Asymptotic Properties of Bridge Estimators in Sparse High-Dimensional Regression Models

Jian Huang   Joel Horowitz   Shuangge Ma

Presenter: Minjing Tao

April 16, 2010

Outline

Outline

Outline

Linear Regression Model

Consider the linear regression model

$$Y_i = \beta_0 + \mathbf{x}_i' \boldsymbol{\beta} + \epsilon_i, \qquad i = 1, \cdots, n,$$

where $Y_i \in \mathbb{R}$ is a response variable, $\mathbf{x}_i$ is a $p_n \times 1$ covariate vector and ϵ_i's are i.i.d. random error terms.

- Assume: $\beta_0 = 0$ (It can be achieved by centering the response and covariates.)
- Interested in: estimating the vector of regression coefficients $\boldsymbol{\beta}$ when p_n may go to **infinity** and $\boldsymbol{\beta}$ is **sparse** (many of its elements are zero).

Bridge Estimator

Penalized least squares objective function

$$L_n(\beta) = \sum_{i=1}^{n}(Y_i - \mathbf{x}_i'\beta)^2 + \lambda_n \sum_{j=1}^{p_n} |\beta_j|^\gamma, \quad (1)$$

where λ_n is a penalty parameter, and $\gamma > 0$.

Definition (Bridge Estimator)

The value $\widehat{\beta}_n$ that minimizes (1) is called a bridge estimator [Frank and Friedman (1993) and Fu (1998)].

- When $\gamma = 2$, it is the ridge estimator [Hoerl and Kennard (1970)].
- When $\gamma = 1$, it is the LASSO estimator [Tibshirani (1996)].

A Property of Bridge Estimator

- Knight and Fu (2000): when $0 < \gamma \leq 1$, some components of the bridge estimator can be **exactly zero** if λ_n is sufficiently large.
  $\Rightarrow$ The bridge estimator for $0 < \gamma \leq 1$ provides a way to achieve variable selection and parameter estimation in a **single** step.
- In this paper: $0 < \gamma < 1$ is concerned.

Outline

Bridge Estimator: Knight and Fu (2000)

Knight and Fu (2000) studies the asymptotic properties of bridge estimators when the number of covariates is **finite.** They showed that, under appropriate regularity conditions,

- the bridge estimator is consistent;
- for $0 < \gamma \leq 1$, the limiting distributions can have positive probability mass at 0 when the true value of the parameter is zero;
- the usage of bridge estimators: distinguish the covariates with coefficients between **exactly zero** and **nonzero.**

Another Penalization Method: SCAD

For the SCAD penalty, Fan and Peng (2004) studied asymptotic properties of penalized likelihood methods. They showed there exist local maximizers that have an **oracle property**:

- correctly select the nonzero coefficients with probability converging to 1;
- the estimators of the nonzero coefficients are asymptotically normal with the **same** means and covariances that they would have if the zero coefficients were known in advance.

Outline

What You Can Expect Is ...

- Extend the results of Knight and Fu (2000) to infinite-dimensional parameter settings. It is proved that bridge estimator is **consistent** for any $\gamma > 0$.
- Show that under $0 < \gamma < 1$, the bridge estimator has the similar **oracle property** as Fan and Peng (2004).

Limitation: the condition that $p_n < n$ is needed, for identification and consistent estimation of the regression parameter.

- In studies of relationships between a phenotype and microarray gene expression profiles, the number of genes (covariates) is typically much greater than the sample size.

The $p_n > n$ Scenario

Motivation: How to deal with the “not identifiable” problem?
If **X** are mutually orthogonal,

- Each regression coefficient can be estimated by univariate regression.
- This assumption is too strong.

Answer: use the **marginal bridge estimator** under a **partial orthogonality condition**.

- The marginal bridge estimator can **consistently** distinguish between zero and nonzero coefficients, although the the estimation is not consistent.
- The good estimator can be obtained by a two-step approach.

Outline

Outline

1. Introduction
   - The Definition of Bridge Estimator
   - Related Work
   - Major Contribution of this Paper
2. Asymptotic Properties of Bridge Estimators
   - Scenario 1: $p_n < n$ (Consistency and Oracle Property)
   - Scenario 2: $p_n > n$ (A Two-Step Approach)
3. Numerical Studies
4. Summary

Notations

- $\boldsymbol{\beta}_0$: true parameter. Let $\boldsymbol{\beta}_0 = (\boldsymbol{\beta}_{10}', \boldsymbol{\beta}_{20}')'$, where $\boldsymbol{\beta}_{10}$ (nonzero coefficients) is a $k_n \times 1$ vector, and $\boldsymbol{\beta}_{20} = \mathbf{0}$ is a $m_n \times 1$ vector.
- $\mathbf{x}_i = (x_{i1}, \cdots, x_{ip_n})'$ is a $p_n \times 1$ vector of covariates of the ith observation.
- $\mathbf{x}_i = (\mathbf{w}_i', \mathbf{z}_i')'$, where $\mathbf{w}_i$ is corresponding to the nonzero coefficients, and $\mathbf{z}_i$ to the zero coefficients.
- $\mathbf{X}_n = (\mathbf{x}_1, \cdots, \mathbf{x}_n)'$, $\mathbf{X}_{1n} = (\mathbf{w}_1, \cdots, \mathbf{w}_n)'$, and $\mathbf{X}_{2n} = (\mathbf{z}_1, \cdots, \mathbf{z}_n)'$.
- $\Sigma_n = n^{-1}\mathbf{X}_n'\mathbf{X}_n$ and $\Sigma_{1n} = n^{-1}\mathbf{X}_{1n}'\mathbf{X}_{1n}$.
- Let ρ_{1n} (ρ_{2n}) and τ_{1n} (τ_{2n}) be the smallest (largest) eigenvalue of Σ_n and Σ_{1n}, respectively.

Assumptions

- The covariates are assumed to be **fixed**. But for the random covariates, the results hold conditionally on the covariates.
- Assume that Y_i's are centered and the covariates are standardized, i.e.,

$$\sum_{i=1}^{n} Y_i = 0, \quad \sum_{i=1}^{n} x_{ij} = 0, \quad \frac{1}{n}\sum_{i=1}^{n} x_{ij}^2 = 1$$

Regularity Conditions

(A1) Error Terms

$\epsilon_1, \epsilon_2, \cdots$ are i.i.d. r.v.'s with mean 0 and variance σ^2, where $0 < \sigma^2 < \infty$.

(A2) Smallest eigenvalue of Σ_n

(a) $\rho_{1n} > 0$ for all n; (b) $(p_n + \lambda_n k_n)(n\rho_{1n})^{-1} \to 0$

Notes:

- (A2)(a) implies that Σ_n is nonsingular for each n, but it allows $\rho_{1n} \to 0$
- (A2)(b) is a condition needed in the proof of consistency. And $\sqrt{(p_n + \lambda_n k_n)/(n\rho_{1n})}$ is part of the consistent rate of the bridge estimator.

Regularity Conditions

(A3) Restrictions on λ_n, k_n and p_n

(a) $\lambda_n(k_n/n)^{1/2} \to 0$; (b) $\lambda_n n^{-\gamma/2}(\rho_{1n}/\sqrt{p_n})^{2-\gamma} \to \infty$

It's needed in the proof of consistency and oracle property. If ρ_{1n} is bounded away from 0 and ∞ for all n, and k_n is finite, then

(A3)' Simplified Version of (A3)

(a) $\lambda_n n^{-1/2} \to 0$; (b) $\lambda_n^2 n^{-\gamma} p_n^{-(2-\gamma)} \to \infty$

- The penalty parameter λ_n must always be $o(n^{1/2})$.
- The smaller the γ, the larger p_n is allowed. For $\gamma = 0$, $p_n = o(n^{1/2})$.
- If $\gamma = 1$, then (A3)'(b) becomes $(\lambda_n^2 n^{-1})/p_n \to \infty$, which is impossible. Therefore, (A3)'(b) excludes $\gamma = 1$ (LASSO).

Regularity Conditions

(A4) Nonzero Coefficients

There exist constants $0 < b_0 < b_1 < \infty$ such that

$$b_0 \le \min\{|\beta_{1j}|, 1 \le j \le k_n\} \le \max\{|\beta_{1j}|, 1 \le j \le k_n\} \le b_1$$

This condition assumes the nonzero coefficients are uniformly bounded away from 0 and ∞.

Regularity Conditions

(A5) Condition on $\Sigma_{1n} = n^{-1}X'_{1n}X_{1n}$

(a) There exist constants $0 < \tau_1 < \tau_2 < \infty$ such that $\tau_1 \le \tau_{1n} \le \tau_{2n} \le \tau_2$ for all n;
(b) $n^{-1/2} \max_{1 \le i \le n} \mathbf{w}_i'\mathbf{w}_i \to 0$.

- (a) assumes Σ_{1n} is strictly positive definite. In the sparse problems, k_n is small relative to n. Then this assumption is reasonable.
- (b) is needed in the proof of asymptotic normality of nonzero coefficients. In fact, if all the covariates corresponding to the nonzero coefficients are bounded by a constant C, then by condition (A3)(a),

$$n^{-1/2} \max_{1 \le i \le n} \mathbf{w}_i'\mathbf{w}_i \le n^{-1/2} k_n C \to 0$$

Consistency

Theorem 1 (Consistency)

Let $\widehat{\beta}_n$ denote the minimizer of (1). Suppose that $\gamma > 0$ and that conditions (A1), (A2), (A3)(a) and (A4) hold. Let $h_n = \rho_{1n}^{-1}(p_n/n)^{1/2}$ and $h'_n = [(p_n + \lambda_n k_n)/(n\rho_{1n})]^{1/2}$. Then $||\widehat{\beta}_n - \beta_0|| = O_p(\min\{h_n, h'_n\})$

Notes:

- Theorem 1 states that the variable selection and coefficient estimation can be achieved in one single step.
- It holds for any $\gamma > 0$, including LASSO and ridge estimators.

Consistency

Discussion: The Convergence Rate.

The convergence rate is $O_p(\min\{h_n, h'_n\})$, where $h_n = \rho_{1n}^{-1}(p_n/n)^{1/2}$ and $h'_n = [(p_n + \lambda_n k_n)/(n\rho_{1n})]^{1/2}$.

- If $\rho_{1n} > \rho_1 > 0$ for all n, then $\min\{h_n, h'_n\} = h_n$, and the convergence rate is $O_p((p_n/n)^{1/2})$.
- Furthermore, if p_n is finite, then the rate is the familiar $n^{-1/2}$.
- If $\rho_{1n} \to 0$, then h_n may not converge to zero faster than h'_n. And the convergence rate will be slower than $n^{-1/2}$.

Oracle Property

Theorem 2 (Oracle Property)

Let $\widehat{\boldsymbol{\beta}}_n = (\widehat{\boldsymbol{\beta}}_{1n}, \widehat{\boldsymbol{\beta}}_{2n})$, where $\widehat{\boldsymbol{\beta}}_{1n}$ and $\widehat{\boldsymbol{\beta}}_{2n}$ are estimators of $\boldsymbol{\beta}_{10}$ and $\boldsymbol{\beta}_{20}$, respectively. Suppose that $0 < \gamma < 1$ and that conditions (A1) to (A5) are satisfied. We have the following:

1. $\widehat{\boldsymbol{\beta}}_{2n} = \mathbf{0}$ with probability converging to 1.
2. Let $s_n^2 = \sigma^2 \boldsymbol{\alpha}_n' \Sigma_{1n}^{-1} \boldsymbol{\alpha}_n$ for any $k_n \times 1$ vector $\boldsymbol{\alpha}_n$ satisfying $||\boldsymbol{\alpha}_n||_2 \leq 1$. Then

$$
\begin{aligned}
& n^{1/2} s_n^{-1} \boldsymbol{\alpha}_n' (\widehat{\boldsymbol{\beta}}_{1n} - \boldsymbol{\beta}_{10}) \\
=& n^{1/2} s_n^{-1} \sum_{i=1}^{n} \epsilon_i \boldsymbol{\alpha}_n' \Sigma_{1n}^{-1} \mathbf{w}_i + o_p(1) \rightarrow_D N(0, 1)
\end{aligned}
$$

where $o_p(1)$ converges to zero in prob uniformly w.r.t. $\boldsymbol{\alpha}_n$.

Oracle Property

Discussion: Asymptotic Normality for $\widehat{\beta}_{1nj}$

- Let $\widehat{\beta}_{1nj}$ and β_{10j} be the jth components of $\widehat{\boldsymbol{\beta}}_{1n}$ and $\boldsymbol{\beta}_{10}$, respectively.
- Set $\alpha_n = \mathbf{e}_j$, where $\mathbf{e}_j$ is the unit vector whose only nonzero element is the jth element. Then $s_n^2 = \sigma^2 \mathbf{e}_j' \Sigma_{1n}^{-1} \mathbf{e}_j$, and denote it as s_{nj}^2.
- Applying Theorem 2 (2), we have

$$n^{1/2} s_{nj}^{-1} (\widehat{\beta}_{1nj} - \beta_{10j}) \rightarrow_D N(0, 1)$$

Outline

Marginal Bridge Objective Function

Recall that:

- In some problems, like the "phenotype and microarray gene expression ", Theorem 1 and 2 are not applicable.
- In the scenario $p_n > n$, extra condition on the design matrix is needed. (partial orthogonality condition).
- A univariate version of bridge estimator, marginal bridge estimator, is studied.

Definition

The marginal bridge objective function is

$$U_n(\beta) = \sum_{j=1}^{p_n} \sum_{i=1}^{n} (Y_i - x_{ij}\beta_j)^2 + \lambda_n \sum_{j=1}^{p_n} |\beta_j|^\gamma$$

New Notations

- $\tilde{\beta}_n$: marginal bridge estimator (the value minimizes U_n). Write $\tilde{\beta}_n = (\tilde{\beta}'_{n1}, \tilde{\beta}'_{n2})'$ according to the partition $\beta_0 = (\beta'_{10}, \beta'_{20})'$.
- Let $K_n = \{1, \cdots, k_n\}$ and $J_n = \{k_n + 1, \cdots, p_n\}$ be the set of indices of nonzero and zero coefficients, respectively.
- ξ_{nj}: the "covariance" between the jth covariate and the response variable.

$$\xi_{nj} = n^{-1} E\left(\sum_{i=1}^{n} Y_i x_{ij}\right) = n^{-1} \sum_{i=1}^{n} (\mathbf{w}'_i \beta_{10}) x_{ij}$$

Therefore, ξ_{nj}/σ is the correlation coefficient.

New Regularity Conditions

(B1) Error Terms

(a) $\epsilon_1, \epsilon_2, \cdots$ are i.i.d. r.v.'s with mean 0 and variance σ^2, where $0 < \sigma^2 < \infty$.
(b) ϵ_i's are sub-Gaussian, i.e., the tail probability satisfying $P(|\epsilon_i| > x) \leq K \exp(-Cx^2)$ for constants C and K.

Note: There is no normality assumption about the error terms. Instead, the tails of the error distribution should behave like normal tails.

New Regularity Conditions

(B2) Partial Orthogonality Condition

(a) There exists a constant $c_0 > 0$ such that

$$\left| n^{-1/2} \sum_{i=1}^{n} x_{ij} x_{ik} \right| \leq c_0, \quad j \in J_n, k \in K_n,$$

for all n sufficiently large.
(b) There exists a constant $\xi_0 > 0$ such that $\min_{k \in K_n} |\xi_{nj}| > \xi_0$.

- Condition (a) assumes that the covariates of the nonzero and zero coefficients are only weekly correlated
- Condition (b) requires the correlations between covariates with nonzero coefficients and response are bounded away from zero.

New Regularity Conditions

(B3) Restrictions on λ_n, k_n and m_n

(a) $\lambda_n/n \to 0$ and $\lambda_n n^{-\gamma/2} k_n^{\gamma-2} \to \infty$;
(b) $\log(m_n) = o(1) \times (\lambda_n n^{-\gamma/2})^{2/(2-\gamma)}$

Notes:

1. $\lambda_n = o(n)$, $k_n = o(n^{1/2})$, and $\log(m_n) = o(n)$.
2. "Sparse" requires $k_n = o(n^{1/2})$.
3. The condition permits $p_n/n \to \infty$.

(B4) Nonzero Coefficients

There exists a constant $0 < b_1 < \infty$ such that $\max_{k \in K_n} |\beta_{1k}| \leq b_1$

Correctly Identify Zero and Nonzero Coefficients

Theorem 3

Suppose that conditions (B1) to (B4) hold and that $0 < \gamma < 1$. Then

$$P(\tilde{\beta}_{n2} = \mathbf{0}) \to 1 \quad \text{and} \quad P(\tilde{\beta}_{n1k} \neq 0, k \in K_n) \to 1.$$

- The estimators of nonzero coefficients are not consistent.
- To get consistent estimators, a two-step approach is needed.
  1. First step: using marginal bridge estimator (by Theorem 3).
  2. Second step: any reasonable regression method can be used.

The Second Step: Bridge Regression

- Assume that only the covariates with nonzero coefficients are included in the model in this step.
- Let $\widehat{\boldsymbol{\beta}}^*_{1n}$ be the estimator. It's defined as the value minimizing

Step 2: Bridge Regression

$$U_n(\boldsymbol{\beta}_1)^* = \sum_{i=1}^{n}(Y_i - \mathbf{w}_i'\boldsymbol{\beta}_1)^2 + \lambda_n^* \sum_{j=1}^{k_n} |\beta_{1j}|^\gamma,$$

where $\boldsymbol{\beta}_1 = (\beta_{11}, \cdots, \beta_{1k_n})$

Additional Regularity Conditions

(B5) Conditions on Σ_{1n}

(a) There exists a constant $\tau_1 > 0$ such that $\tau_{1n} \geq \tau_1$ for all n sufficiently large;
(b) The covariates of nonzero coefficients satisfy $n^{-1/2} \max_{1 \leq i \leq n} \mathbf{w}_i' \mathbf{w}_i \to 0$.

It's similar to condition (A5).

(B6) Restrictions on k_n and λ_n^*

(a) $k_n(1 + \lambda_n^*)/n \to 0$; (b) $\lambda_n^*(k_n/n)^{1/2} \to 0$.

Note: From (B6), one can set $\lambda_n^* = 0$ for all n. Then $\widehat{\beta}_{1n}^*$ is the OLS estimator.

Asymptotic Normality of $\widehat{\beta}^*_{1n}$

Theorem 4

Suppose that conditions (B1) to (B6) hold and that $0 < \gamma < 1$. Let $s_n^2 = \sigma^2 \alpha_n' \Sigma_{1n}^{-1} \alpha_n$ for any $k_n \times 1$ vector α_n satisfying $||\alpha_n||_2 \leq 1$. Then

$$
\begin{aligned}
& n^{1/2} s_n^{-1} \alpha_n' (\widehat{\beta}^*_{1n} - \beta_{10}) \\
= & n^{1/2} s_n^{-1} \sum_{i=1}^{n} \epsilon_i \alpha_n' \Sigma_{1n}^{-1} \mathbf{w}_i + op(1) \rightarrow_D N(0,1)
\end{aligned}
$$

where $o_p(1)$ is a term that converges to zero in probability uniformly w.r.t. α_n.

Note: This is the same result as Theorem 2 (2).

Outline

Simulation

There are six examples simulated in the paper, all the data from the model

$$y = \mathbf{x}'\beta + \epsilon, \quad \epsilon \sim N(0, \sigma^2)$$

where

- $\sigma = 1.5$
- $\mathbf{x}$ is generated from a multivariate normal with marginal distributions being standard normal $N(0, 1)$.
- $n = 100$
- Number of covariates with nonzero coefficients is 15.

Six Examples

Example 1

- $p = 30$
- The pairwise correlation between the ith and the jth components of $\mathbf{x}$ is $r^{|i-j|}$ with $r = 0.5$
- The true β is $(\underbrace{2.5, \cdots, 2.5}_{5}, \underbrace{1.5, \cdots, 1.5}_{5}, \underbrace{0.5, \cdots, 0.5}_{5}, 0, \cdots)$

Example 2

The same as Example 1, except that $r = 0.95$

Six Examples

Example 3

- $p = 30$
- The covariates are generated as follows:

$$\begin{array}{lll} x_i = Z_1 + e_i, & Z_1 \sim N(0,1), & i = 1, \ldots, 5 \\ x_i = Z_2 + e_i, & Z_2 \sim N(0,1), & i = 6, \ldots, 10 \\ x_i = Z_3 + e_i, & Z_3 \sim N(0,1), & i = 11, \ldots, 15 \\ x_i \sim N(0,1), & x_i \text{ i.i.d.} & i = 16, \ldots, 30 \end{array}$$

  where e_i are i.i.d. $N(0, 0.01), i = 1, \ldots, 15$
- The true β is $(\underbrace{1.5, \cdots, 1.5}_{15}, 0, \cdots)$

Six Examples

Example 4

- $p = 200$
- The first 15 covariates and the remaining 185 covariates (two groups) are independent.
- The pairwise correlation between the ith and the jth components within two groups is $r^{|i-j|}$ with $r = 0.5$
- The true β is $(\underbrace{2.5, \cdots, 2.5}_{5}, \underbrace{1.5, \cdots, 1.5}_{5}, \underbrace{0.5, \cdots, 0.5}_{5}, 0, \cdots)$

Example 5

The same as Example 4, except that $r = 0.95$

Six Examples

Example 6

- $p = 500$
- The first 15 covariates are generated the same way as in Example 5.
- The remaining 485 covariates are independent of the first 15 covariates and are generated independently from $N(0,1)$.
- The true β is $(\underbrace{1.5, \cdots, 1.5}_{15}, 0, \cdots)$

Result 1: Prediction MSE

TABLE 1

Simulation study: comparison of OLS, RR, LASSO, Elastic net and the bridge estimator with $\gamma = 1/2$. PMSE: median of PMSE, inside "(·)" are the corresponding standard deviations. Covariate: median of number of covariates with nonzero coefficients

Example		OLS	RR	LASSO	ENet	Bridge
1	PMSE	3.32 (0.58)	3.51 (0.69)	2.92 (0.51)	2.80 (0.47)	2.95 (0.51)
	Covariate	30	30	23	22	17
2	PMSE	3.21 (0.53)	2.65 (0.41)	2.60 (0.40)	2.46 (0.35)	2.37 (0.36)
	Covariate	30	30	18	16	15
3	PMSE	3.26 (0.58)	3.34 (0.58)	2.66 (0.40)	2.38 (0.33)	2.31 (0.34)
	Covariate	30	30	18	15	15
4	PMSE	–	20.45 (2.02)	3.55 (0.64)	3.30 (0.53)	3.98 (0.83)
	Covariate	–	200	37	37	29
5	PMSE	–	5.80 (1.31)	2.71 (0.42)	2.50 (0.36)	2.64 (0.44)
	Covariate	–	200	25	16	15
6	PMSE	–	43.10 (2.23)	3.51 (0.57)	2.70 (0.49)	2.68 (0.39)
	Covariate	–	500	43	20	17

Result 2: Probability of Correctly Identified

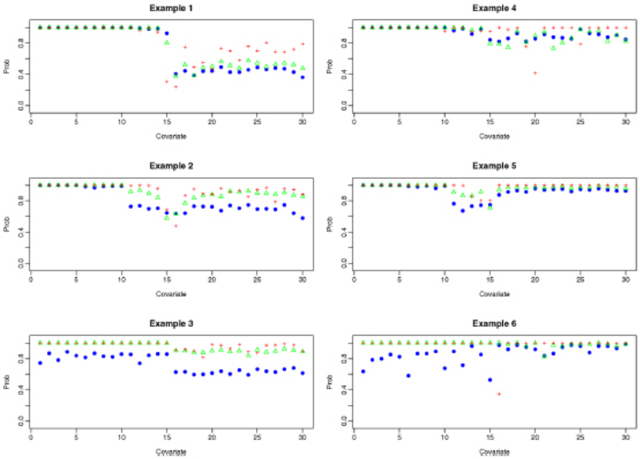

FIG. 1. *Simulation study (Examples 1–6): probability of individual covariate effect being correctly identified. Circle: LASSO; Triangle: ENet; Plus sign: Bridge estimate.*

Outline

Summary

- The asymptotic properties of bridge estimators is studied when p_n and k_n increase to infinity.
- When $0 < \gamma < 1$, bridge estimators correctly identify zero coefficients with probability converging to one, and that the estimators of nonzero coefficients are asymptotically normal and oracle efficient, under the scenario $p_n < n$.
- For the scenario $p_n > n$, a marginal bridge estimator is considered under the partial orthogonality condition. It can consistently distinguish covariates of zero and nonzero coefficients.
- In this scenario, the number of zero coefficients can be in the order of $o(e^n)$.