Asymptotics for LASSO-type Estimators

Ganchao Wei

December 6, 2021

Overview

Introduction

Consider the linear regression model:

$$Y_i = \beta_0 + \boldsymbol{x}_i^T \boldsymbol{\beta} + \epsilon_i$$

We estimate $\boldsymbol{\beta}$ by minimizing the penalized least squares criterion:

$$\sum_{i=1}^{n} (Y_i - \boldsymbol{x}_i^T \phi)^2 + \lambda_n \sum_{j=1}^{p} |\phi_j|^\gamma$$

Such estimators were called Bridge estimators. When $\lambda_n = 0$, it corresponds to the OLS estimator, denoted by $\hat{\boldsymbol{\beta}}_n^{(0)}$.
Some special cases:

- $\gamma = 2$, ridge regression
- $\gamma = 1$, LASSO regression
- $\gamma \to 0$, penalize by the number of nonzero parameters, e.g. AIC & BIC.

Introduction

Regularity conditions:

- $C_n = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i^T \boldsymbol{x}_i \to C$
- C_n is not singular (although this can be relaxed by equivalence class)
- First assume C is nonsingular. It will be further relaxed when discussing "nearly singular" design.
- $\frac{1}{n}\max_{1\leq i\leq n} \boldsymbol{x}_i^T \boldsymbol{x}_i \to 0$

Under the regularity conditions, we know the OLS estimator is consistent and:

$$\sqrt{n}(\hat{\beta}_n^{(0)} - \beta) \to_d N(0, \sigma^2 C^{-1})$$

Limiting distributions

Assume C is nonsingular. Define the (random) function:

$$Z_n(\phi) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \mathbf{x}_i^T\phi)^2 + \frac{\lambda_n}{n}\sum_{j=1}^{p}|\phi_j|^\gamma$$

, which is minimized at $\phi = \hat{\beta}_n$. The following result shows that $\hat{\beta}_n$ is consistent provided $\lambda_n = o(n)$

THEOREM 1. *If C in (3) is nonsingular and $\lambda_n/n \to \lambda_0 \geq 0$, then $\hat{\boldsymbol{\beta}}_n \to_p$ argmin(Z) where*

$$Z(\boldsymbol{\phi}) = (\boldsymbol{\phi} - \boldsymbol{\beta})^T C(\boldsymbol{\phi} - \boldsymbol{\beta}) + \lambda_0 \sum_{j=1}^{p} |\phi_j|^\gamma.$$

Thus if $\lambda_n = o(n)$, argmin$(Z) = \boldsymbol{\beta}$ *and so $\hat{\boldsymbol{\beta}}_n$ is consistent.*

Limiting distributions

Although $\lambda_n = o(n)$ is sufficient for consistency, we require that λ_n grow slowly for $\sqrt{n}$-consistency, but not too small to reduce to OLS.
For $\gamma \geq 1$, we need $\lambda_n = O(\sqrt{n})$

THEOREM 2. *Suppose that $\gamma \geq 1$. If $\lambda_n/\sqrt{n} \to \lambda_0 \geq 0$ and C is nonsingular then*

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \to_d \operatorname{argmin}(V),$$

where

$$V(\mathbf{u}) = -2\mathbf{u}^T\mathbf{W} + \mathbf{u}^T C\mathbf{u} + \lambda_0 \sum_{j=1}^{p} u_j \operatorname{sgn}(\beta_j)|\beta_j|^{\gamma-1}$$

if $\gamma > 1$,

$$V(\mathbf{u}) = -2\mathbf{u}^T\mathbf{W} + \mathbf{u}^T C\mathbf{u} + \lambda_0 \sum_{j=1}^{p} [u_j \operatorname{sgn}(\beta_j) I(\beta_j \neq 0) + |u_j| I(\beta_j = 0)]$$

if $\gamma = 1$, and $\mathbf{W}$ has a $N(\mathbf{0}, \sigma^2 C)$ distribution.

Limiting distributions

Although $\lambda_n = O(\sqrt{n})$ suffices for $\gamma < 1$, we further suggests $\lambda_n = O(n^{\gamma/2})$ for $\gamma < 1$

THEOREM 3. *Suppose that $\gamma < 1$. If $\lambda_n / n^{\gamma/2} \to \lambda_0 \geq 0$ then*

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \to_d \operatorname{argmin}(V),$$

where

$$V(\mathbf{u}) = -2\mathbf{u}^T\mathbf{W} + \mathbf{u}^T C\mathbf{u} + \lambda_0 \sum_{j=1}^{p} |u_j|^\gamma I(\beta_j = 0).$$

This is interesting:

- Estimate nonzero regression parameters at the usual rate, without asymptotic bias
- Shrink the estimates of zeros regression parameters to 0, with positive probability
- This is in contrast to theorem 2 ($\gamma \geq 1$): bias is $\lambda_0 > 0$

In the following example, $\beta_1 > 0$ and $\beta_2 = 0$

EXAMPLE 1. Consider a quadratic regression model

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 \frac{(x_i - a_n^{(1)})}{s_n^{(1)}} + \beta_2 \frac{(x_i^2 - a_n^{(2)})}{s_n^{(2)}} + \varepsilon_i \\ &= \beta_0 + \beta_1 z_{1i} + \beta_2 z_{2i} + \varepsilon_i \quad \text{for } i = 1, \ldots, n, \end{aligned}$$

Limiting distributions

In this example, we will consider the cases $\gamma = 1$ and $\gamma = 1/2$ with $\lambda_n/n^{\gamma/2} \to \lambda_0 > 0$ and $\beta_1 > 0$, $\beta_2 = 0$. Then

$$\sqrt{n}\begin{pmatrix} \hat{\beta}_{nl} - \beta_1 \\ \hat{\beta}_{n2} \end{pmatrix} \to_d \text{argmin}(V),$$

where

$$V(u_1, u_2) = -2(u_1 W_1 + u_2 W_2) + u_1^2 + \frac{\sqrt{15}}{2} u_1 u_2 + u_2^2 + \lambda_0 [u_1 + |u_2|]$$

for $\gamma = 1$,

$$V(u_1, u_2) = -2(u_1 W_1 + u_2 W_2) + u_1^2 + \frac{\sqrt{15}}{2} u_1 u_2 + u_2^2 + \lambda_0 |u_2|^{1/2}$$

for $\gamma = 1/2$, and (W_1, W_2) is a zero mean bivariate Normal random vector with covariance matrix $\sigma^2 C$.

Limiting distributions

For $\gamma = 1$:

- the asymptotic variances decreases as λ_0 increases.
- As λ_0 increases, the asymptotic bias of $\hat{\beta}_{n1}$ becomes increasingly negative, while $\hat{\beta}_{n2}$ increase away from 0 then decreases to 0.

TABLE 1

Properties of the distribution of argmin(V) *for* $\gamma = 1$ *and various values of* λ_0

$\frac{\lambda_0}{\sigma}$	$\frac{E(\widehat{U}_1)}{\sigma}$	$\frac{E(\widehat{U}_2)}{\sigma}$	$\frac{\mathrm{Var}(\widehat{U}_1)}{\sigma^2}$	$\frac{\mathrm{Var}(\widehat{U}_2)}{\sigma^2}$	$\mathrm{Corr}(\widehat{U}_1, \widehat{U}_2)$	$P(\widehat{U}_2 = 0)$
0.0	0.00	0.00	16.00	16.00	−0.968	0.000
0.1	−0.68	0.65	11.90	11.62	−0.957	0.156
0.2	−1.14	1.07	8.89	8.49	−0.944	0.290
0.5	−1.71	1.50	6.16	5.53	−0.915	0.488
1.0	−1.93	1.47	5.78	5.10	−0.909	0.525
2.0	−2.33	1.37	5.36	4.71	−0.901	0.550
5.0	−3.51	1.04	4.40	3.63	−0.876	0.624

Limiting distributions

But for $\gamma = 0.5$, things are different (in terms of the asymptotic bias):

TABLE 2

Properties of the distribution of argmin(V) *for* $\gamma = 0.5$ *and various values of* λ_0

$\frac{\lambda_0}{\sigma^{3/2}}$	$\frac{E(\widehat{U}_1)}{\sigma}$	$\frac{E(\widehat{U}_2)}{\sigma}$	$\frac{\mathrm{Var}(\widehat{U}_1)}{\sigma^2}$	$\frac{\mathrm{Var}(\widehat{U}_2)}{\sigma^2}$	$\mathrm{Corr}(\widehat{U}_1, \widehat{U}_2)$	$P(\widehat{U}_2 = 0)$
0.0	0.00	0.00	16.00	16.00	−0.968	0.000
0.1	0.00	0.00	14.86	14.78	−0.966	0.193
0.2	0.00	0.00	13.73	13.57	−0.963	0.303
0.5	0.00	0.00	10.77	10.41	−0.952	0.529
1.0	0.00	0.00	7.06	6.46	−0.926	0.745
2.0	0.00	0.00	3.09	2.21	−0.821	0.930
5.0	0.00	0.00	1.05	0.04	−0.197	0.999

Limiting distributions

Scatter plot of random samples (500) from limiting distribution. OLS, $\lambda_0 = 0$

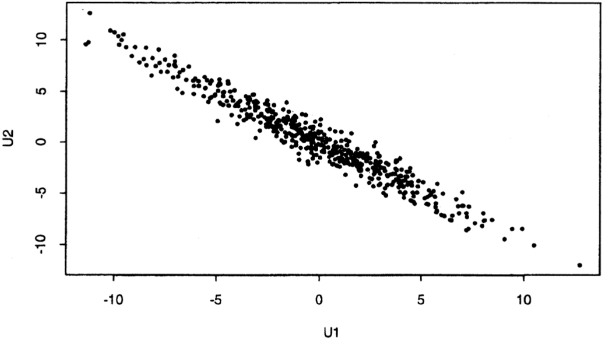

FIG. 1. *Sample of* 500 *from the limiting distribution of the LS estimator in Example* 1.

Strong correlation: overestimation of β_1 always accompanied by underestimation of β_2 (and vice versa).

Limiting distributions

LASSO, $\lambda = 1$

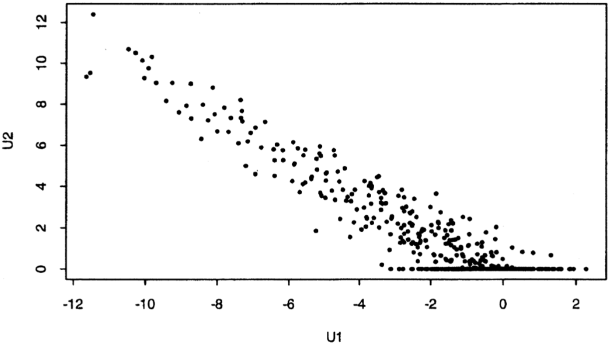

FIG. 2. *Sample of* 500 *from the limiting distribution of the Bridge estimator in Example* 1 *with* $\gamma = 1$ *and* $\lambda_0 = 1$. *The probability that* $\widehat{U}_2$ *is strictly less than* 0 *is approximately* 4.1×10^{-5}, *which explains the absence of negative* $\widehat{U}_2$ *values.*

Effectively sets the estimate of β_2 to 0, if β_1 is overestimated.

Limiting distributions

$\lambda = 0.5$

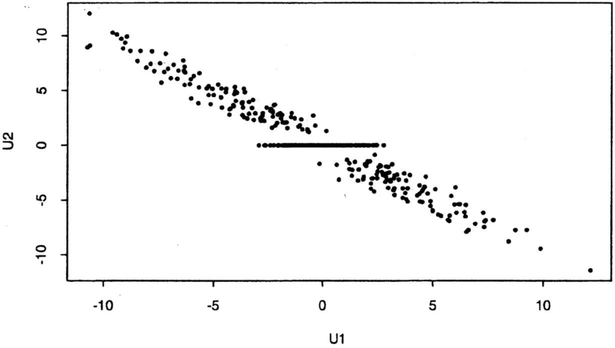

FIG. 3. *Sample of* 500 *from the limiting distribution of the Bridge estimator in Example* 1 *with* $\gamma = 1/2$ *and* $\lambda_0 = 1/2$.

The shrinkage is more selective. And there's a gap in the distribution of $\hat{U}_2$ ("no man's land")

Local asymptotics and small parameters

They further consider the performance for finite sample. They show how the "exact 0" phenomenon occur in finite samples, when the true parameter is small but nonzero ($\gamma \leq 1$). The statement & proofs are similar.

THEOREM 4. *Assume the model* (11) *with* $\boldsymbol{\beta}_n = \boldsymbol{\beta} + t/\sqrt{n}$ *and assume that* (12) *and* (13) *are satisfied. Let* $\hat{\boldsymbol{\beta}}_n$ *minimize* (14) *for some* $\gamma > 1$.

(a) *If* $\lambda_n/\sqrt{n} \to \lambda_0 \geq 0$ *then*

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_n) \to_d \operatorname{argmin}(V),$$

where

$$V(\mathbf{u}) = -2\mathbf{u}^T\mathbf{W} + \mathbf{u}^T C\mathbf{u} + \lambda_0 \sum_{j=1}^{p} u_j \operatorname{sgn}(\beta_j)|\beta_j|^{\gamma-1}.$$

(b) *If* $\boldsymbol{\beta} = \mathbf{0}$ *and* $\lambda_n/n^{\gamma/2} \to \lambda_0 \geq 0$ *then*

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_n) \to_d \operatorname{argmin}(V),$$

where

$$V(\mathbf{u}) = -2\mathbf{u}^T\mathbf{W} + \mathbf{u}^T C\mathbf{u} + \lambda_0 \sum_{j=1}^{p} |u_j + t_j|^{\gamma}.$$

Local asymptotics and small parameters

THEOREM 5. *Assume the model* (11) *with* $\boldsymbol{\beta}_n = \boldsymbol{\beta} + \mathbf{t}/\sqrt{n}$ *and assume that* (12) *and* (13) *are satisfied. Suppose that* $\hat{\boldsymbol{\beta}}_n$ *minimizes* (14) *for* $\gamma \le 1$ *where* $\lambda_n / n^{\gamma/2} \to \lambda_0 \ge 0$.

(a) *For* $\gamma = 1$,

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_n) \to_d \operatorname{argmin}(V)$$

where

$$V(\mathbf{u}) = -2\mathbf{u}^T \mathbf{W} + \mathbf{u}^T C \mathbf{u} + \lambda_0 \sum_{j=1}^{p} [u_j \operatorname{sgn}(\beta_j) I(\beta_j \ne 0) + |u_j + t_j| I(\beta_j = 0)]$$

(b) *For* $\gamma < 1$,

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_n) \to_d \operatorname{argmin}(V),$$

where

$$V(\mathbf{u}) = -2\mathbf{u}^T \mathbf{W} + \mathbf{u}^T C \mathbf{u} + \lambda_0 \sum_{j=1}^{p} |u_j + t_j|^{\gamma} I(\beta_j = 0).$$

Bootstrapping

Estimating the standard error to bridge parameter estimates is nontrivial, especially when $\gamma \leq 1$. One natural idea is to use bootstrapping. But the

asymptotic results show that the bootstrap may introduce bias that doesn't vanish asymptotically, when

- $\gamma < 1$
- some true parameters are either exactly 0 or close to 0.

Nearly singular design

$$C_n = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i^T \boldsymbol{x}_i \to C$$

In the last section, they consider the nearly singular design. That is, C_n is not singular, but C is singular.

Nearly singular design

THEOREM 6. *Assume a nearly singular model with C_n satisfying* (15). *Let* $\mathbf{W}$ *be a zero mean multivariate Normal random vector such that* $\operatorname{Var}(\mathbf{u}^T\mathbf{W}) = \mathbf{u}^T D\mathbf{u} > 0$ *for each nonzero* $\mathbf{u}$ *satisfying* $C\mathbf{u} = \mathbf{0}$.

(a) *If* $\gamma > 1$ *and* $\lambda_n/b_n \to \lambda_0 \geq 0$, *then*

$$b_n(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \to_d \operatorname{argmin}\{V(\mathbf{u}): C\mathbf{u} = \mathbf{0}\},$$

where

$$V(\mathbf{u}) = -2\mathbf{u}^T\mathbf{W} + \mathbf{u}^T D\mathbf{u} + \lambda_0 \sum_{j=1}^{p} u_j \operatorname{sgn}(\beta_j)|\beta_j|^{\gamma-1}.$$

(b) *If* $\gamma = 1$ *and* $\lambda_n/b_n \to \lambda_0 \geq 0$, *then*

$$b_n(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \to_d \operatorname{argmin}\{V(\mathbf{u}): C\mathbf{u} = \mathbf{0}\},$$

where

$$V(\mathbf{u}) = -2\mathbf{u}^T\mathbf{W} + \mathbf{u}^T D\mathbf{u} + \lambda_0 \sum_{j=1}^{p}[u_j \operatorname{sgn}(\beta_j)\mathbf{I}(\beta \neq 0) + |u_j|\mathbf{I}(\beta_j = 0)].$$

(c) *If* $\gamma < 1$ *and* $\lambda_n/b_n^\gamma \to \lambda_0 \geq 0$ *then*

$$b_n(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \to_d \operatorname{argmin}\{V(\mathbf{u}): C\mathbf{u} = \mathbf{0}\},$$

where

$$V(\mathbf{u}) = -2\mathbf{u}^T\mathbf{W} + \mathbf{u}^T D\mathbf{u} + \lambda_0 \sum_{j=1}^{p} |u_j|^\gamma I(\beta_j = 0).$$

Nearly singular design

One example for nearly singular design...

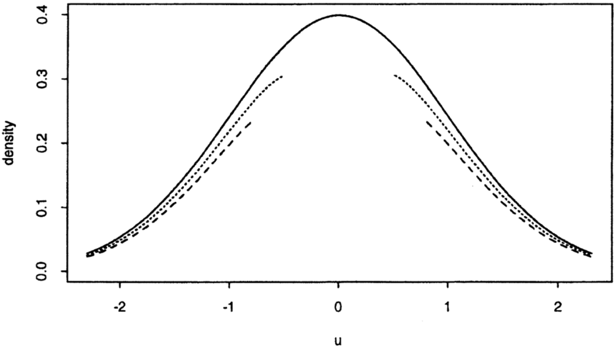

FIG. 4. *Densities for* $\lambda = 0$ *(solid line),* $\lambda = 0.5$ *(dotted line) and* $\lambda = 1$ *(dashed line); for* $\lambda = 0.5$ *and* $\lambda = 1$*; these are the densities of the absolute continuous part of the distribution as the distribution in these cases has positive probability mass at* 0.

$P(\hat{U} = 0) = 0, 0.448, 0.655$ for $\gamma = 0, 0.5, 1$