf{X}_n(\alpha_p) = \mathbf{X}_n \qquad \beta(\alpha_p) = \beta$$

$$\mathbf{X}_n(\alpha_1) = \mathbf{1}_n \qquad \beta(\alpha_1) = \mu_1$$

Criterion for Model Selection

- $\boldsymbol{\mu}_n$ is estimated by $\widehat{\boldsymbol{\mu}}_n(\alpha)$ under model α
- Minimize the squared error loss

$$L_n(\alpha) = \frac{\|\boldsymbol{\mu}_n - \widehat{\boldsymbol{\mu}}_n(\alpha)\|^2}{n} \qquad \text{over } \alpha \in \mathscr{A}_n$$

  Equivalent to minimizing the average prediction error

$$\frac{E\left[\|\mathbf{z}_n - \widehat{\boldsymbol{\mu}}_n(\alpha)\|^2 \mid \mathbf{y}_n\right]}{n} \qquad \text{over } \alpha \in \mathscr{A}_n$$

  $\mathbf{z}_n$: a future independent copy of $\mathbf{y}_n$
- Optimal model α_n^L:

$$L_n(\alpha_n^L) = \min_{\alpha \in \mathscr{A}_n} L_n(\alpha)$$

  α_n^L may be random

Assessment of Model Selection Procedures

- $\widehat{\alpha}_n$: a model selected based on a model selection procedure
- The selection procedure is consistent if

$$\lim_{n\to\infty} P\{\widehat{\alpha}_n = \alpha_n^L\} = 1$$

which implies

$$\lim_{n\to\infty} P\{L_n(\widehat{\alpha}_n) = L_n(\alpha_n^L)\} = 1$$

$\widehat{\mu}_n(\alpha_n)$ is asymptotically efficient, i.e., it is asymptotically as efficient as $\widehat{\mu}_n(\alpha_n^L)$
The two results are the same if $L_n(\alpha)$ has a unique minimum for all large n

- The selection procedure is asymptotically loss efficient if

$$L_n(\widehat{\alpha}_n)/L_n(\alpha_n^L) \to_P 1$$

which is weaker than consistency

Model Selection Procedures

Methods for fixed p or $p_n/n \to 0$

- Information criterion
  - AIC (Akaike, 1970), C_p (Mallows, 1973), BIC (Schwarz, 1978)
  - FPE_λ (Shibata, 1984)
  - GIC (Nishii, 1984, Rao and Wu, 1989, Potscher, 1989)
- Cross-Validation (CV)
  - Delete-1 CV (Allen, 1974, Stone, 1974)
  - GCV (Craven and Wahba, 1979)
  - Delete-d CV (Geisser, 1975, Burman, 1986, Shao, 1993)
- Bootstrap (Efron, 1983, Shao, 1996)
- Methods for Time Series
  - PMDL and PLS (Rissanen, 1986, Wei, 1992)
- LASSO (Tibshirani, 1996)
- Methods after 1997?
- Thresholding
- Methods for $p_n/n \not\to 0$?

Asymptotic Theory for GIC

The GIC in linear models

Consider linear models

$$\mu_n = \mathbf{X}_n(\alpha)\beta(\alpha) \qquad \alpha \in \mathscr{A}_n$$

- $\mathbf{X}_n$ is of full rank ($p_n < n$)
- $\mathbf{e}_n = \mathbf{y}_n - \mu_n$ has iid components, $V(\mathbf{e}_n|\mathbf{X}_n) = \sigma^2 \mathbf{I}_n$
- Under model α, $\beta(\alpha)$ is estimated by the LSE
- $\widehat{\mu}_n(\alpha) = \mathbf{H}_n(\alpha)\mathbf{y}_n$, $\mathbf{H}_n(\alpha) = \mathbf{X}_n(\alpha)[\mathbf{X}_n(\alpha)'\mathbf{X}_n(\alpha)]^{-1}\mathbf{X}_n(\alpha)$
- Correct models

  $$\mathscr{A}_n^c = \{\alpha \in \mathscr{A}_n : \mu_n = \mathbf{X}_n(\alpha)\beta(\alpha) \text{ is true } \}$$

  Wrong models
  $$\mathscr{A}_n^w = \{\alpha \in \mathscr{A}_n : \alpha \notin \mathscr{A}_n^c\}$$
- The loss is equal to

$$L_n(\alpha) = \Delta_n(\alpha) + \mathbf{e}_n'\mathbf{H}_n(\alpha)\mathbf{e}_n/n$$

$$\Delta_n(\alpha) = \|\mu_n - \mathbf{H}_n(\alpha)\mu_n\|^2/n \ \ (= 0 \text{ if } \alpha \in \mathscr{A}_n^c)$$

The GIC

A model $\widehat{\alpha}_n \in \mathscr{A}_n$ is selected by minimizing

$$\Gamma_{n,\lambda_n}(\alpha) = \frac{S_n(\alpha)}{n} + \frac{\lambda_n \widehat{\sigma}_n^2 p_n(\alpha)}{n} \quad \text{over } \alpha \in \mathscr{A}_n$$

$S_n(\alpha) = \|\mathbf{y}_n - \widehat{\mu}_n(\alpha)\|^2$ (measuring goodness-of-fit)
$p_n(\alpha)$: dimension of α
λ_n: non-random positive penalty
$\widehat{\sigma}_n^2$: an estimator of σ^2, e.g., $\widehat{\sigma}_n^2 = \|\mathbf{y}_n - \widehat{\mu}_n\|^2/(n - p_n)$

- If $\lambda_n = 2$, this is the C_p method
- If $\lambda_n = \lambda$, a constant larger than 2, this is the FPE$_\lambda$ method
- If $\lambda_n = \log n$, this is almost the BIC
- In general, λ_n can be any sequence with $\lambda_n \to \infty$
- If $\lambda_n = 2$, the GIC is asymptotically equivalent to the delete-1 CV and GCV
- If $\lambda_n = n/(n-d)$, then the GIC is asymptotically equivalent to the delete-d CV.

Is the GIC asymptotically loss efficient or consistent?

$$\frac{S_n(\alpha)}{n} = \frac{\|\mathbf{y}_n - \mathbf{H}_n(\alpha)\mathbf{y}_n\|^2}{n} = \frac{\|\boldsymbol{\mu}_n - \mathbf{H}_n(\alpha)\boldsymbol{\mu}_n + \mathbf{e}_n - \mathbf{H}_n(\alpha)\mathbf{e}_n\|^2}{n}$$

$$= \Delta_n(\alpha) + \frac{\|\mathbf{e}_n\|^2}{n} - \frac{\mathbf{e}_n'\mathbf{H}_n(\alpha)\mathbf{e}_n}{n} + \frac{2\mathbf{e}_n'[\mathbf{I}_n - \mathbf{H}_n(\alpha)]\boldsymbol{\mu}_n}{n}$$

$\alpha \in \mathscr{A}_n^c$

$[\mathbf{I}_n - \mathbf{H}_n(\alpha)]\boldsymbol{\mu}_n = \mathbf{X}_n(\alpha)\beta(\alpha) - \mathbf{X}_n(\alpha)\beta(\alpha) = 0$

$\Delta_n(\alpha) = 0$

$L_n(\alpha) = \Delta_n(\alpha) + \mathbf{e}_n'\mathbf{H}_n(\alpha)\mathbf{e}_n/n = \mathbf{e}_n'\mathbf{H}_n(\alpha)\mathbf{e}_n/n$

$$\Gamma_{n,\lambda_n}(\alpha) = \frac{S_n(\alpha)}{n} + \frac{\lambda_n\hat{\sigma}_n^2 p_n(\alpha)}{n} = \frac{\|\mathbf{e}_n\|^2}{n} - \frac{\mathbf{e}_n'\mathbf{H}_n(\alpha)\mathbf{e}_n}{n} + \frac{\lambda_n\hat{\sigma}_n^2 p_n(\alpha)}{n}$$

$$= \frac{\|\mathbf{e}_n\|^2}{n} + L_n(\alpha) + \frac{\lambda_n\hat{\sigma}_n^2 p_n(\alpha)}{n} - \frac{2\mathbf{e}_n'\mathbf{H}_n(\alpha)\mathbf{e}_n}{n}$$

Is the GIC asymptotically loss efficient or consistent?

$$\frac{S_n(\alpha)}{n} = \frac{\|\mathbf{y}_n - \mathbf{H}_n(\alpha)\mathbf{y}_n\|^2}{n} = \frac{\|\boldsymbol{\mu}_n - \mathbf{H}_n(\alpha)\boldsymbol{\mu}_n + \mathbf{e}_n - \mathbf{H}_n(\alpha)\mathbf{e}_n\|^2}{n}$$
$$= \Delta_n(\alpha) + \frac{\|\mathbf{e}_n\|^2}{n} - \frac{\mathbf{e}_n'\mathbf{H}_n(\alpha)\mathbf{e}_n}{n} + \frac{2\mathbf{e}_n'[\mathbf{I}_n - \mathbf{H}_n(\alpha)]\boldsymbol{\mu}_n}{n}$$

$\alpha \in \mathscr{A}_n^c$

$[\mathbf{I}_n - \mathbf{H}_n(\alpha)]\boldsymbol{\mu}_n = \mathbf{X}_n(\alpha)\beta(\alpha) - \mathbf{X}_n(\alpha)\beta(\alpha) = 0$

$\Delta_n(\alpha) = 0$

$L_n(\alpha) = \Delta_n(\alpha) + \mathbf{e}_n'\mathbf{H}_n(\alpha)\mathbf{e}_n/n = \mathbf{e}_n'\mathbf{H}_n(\alpha)\mathbf{e}_n/n$

$$\Gamma_{n,\lambda_n}(\alpha) = \frac{S_n(\alpha)}{n} + \frac{\lambda_n\widehat{\sigma}_n^2 p_n(\alpha)}{n} = \frac{\|\mathbf{e}_n\|^2}{n} - \frac{\mathbf{e}_n'\mathbf{H}_n(\alpha)\mathbf{e}_n}{n} + \frac{\lambda_n\widehat{\sigma}_n^2 p_n(\alpha)}{n}$$
$$= \frac{\|\mathbf{e}_n\|^2}{n} + L_n(\alpha) + \frac{\lambda_n\widehat{\sigma}_n^2 p_n(\alpha)}{n} - \frac{2\mathbf{e}_n'\mathbf{H}_n(\alpha)\mathbf{e}_n}{n}$$

When $\mathscr{A}_n = \mathscr{A}_n^c$

- $\alpha_n^L = \alpha \in \mathscr{A}_n^c$ with the smallest $p_n(\alpha)$

$$\Gamma_{n,\lambda_n}(\alpha) = \frac{\|\mathbf{e}_n\|^2}{n} + L_n(\alpha) + \frac{\lambda_n \widehat{\sigma}_n^2 p_n(\alpha)}{n} - \frac{2\mathbf{e}_n' \mathbf{H}_n(\alpha)\mathbf{e}_n}{n}$$

- If $\lambda_n = 2$ (the C_p method, AIC, delete-1 CV, or GCV), the term

$$\frac{2\widehat{\sigma}_n^2 p_n(\alpha)}{n} - \frac{2\mathbf{e}_n' \mathbf{H}_n(\alpha)\mathbf{e}_n}{n}$$

  is of the same order as $L_n(\alpha) = \mathbf{e}_n' \mathbf{H}_n(\alpha)\mathbf{e}_n / n$ unless $p_n(\alpha) \to \infty$ for all but one model in $\mathscr{A}_n^c$
- Under some conditions, the GIC with $\lambda_n = 2$ is asymptotically loss efficient if and only if $\mathscr{A}_n^c$ does not contain two models with fixed dimensions
- If $\lambda_n \to \infty$, the dominating term in $\Gamma_{n,\lambda_n}(\alpha)$ is $\lambda_n \widehat{\sigma}_n^2 p_n(\alpha)/n$
  The GIC selects a model by minimizing $p_n(\alpha)$
  Hence, the GIC is consistent
- The case of $\lambda_n = \lambda$ is similar to the case of $\lambda_n = 2$

When $\mathscr{A}_n = \mathscr{A}_n^w$

$$\Gamma_{n,\lambda_n}(\alpha) = \frac{\|\mathbf{e}_n\|^2}{n} + \Delta_n(\alpha) - \frac{\mathbf{e}_n'\mathbf{H}_n(\alpha)\mathbf{e}_n}{n} + \frac{\lambda_n \widehat{\sigma}_n^2 p_n(\alpha)}{n} + O_P\left(\frac{\Delta_n(\alpha)}{n}\right)$$

$$= \frac{\|\mathbf{e}_n\|^2}{n} + L_n(\alpha) + O_P\left(\frac{\lambda_n p_n(\alpha)}{n}\right) + O_P\left(\frac{L_n(\alpha)}{n}\right)$$

Assume that

$$\liminf_{n\to\infty} \min_{\alpha\in\mathscr{A}_n^w} \Delta_n(\alpha) > 0 \quad \text{and} \quad \frac{\lambda_n p_n}{n} \to 0$$

(The first condition impies that a wrong model is always worse than a correct model)
Then

$$\Gamma_{n,\lambda_n}(\alpha) = \frac{\|\mathbf{e}_n\|^2}{n} + L_n(\alpha) + o_P(L_n(\alpha))$$

Minimizing $\Gamma_{n,\lambda_n}(\alpha)$ is asymptotically the same as minimizing $L_n(\alpha)$

Hence, the GIC is asymptotically loss efficient
The GIC can select the best model in $\mathscr{A}_n^w$

Conclusions (under the given conditions)

According to their asymptotic behavior, the model selection methods can be classfied into three classes

(1) The GIC with $\lambda_n = 2$, C_p, AIC, delete-1 CV, GCV

(2) The GIC with $\lambda_n \to \infty$, delete-d CV with $d/n \to 1$, BIC, PMDL, PLS
$\lambda_n p_n / n \to 0$

(3) The GIC with $\lambda_n = \lambda$, delete-d CV with $d/n \to \tau \in (0,1)$

Key properties

- Methods in class (1) are useful when there is no fixed-dimension correct model
- Methods in class (2) are useful whene there are fixed-dimension correct models
- Methods in class (3) are compromises and are not recommended

Conclusions (under the given conditions)

According to their asymptotic behavior, the model selection methods can be classfied into three classes

(1) The GIC with $\lambda_n = 2$, C_p, AIC, delete-1 CV, GCV

(2) The GIC with $\lambda_n \to \infty$, delete-d CV with $d/n \to 1$, BIC, PMDL, PLS $\lambda_n p_n/n \to 0$

(3) The GIC with $\lambda_n = \lambda$, delete-d CV with $d/n \to \tau \in (0,1)$

Key properties

- Methods in class (1) are useful when there is no fixed-dimension correct model
- Methods in class (2) are useful whene there are fixed-dimension correct models
- Methods in class (3) are compromises and are not recommended

Example 2. 1-mean vs p-mean

$\mathscr{A}_n = \{\alpha_1, \alpha_p\}$
p_n groups, each with r_n observations
$\Delta_n(\alpha_p) = \sum_{j=1}^{p}(\mu_j - \overline{\mu})^2/p$, $\overline{\mu} = \sum_{j=1}^{p}\mu_j/p$
$n = p_n r_n \to \infty$ means that either $p_n \to \infty$ or $r_n \to \infty$

1. $p_n = p$ is fixed and $r_n \to \infty$

- The dimensions of correct models are fixed
- The GIC with $\lambda_n \to \infty$ and $\lambda_n/n \to 0$ is consistent
- The GIC with $\lambda_n = 2$ is inconsistent

2. $p_n \to \infty$ and $r_n = r$ is fixed

- Only one correct model has a fixed dimension
- The GIC with $\lambda_n = 2$ is consistent
- The GIC with $\lambda_n \to \infty$ is inconsistent, because $\lambda_n p_n/n = \lambda_n/r \to \infty$

3. $p_n \to \infty$ and $r_n \to \infty$

- Only one correct model has a fixed dimension
- The GIC is consistent, provided that $\lambda_n/r_n \to 0$

Example 2. 1-mean vs p-mean

$\mathscr{A}_n = \{\alpha_1, \alpha_p\}$

p_n groups, each with r_n observations

$\Delta_n(\alpha_p) = \sum_{j=1}^{p} (\mu_j - \overline{\mu})^2 / p$, $\overline{\mu} = \sum_{j=1}^{p} \mu_j / p$

$n = p_n r_n \to \infty$ means that either $p_n \to \infty$ or $r_n \to \infty$

1. $p_n = p$ is fixed and $r_n \to \infty$

- The dimensions of correct models are fixed
- The GIC with $\lambda_n \to \infty$ and $\lambda_n / n \to 0$ is consistent
- The GIC with $\lambda_n = 2$ is inconsistent

2. $p_n \to \infty$ and $r_n = r$ is fixed

- Only one correct model has a fixed dimension
- The GIC with $\lambda_n = 2$ is consistent
- The GIC with $\lambda_n \to \infty$ is inconsistent, because $\lambda_n p_n / n = \lambda_n / r \to \infty$

3. $p_n \to \infty$ and $r_n \to \infty$

- Only one correct model has a fixed dimension
- The GIC is consistent, provided that $\lambda_n / r_n \to 0$

Example 2. 1-mean vs p-mean

$\mathscr{A}_n = \{\alpha_1, \alpha_p\}$

p_n groups, each with r_n observations

$\Delta_n(\alpha_p) = \sum_{j=1}^{p}(\mu_j - \bar{\mu})^2/p$, $\bar{\mu} = \sum_{j=1}^{p}\mu_j/p$

$n = p_n r_n \to \infty$ means that either $p_n \to \infty$ or $r_n \to \infty$

1. $p_n = p$ is fixed and $r_n \to \infty$

- The dimensions of correct models are fixed
- The GIC with $\lambda_n \to \infty$ and $\lambda_n/n \to 0$ is consistent
- The GIC with $\lambda_n = 2$ is inconsistent

2. $p_n \to \infty$ and $r_n = r$ is fixed

- Only one correct model has a fixed dimension
- The GIC with $\lambda_n = 2$ is consistent
- The GIC with $\lambda_n \to \infty$ is inconsistent, because $\lambda_n p_n/n = \lambda_n/r \to \infty$

3. $p_n \to \infty$ and $r_n \to \infty$

- Only one correct model has a fixed dimension
- The GIC is consistent, provided that $\lambda_n/r_n \to 0$

Example 2. 1-mean vs p-mean

$\mathscr{A}_n = \{\alpha_1, \alpha_p\}$

p_n groups, each with r_n observations

$\Delta_n(\alpha_p) = \sum_{j=1}^{p}(\mu_j - \bar{\mu})^2/p$, $\bar{\mu} = \sum_{j=1}^{p}\mu_j/p$

$n = p_n r_n \to \infty$ means that either $p_n \to \infty$ or $r_n \to \infty$

1. $p_n = p$ is fixed and $r_n \to \infty$

- The dimensions of correct models are fixed
- The GIC with $\lambda_n \to \infty$ and $\lambda_n/n \to 0$ is consistent
- The GIC with $\lambda_n = 2$ is inconsistent

2. $p_n \to \infty$ and $r_n = r$ is fixed

- Only one correct model has a fixed dimension
- The GIC with $\lambda_n = 2$ is consistent
- The GIC with $\lambda_n \to \infty$ is inconsistent, because $\lambda_n p_n/n = \lambda_n/r \to \infty$

3. $p_n \to \infty$ and $r_n \to \infty$

- Only one correct model has a fixed dimension
- The GIC is consistent, provided that $\lambda_n/r_n \to 0$

Variable Selection by Thresholding

Assumption A

- $\mathbf{y}_n$ is normally distributed
- $\min_{j:\beta_j \neq 0} |\beta_j| >$a positive constant, $\beta = (\beta_1, ..., \beta_p)$
- $\mathbf{X}_n'\mathbf{X}_n$ is of rank p ($p < n$)
- λ_{in} = the ith eigenvalue of $\mathbf{X}_n'\mathbf{X}_n$, $i = 1, ..., n$
  $\lambda_{in} = b_i \zeta_n$, $0 < b_i \leq b < \infty$, $0 < \zeta_n \to \infty$
- $p_n \to \infty$ but $(\log p_n)/\zeta_n \to 0$

Thresholding

- $\widehat{\beta} = (\mathbf{X}_n'\mathbf{X}_n)^{-1}\mathbf{X}_n'\mathbf{y}_n = (\widehat{\beta}_1, ..., \widehat{\beta}_p)$ (the LSE)
  $\widehat{\beta} \sim N(\beta, \sigma^2(\mathbf{X}_n'\mathbf{X}_n)^{-1})$
- $a_n = [(\log p_n)/\zeta_n]^\alpha$, $\alpha \in (0, 1/2)$, $a_n \to 0$
- Variable $\mathbf{x}_i$ is selected if and only if $|\widehat{\beta}_i| > a_n$

Variable Selection by Thresholding

Assumption A

- $\mathbf{y}_n$ is normally distributed
- $\min_{j:\beta_j \neq 0} |\beta_j| >$a positive constant, $\beta = (\beta_1, ..., \beta_p)$
- $\mathbf{X}_n'\mathbf{X}_n$ is of rank p ($p < n$)
- λ_{in} = the ith eigenvalue of $\mathbf{X}_n'\mathbf{X}_n$, $i = 1, ..., n$
  $\lambda_{in} = b_i \zeta_n$, $0 < b_i \leq b < \infty$, $0 < \zeta_n \to \infty$
- $p_n \to \infty$ but $(\log p_n)/\zeta_n \to 0$

Thresholding

- $\widehat{\beta} = (\mathbf{X}_n'\mathbf{X}_n)^{-1}\mathbf{X}_n'\mathbf{y}_n = (\widehat{\beta}_1, ..., \widehat{\beta}_p)$ (the LSE)
  $\widehat{\beta} \sim N(\beta, \sigma^2(\mathbf{X}_n'\mathbf{X}_n)^{-1})$
- $a_n = [(\log p_n)/\zeta_n]^\alpha$, $\alpha \in (0, 1/2)$, $a_n \to 0$
- Variable $\mathbf{x}_i$ is selected if and only if $|\widehat{\beta}_i| > a_n$

Asymptotic properties

1. $\lim_{n\to\infty} P(|\widehat{\beta}_i| \le a_n \text{ for all } i \text{ with } \beta_i = 0) = 1$
2. $\lim_{n\to\infty} P(|\widehat{\beta}_i| > a_n \text{ for all } i \text{ with } \beta_i \neq 0) = 1$

Proof

$$\begin{aligned}
1 - P(|\widehat{\beta}_i| \le a_n \text{ for all } i \text{ with } \beta_i = 0) &= P\left(\cup_{i:\beta_i=0}\{|\widehat{\beta}_i - \beta_i| > a_n\}\right) \\
&\le \sum_{i:\beta_i=0} P\left(\{|\widehat{\beta}_i - \beta_i| > a_n\}\right) \\
&= 2\sum_{i:\beta_i=0} \Phi\left(-\frac{a_n}{\tau_i}\right) \\
&\le \sum_{i:\beta_i=0} e^{-a_n^2/(2\tau_i^2)}
\end{aligned}$$

$\tau_i^2 = \text{var}(\widehat{\beta}_i)$

Asymptotic properties

1. $\lim_{n\to\infty} P(|\widehat{\beta}_i| \le a_n \text{ for all } i \text{ with } \beta_i = 0) = 1$

2. $\lim_{n\to\infty} P(|\widehat{\beta}_i| > a_n \text{ for all } i \text{ with } \beta_i \neq 0) = 1$

Proof

$$\begin{aligned} 1 - P(|\widehat{\beta}_i| \le a_n \text{ for all } i \text{ with } \beta_i = 0) &= P\left(\cup_{i:\beta_i=0}\{|\widehat{\beta}_i - \beta_i| > a_n\}\right) \\ &\le \sum_{i:\beta_i=0} P\left(\{|\widehat{\beta}_i - \beta_i| > a_n\}\right) \\ &= 2\sum_{i:\beta_i=0} \Phi\left(-\frac{a_n}{\tau_i}\right) \\ &\le \sum_{i:\beta_i=0} e^{-a_n^2/(2\tau_i^2)} \end{aligned}$$

$\tau_i^2 = \text{var}(\widehat{\beta}_i)$

$\tau_i \le c\zeta_n^{-1}$ for a constant c

$$\frac{a_n^2}{2\tau_i^2} \ge \frac{a_n^2 \zeta_n}{2c} = \frac{1}{2c}\left(\frac{\log p_n}{\zeta_n}\right)^{2\alpha-1} \log p_n \ge M \log p_n$$

for any $M > 0$, since $(\log p_n)/\zeta_n \to 0$ and $\alpha < 1/2$

Then

$$\begin{aligned} 1 - P(|\widehat{\beta}_i| \le a_n \text{ for all } i \text{ with } \beta_i = 0) &\le \sum_{i:\beta_i=0} e^{-M\log p} \\ &\le p e^{-M\log p} \\ &= p^{1-M} \\ &\to 0 \end{aligned}$$

This proves property 1

The proof for property 2 is similar

Topics of Covered in 992

- LASSO and its asymptotic properties
- Nonconcave penalized likelihood method
- Sure independence screening
- High dimensional variable selection by Wasserman and Roeder
- Bayesian model/variable selection
- A review by Fan and Lv
- Others