Asymptotic oracle properties of SCAD-penalized least squares estimators

San-Teng Huang

National Dong Hwa University

2018/09/25

Outline

Introduction

Model

- Consider a linear model:

$$y_i = \beta_0 + \tilde{\mathbf{x}}_i \boldsymbol{\beta} + \varepsilon_i \qquad , i = 1, 2, ..., n \tag{1}$$

where $\boldsymbol{\beta}$ is $p_n \times 1$ vector,
$\tilde{\mathbf{x}}_i$ is $1 \times p_n$ row vector of $\mathbf{X}$,
ε_i *i.i.d* with mean 0, variance σ^2, for $i = 1, ..., n$.
- For simplicity, assume $\beta_0 = 0$.

Of interest

- Assume $p_n \to \infty$ as $n \to \infty$.
  In biomedical studies investiong the relationship between phenotype and genomic.
- Assume only some parameters are non-zero.
  the true $\boldsymbol{\beta_0} = (\beta_{01}, \beta_{02}, \dots, \beta_{0p_n}) = (\boldsymbol{\beta_{01}^T}, \boldsymbol{\beta_{02}^T})$
  with $\boldsymbol{\beta_{01}^T} = (\beta_{01}, \dots, \beta_{0k_n})$, $\boldsymbol{\beta_{02}^T} = (0_{01}, \dots, 0_{0m_n})$,
  $p_n = k_n + m_n$.
- estimate $\boldsymbol{\beta_0}$ where $\boldsymbol{\beta_0}$ is sparse.

Variable selection

- Classical variable selection method: best subset selection, forward/backward-stepwise selection.

  Disadvantage:
  (i) for high-dimensional data, computation is not feasible. (p is large $\Rightarrow$ lots of combinations)
  (ii) discrete process: take some subsets of full model and estimate coefficients. the variables are either removed or retained.

Variable selection

- Penalized method: SCAD, LASSO, ridge regression.
  (i) It's more computationally feasible for high-dimensional data.
  (ii) continuous process: fit the model with all variables but regularizing the estimated coefficients. achieve variable selection and estimations simultaneously.

SCAD-penalized Least Squares(LS-SCAD)

- To minimize

$$Q_n(\boldsymbol{\beta};\lambda_n)=\sum_{i=1}^{n}(y_i-\tilde{\mathbf{x}}_i\boldsymbol{\beta})^2+n\sum_{j=1}^{p_n}p_{\lambda_n}(|\beta_j|) \tag{2}$$

where $p_{\lambda_n}(|\beta_j|)$ is SCAD penalty function.

- The LS-SCAD estimator is

$$\hat{\boldsymbol{\beta}}_n=\arg\min_{\boldsymbol{\beta}} Q_n(\boldsymbol{\beta};\lambda_n)$$

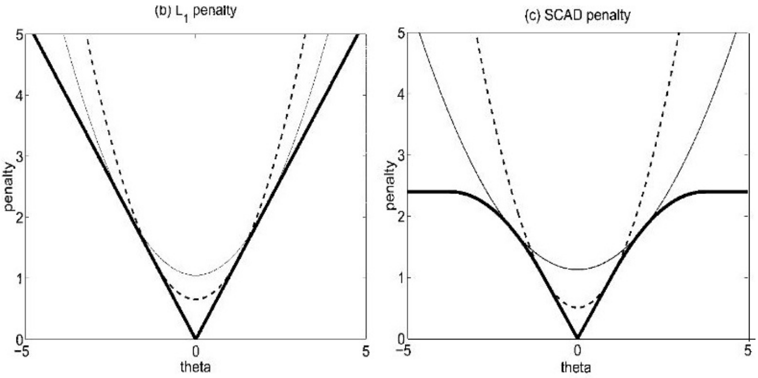
SCAD penalty function

- Smoothly Clipped Absolute Deviation (SCAD):

$$p_{\lambda_n}(|\beta|) = \begin{cases} \lambda_n|\beta| & , |\beta| \le \lambda_n \\ -(\dfrac{|\beta|^2 - 2a\lambda_n|\beta| + \lambda_n^2}{2(a-1)}), & \lambda_n < |\beta| \le a\lambda_n \\ \dfrac{(a+1)\lambda_n^2}{2} & , |\beta| > a\lambda_n \end{cases}$$

where $a = 3.7$(Fan and Li, 2001).

Plot of $p_\lambda(|\theta|)$:



Asymptotic properties of the LS-SCAD estimator

Assumptions(for fixed covariates)

Let k_n be the number of nonzero coefficients, $\rho_{n,1}$ be the smallest eigenvalue of $\frac{1}{n}\mathbf{X}^T\mathbf{X}$. let $\pi_{n,k_n}, \omega_{n,m_n}$ be the largest eigenvalues of $\frac{1}{n}\mathbf{X_1}^T\mathbf{X_1}$, $\frac{1}{n}\mathbf{X_2}^T\mathbf{X_2}$, respectively.

(A0) (a) ε_i's are *i.i.d* with mean 0 and variance σ^2;
(b) For $j \in \{1, ..., p_n\}$, $||\mathbf{x_j}||^2 = n$.($\mathbf{x_j}$ column vector of $\mathbf{X}$)

(A1) (a) $\lim_{n\to\infty} \sqrt{k_n}\lambda_n / \sqrt{\rho_{n,1}} = 0$;
(b) $\lim_{n\to\infty} \sqrt{p_n} / \sqrt{n\rho_{n,1}} = 0$.

(A2) (a) $\lim_{n\to\infty} \sqrt{k_n}\lambda_n / (\sqrt{\rho_{n,1}} \min_{1\le j\le k_n} |\beta_j|) = 0$;
(b) $\lim_{n\to\infty} \sqrt{p_n} / (\sqrt{n\rho_{n,1}} \min_{1\le j\le k_n} |\beta_j|) = 0$.

(A3) $\lim_{n\to\infty} \sqrt{\max(\pi_{n,k_n}, \omega_{n,m_n}) p_n} / (\sqrt{n}\rho_{n,1}\lambda_n) = 0$.

(A4) $\lim_{n\to\infty} \max_{1\le i\le n} \tilde{\mathbf{x}}_{i1} (\sum_{i=1}^{n} \tilde{\mathbf{x}}_{i1}^T \tilde{\mathbf{x}}_{i1})^{-1} \tilde{\mathbf{x}}_{i1}^T = 0$.($\tilde{\mathbf{x}}_{i1}$ row vector of $\mathbf{X_1}$)

Assumptions(for random covariates)

Let ρ_1 be the smallest eigenvalue of $E[\mathbf{x}\mathbf{x}^T]$. let π_{k_n}, ω_{m_n} be the largest eigenvalues of $E[\mathbf{X_1}^T\mathbf{X_1}]$, $E[\mathbf{X_2}^T\mathbf{X_2}]$, respectively.

(B0) $(\mathbf{x}_i^T, \varepsilon_i) = (X_{i1}, ..., X_{ip_n}, \varepsilon_i)$, $i = 1, ..., n$ are *i.i.d* with

(a) $E[X_{ij}] = 0$, $Var(X_{ij}) = 1$;

(b) $E[\varepsilon|\mathbf{X}] = 0$, $Var(\varepsilon|\mathbf{X}) = \sigma^2$.

(B1) (a) $\lim_{n\to\infty} p_n^2/(n\rho_1^2) = 0$;

(b) $\lim_{n\to\infty} k_n\lambda_n^2/\rho_1 = 0$.

(B2) (a) $\lim_{n\to\infty} \sqrt{p_n}/(\sqrt{n\rho_1}\min_{1\le j\le k_n}|\beta_j|) = 0$;

(b) $\lim_{n\to\infty} \lambda_n\sqrt{k_n}/(\sqrt{\rho_1}\min_{1\le j\le k_n}|\beta_j|) = 0$.

(B3) $\lim_{n\to\infty} \sqrt{\max(\pi_{k_n}, \omega_{m_n})p_n}/(\sqrt{n}\rho_1\lambda_n) = 0$.

Consistency

Theorem 1 (Consistency in the fixed design setting)

Under $(\mathrm{A0})-(\mathrm{A1})$,

$$||\hat{\beta}_n - \beta|| \xrightarrow{P} 0 \quad \text{as } n \to \infty.$$

Theorem 2 (Consistency in the random design setting)

Suppose that there exists an absolute constant M_4 *such that for all* n, $\max_{1 \leq j \leq p_n} E[X_j^4] \leq M_4 < \infty$. *Then under* $(\mathrm{B0})-(\mathrm{B1})$,

$$||\hat{\beta}_n - \beta|| \xrightarrow{P} 0 \quad \text{as } n \to \infty.$$

Convergency rate

Lemma 1 (Convergency rate in the fixed design setting)

Under $(\mathrm{A0}) - (\mathrm{A2})$,

$$||\hat{\beta}_n - \beta|| = O_P\left(\frac{\sqrt{p_n}}{\sqrt{n\rho_{n,1}}}\right).$$

Lemma 2 (Convergency rate in the random design setting)

Under $(\mathrm{B0}) - (\mathrm{B2})$,

$$||\hat{\beta}_n - \beta|| = O_P\left(\frac{\sqrt{p_n}}{\sqrt{n\rho_1}}\right).$$

Variable selection

Theorem 3 (Variable selection in the fixed design setting)

Under $(\mathrm{A0}) - (\mathrm{A3})$,

$$\hat{\beta}_{2n} = \mathbf{0}_{m_n} \textit{ with probability tending to 1.}$$

Theorem 4 (Variable selection in the random design setting)

Suppose there exists an absolute constant M such that $\max_{1 \le j \le p_n} |X_j| \le M < \infty$. *Then under* $(\mathrm{B0}) - (\mathrm{B3})$,

$$\hat{\beta}_{2n} = \mathbf{0}_{m_n} \textit{ with probability tending to 1.}$$

Asymptotic normality

Let $\{\mathbf{A}_n, n = 1, 2, ...\}$ be sequence of $d \times k_n$ matrices with full rank.

Theorem 5 (Asymptotic normality in the fixed design setting)

Under $(\mathrm{A0}) - (\mathrm{A4})$,

$$\sqrt{n}\Sigma_n^{-1/2}\mathbf{A}_n(\hat{\beta}_{1n} - \beta_1) \xrightarrow{D} N(\mathbf{0}_d, \mathbf{I}_d),$$

where $\Sigma_n = \sigma^2\mathbf{A}_n(\sum_{i=1}^n \tilde{\mathbf{x}}_{i1}^T\tilde{\mathbf{x}}_{i1}/n)^{-1}\mathbf{A}_n^T$.

Theorem 6 (Asymptotic normality in the random design setting)

Suppose that there exists an absolute constant M such that $\max_{1\le j\le p_n} ||X_j|| \le M < \infty$ *and a* σ_4 *such that* $E[\varepsilon^4|X_{11}] \le \sigma_4 < \infty$ *for all n. Then under* $(\mathrm{B0}) - (\mathrm{B3})$,

$$n^{-1/2}\Sigma_n^{-1/2}\mathbf{A}_n E^{-1/2}[\tilde{\mathbf{x}}_{i1}^T\tilde{\mathbf{x}}_{i1}]\sum_{i=1}^n \tilde{\mathbf{x}}_{i1}^T\tilde{\mathbf{x}}_{i1}(\hat{\beta}_{1n} - \beta_1) \xrightarrow{D} N(\mathbf{0}_d, \mathbf{I}_d),$$

where $\Sigma_n = \sigma^2\mathbf{A}_n\mathbf{A}_n^T$.

Lindeberg-Feller multivariate CLT

Theorem

Suppose $\mathbf{X}_i$ *is a sequence of independent random vectors with mean* $\boldsymbol{\mu}_i$*, covariance matrix* $\boldsymbol{\Sigma}_i$*. Assume that* $\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\Sigma}_i \rightarrow \Sigma$ *as* $n \rightarrow \infty$*.*

If for any $\epsilon > 0$*,*

$$\frac{1}{n}\sum_{i=1}^{n} E[||\mathbf{X}_i - \boldsymbol{\mu}_i||^2 I_{\{||\mathbf{X}_i-\boldsymbol{\mu}_i||>\epsilon\sqrt{n}\}}] \rightarrow 0 \text{ as } n \rightarrow \infty$$

then

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(\mathbf{X}_i - \boldsymbol{\mu}_i) \xrightarrow{D} N(\mathbf{0}, \boldsymbol{\Sigma}).$$

Order of p_n

- $(\mathrm{A2.a})$ and $(\mathrm{A2.b})$ are identical to $(\mathrm{A1.a})$ and $(\mathrm{A1.b})$, if $\liminf_{n\to\infty} \min_{1\le j\le k_n} |\beta_j| > 0$.

$(\mathrm{A1})$ (a) $\lim_{n\to\infty} \frac{\sqrt{k_n}\lambda_n}{\sqrt{\rho_{n,1}}} = 0$;

(b) $\lim_{n\to\infty} \frac{\sqrt{p_n}}{\sqrt{n\rho_{n,1}}} = 0$.

$(\mathrm{A2})$ (a) $\lim_{n\to\infty} \frac{\sqrt{k_n}\lambda_n}{\sqrt{\rho_{n,1}}\min_{1\le j\le k_n}|\beta_j|} = 0$;

(b) $\lim_{n\to\infty} \frac{\sqrt{p_n}}{\sqrt{n\rho_{n,1}}\min_{1\le j\le k_n}|\beta_j|} = 0$.

Order of p_n

- $(\mathrm{A3})\ \lim_{n\to\infty} \frac{\sqrt{\max(\pi_{n,k_n},\omega_{n,m_n})p_n}}{\sqrt{n}\rho_{n,1}\lambda_n} = 0$ is implied by $\lim_{n\to\infty} \frac{p_n}{\sqrt{n}\rho_{n,1}\lambda_n} = 0$ (since $\pi_{n,k_n} \leq k_n$, $\omega_{n,m_n} \leq m_n$)

- Suppose $\liminf_{n\to\infty} \min_{1\leq j\leq k_n} |\beta_j| > 0$ and $\liminf_{n\to\infty} \rho_{n,1} > 0$.

Then,

(i) if $p_n = o(n^{\frac{1}{3}})$ and take suitable λ_n (e.g. $\lambda_n = O(n^{-\frac{1}{6}})$), then $(\mathrm{A1}) - (\mathrm{A3})$ can be satisfied.

(However, if $p_n = o(n^{\frac{1}{2}})$, then $\lambda_n = O(1) \Rightarrow (\mathrm{A1.a})$ fails)

(ii) Futhermore, suppose either k_n is fixed, or the largest eigenvalue of $\frac{1}{n}\mathbf{X}^T\mathbf{X}$ is bounded above.

Then $p_n = o(n^{\frac{1}{2}})$ is sufficient.

Conclusion

Conclusion

- Asymptotic properties of LS-SCAD estimator:
  consistency, convergence rate, sparsity, asymptotic normality.
- Order of p_n that is sufficient for asymptotic properties.

Reference :

[1]Fan, Jianqing, and Runze Li. "Variable selection via nonconcave penalized likelihood and its oracle properties." Journal of the American statistical Association 96.456 (2001): 1348-1360.
[2]TIBSHIRANI, R.(1996)Regression Shrinkage and Selection via the Lasso. J.R.Stat.Soc.B.58, No.1, 267-288.