Bootstapping Lasso Estimators

Alberto Quaini

June 2017

Presentation

- *Presentation of:*
  "Bootstrapping Lasso Estimator" – A. Chatterjee, S. N. Lahiri [2011], JASA.
- *for PEF UNISG course:*
  "Resampling methods and forecasting" – L. Camponovo
- *Additional literature:*
  1
  2

Outline

Introduction

Linear regression model with iid errors:

$$y_i = x_i^T \beta + \epsilon_i, \qquad i = 1, ..., n \tag{1}$$

Lasso estimator:

$$\hat{\beta}_n = argmin_{u \in R^p} \sum_{i=1}^{n} (y_i - x_i^T u)^2 + \lambda_n \sum_{j=1}^{p} |u_j| \tag{2}$$

- estimation and variable selection method (Tibshirani [1996])
- computationally feasible (Friedman et Al. [2007])
- model consistency (Wainwright [2006], Zhao and Yu [2006] and Zou [2006])
- estimation consistency (Knight and Fu [2000])

Introduction

Problems:

1. Consistency
   - Knight and Fu [2000] show that the limiting distribution of the Lasso estimator is complicated
   - in practice alternative approximations are needed to carry on inference for the Lasso
   - The two autors consider the residual-based Bootstrap method
   - Chatterjee and Lahiri [2010] show that the Bootstrapped Lasso estimator is inconsistent whenever at least one component of the parameter vector is zero
2. Confidence intervals and testing
   - proposals of Tibshirani [1996] and Osborne et Al. [2000] have the drawback of considering the Lasso an approximately linear transformation
   - proposals of Tibshirani [1996], Fan and Li [2001] and Fan and Peng [2004] only provide CI for underlying non-zero parameters

Introduction

Results and proposals in Chatterjee and Lahiri [2011]:

1. Consistency
   - construct a suitable modification to the residual-based Bootstrap
   - show consistency under mild regularity conditions even when some of the underlying parameters are zero
2. Confidence interval and testing
   - the modified Bootstrap method provides consistent estimate of the variance of the Lasso estimator for both zero and non-zero parameter components

Introduction

3. choice of the regularization parameter λ_n
   - accuracy of the lasso critically depends on the regularization parameter
   - the modified Bootstrap is consistent for the MSE of the Lasso
   - the modified Bootstrap estimator of the MSE can be used for the choice of λ_n
4. Adaptive Lasso estimator (Zou [2006])
   - adaptive weights are used for penalizing different coefficients in the L_1 penalty
   - it enjoys the oracle property, i.e. performs as well as if the true underlying model were given in advance
   - the authors show that the simple residual Bootstrap can consistently estimate the distribution of the adaptive Lasso estimator

The Modified Bootstrap method

background and motivation

The **residual Bootstrap** method (standard in linear regression setting with nonrandom x_i, see Efron [1979], Freedman [1981]) proceeds as follows in the context of the Lasso (Knight and Fu [2000]):

1. Consider the set of centered residuals $E = \{e_i = \tilde{e}_i - \bar{e}, \text{for } i = 1, ..., n\}$, where $\bar{e} = n^{-1} \sum_i \tilde{e}_i$ and $\tilde{e}_i$'s are the residuals of the Lasso fit on the original sample.
2. Construct B bootstrap samples of size n selecting with replacement form E: $E^*_b = \{e^*_{i,b} : i = 1, ..., n\}$ and compute $y^*_{i,b} = {x_i}^T \hat{\beta}_n + e^*_{i,b}$, for $i = 1, ..., n,$ and $b = 1, ..., B$, where $\hat{\beta}_n$ is the Lasso estimator for the original sample.

The Modified Bootstrap method

3. Compute the bootstrap version of $T_n = n^{1/2}(\hat{\beta}_n - \beta)$, i.e. $T_n^* = n^{1/2}(\hat{\beta}^*_{n,b} - \hat{\beta}_n)$, where $\hat{\beta}^*_{n,b}$ is the Lasso estimator for bootstrap sample b.
4. The residual Bootstrap estimator of the distribution G_n of T_n is $\hat{G}_n(B) = P_*(T_n^* \in B)$, where $B \in \mathscr{B}(R^p)$ and P_* is the probability of T_n^* given errors ϵ_i's.

Chatterjee and Lahiri [2010] show that:

- the estimators of the zero parameters fail to capture the target sign value, which is zero
- because of that, $\hat{G}n$, instead of converging to the deterministic limit of Gn converges weakly to a random probability measure
- i.e. it fails to provide a valid approximation to G_n

The Modified Bootstrap method

A Modified Bootstrap method

Objective: capture the signs of the parameters, expecially the zero components, with probability tending to 1, as the sample size n goes to infinity.

Idea: force components of the Lasso estimator $\hat{\beta}_n$ to be exactly zero whenever they are close to zero using the fact that the Lasso estimator is root-n consistent.

To this end:

1. Form a sequence $\{a_n\}$ of real numbers such that $a_n + (n^{-1/2} log(n))a_n^{-1} \to 0$ asymptotically.
2. Threshold the components of the Lasso estimator β_n at a_n, and define the modified Lasso estimator

$$\tilde{\beta}_{n,j} = \beta_{n,j} \mathbb{1}(\beta_{n,j} \geq a_n), \text{for } j = 1, ..., p. \tag{3}$$

The Modified Bootstrap method

Note that with probability tending to 1 (as $n \to \infty$):

- $|\hat{\beta}_{n,j}| = |\beta_j| + O(n^{-1/2}) > |\beta_j|/2 \geq a_n$, for a nonzero component β_j
- $|\hat{\beta}_{n,j}| = |\beta_j| + O(n^{-1/2}) = O(n^{-1/2}) \in [-a_n, a_n]$, for a zero component β_j

Then proceed as before:

3. Consider the set of centered residuals $R = \{r_i = \tilde{r}_i - \bar{r}, \text{for } i = 1, \ldots, n\}$, where $\bar{r} = n^{-1} \sum_i \tilde{r}_i$ and $\tilde{r}_i$'s are the residuals of the modified Lasso fit on the original sample.
4. Construct B bootstrap samples of size n selecting with replacement form R: $R_b^{**} = \{r_{i,b}^{**} : i = 1, \ldots, n\}$ and compute $y_{i,b}^{**} = x_i^T \tilde{\beta}_n + r_{i,b}^{**}$, for $i = 1, \ldots, n, \text{and } b = 1, \ldots, B$, where $\tilde{\beta}_n$ is the modified Lasso estimator for the original sample.

The Modified Bootstrap method

5. Compute the bootstrap version of $T_n = n^{1/2}(\hat{\beta}_n - \beta)$, i.e. $T_n^{**} = n^{1/2}(\hat{\beta}_{n,b}^{**} - \tilde{\beta}_n)$, where $\hat{\beta}_{n,b}^{**}$ is the Lasso estimator for bootstrap sample b.
6. The residual Bootstrap estimator of the distribution G_n of T_n is $\tilde{G}_n(B) = P_{**}(T_n^{**} \in B)$, where $B \in \mathscr{B}(R^p)$ and P_{**} is the probability of T_n^{**} given errors ϵ_i's.

Remarks:

- Centering the residuals ensures the Bootstrap analogue of the condition $E[e_i] = 0$
- A rescaling factor $(1 - p/n) - 1/2$ is sometimes used in the construction of the residuals (see Efron [1982]) to improve finite sample accuracy
- It is possible to replace $\hat{\beta}_n$ by any other $\sqrt{n}$-consistent estimator of β, e.g. least squares

Bootstrapping the Lasso estimator

Consistency and the distributional approximation

Theorem 1: Consistency of Modified Bootstrap

Assume:

- (C1) $n^{-1} \sum_i x_i x_i^T \to C$, p.d. matrix. Furthemore $n^{-1} \sum_i \|x_i\|^3 \to O(1)$.
- (C2) $\lambda_n n^{-1/2} \to \lambda_0 \geq 0$.
- (C3) errors ϵ_i's are iid with $E[\epsilon_i] = 0$ and $VAR[\epsilon_i] = \sigma^2 < \infty$.

Then:

$$\mathscr{P}(\tilde{G}_n, G_n) \to 0, \text{as } n \to \infty, \text{with probability 1},$$

where $\mathscr{P}(\cdot, \cdot)$ denotes the Prohorov probability metric.

Bootstrapping the Lasso estimator

Remarks:

- Chatterjee and Lahiri [2010] shows that under the same set of regularity assumptions, if β has at least one zero component and if $\hat{G}_n$ is the residual bootstrap estimate of G_n, then

$$\mathscr{P}(\hat{G}_n, G_n) \not\to 0, \text{in probability, as } n \to \infty$$

- Theorem 1 states strong consistency of the modified Bootstrap distribution estimator
- From Theorem 1 it follows that the modified bootstrap method can be used to approximate the distribution of the Lasso estimator T_n for any $\beta \in R^p$. Hence, it can be used to construct valid large sample confidence set estimators of β

Bootstrapping the Lasso estimator

Definitions:

- let $t(\alpha)$ denote the $\alpha \in (0,1)$ quantile of $\| T_\infty \|$, where T_∞ denotes the limiting random vector such that $T_n \to T_\infty$ and has distribution G_∞.
- let $\hat{t}_n(\alpha)$ denote the $\alpha \in (0,1)$ quantile of the bootstrap distribution of $\| T_n^{**} \|$. Then the set

$$I_{n,\alpha} \equiv \{ t \in R^p : \| t - \hat{\beta}_n \| \leq n^{-1/2} \hat{t}_n(\alpha) \}.$$

Bootstrapping the Lasso estimator

Corollary 1: Modified Bootstrap Confidence Interval

Assume (C1), (C2) and (C3) hold. Then:

i if $\alpha \in (0,1)$ is such that $P(\|T_\infty\| \leq t(\alpha) + \eta) > \alpha, \forall \eta > 0$, then for all $\beta \in R^p$:

$$P(\beta \in I_{n,\alpha}) \to \alpha, \qquad \text{as } n \to \infty \tag{4}$$

ii if there is at least one nonzero component of β, then (4) holds for all $\alpha \in (0,1)$.

- Corollary 1 justifies the use of the modified Bootstrap method to construct valid large sample confidence regions for β
- Corollary 1 can also be used to test the null hypothesis $H_0 : \beta_j = 0$ for all $j \in J$ for a given $J \subset \{1, ..., p\}$

Bootstrapping the Lasso estimator

Remarks:

- Leeb and Pötscher [2006, 2008] and Pötscher and Schneider [2009] show that it is impossible to consistently estimate the distribution function of the Lasso estimator in a uniform sense
- Problems arise expecially when some underlying nonzero parameters get close to zero as n gets large
- Theorem 1 provides a method to obtain a consistent estimator in case the underlying parameters are fixed
- Andrews and Guggenberger [2009] show that uniform consistency is not necessary for producing uniformly valid confidence intervals
- Corollary 1 asserts that the modified Bootstrap method can control the asymptotic size of confidence intervals, however it is not clear if the latters are uniformly valid in the parameter values

Bootstrapping the Lasso estimator

Bootstrap bias and variance estimation

Theorem 2: Bias and Variance Consistency

Assume (C1), (C2) and (C3) hold. Then with probability 1:

$$E_*[T_n^{**}] \to E[T_\infty] \qquad \text{and} \tag{5}$$

$$(VAR_*[T_n^{**}])_{p\times p} \to (VAR_*[T_\infty])_{p\times p} \tag{6}$$

- for $\lambda_0 \neq 0$ in assumption (C2), T_n may be asymptotically biased and standard MSE estimation methods are unreliable
- The modified Bootstrap method produces strongly consistent estimators of the asymptotic bias and variance of T_n
- hence, Theorem 2 allows to estimate the MSE of a Lasso estimate and quantify the associated uncertainty for all values of β

Bootsrapping the Adaptive Lasso estimator

The **adaptive Lasso** estimator (Zou [2006]):

$$\breve{\beta}_n = argmin_{u \in R^p} \sum_{i=1}^{n} (y_i - x_i^T u)^2 + \lambda_n \sum_{j=1}^{p} \frac{|u_j|}{|\bar{\beta}_{j,n}|^\gamma}, \qquad (7)$$

where $\bar{\beta}_n$ denote an initial consistent estimator of β (e.g. least squares), $\lambda_n \geq 0$ is the penalty and $\gamma > 0$.

Oracle property (Zou [2006]):

$$P(B_n = A) \rightarrow 1, \quad \text{as } n \rightarrow 1 \qquad (8)$$

$$\sqrt{n}(\breve{\beta}_n^{nz} - \beta^{nz}) \rightarrow N(0, \sigma^2 C_{nz}) \qquad (9)$$

where $A = \{j : \beta_j = 0\}$, $B_n = \{j : \breve{\beta}_{j,n} = 0\}$ and $\sigma^2 C_{nz}$ is the var-cov matrix between nonzero estimated and underlying parameters

Bootsrapping the Adaptive Lasso estimator

A residual Bootstrap method for Adaptive Lasso

The algorithm is similar to the residual Bootstrap described earlier with few adjustments:

- E becomes the set of centered residuals of the adaptive Lasso fit on the original sample
- for each bootstrap sample, the adaptive Lasso estimator becomes:

$$\breve{\beta}_n^{+} = argmin_{u \in R^p} \sum_{i=1}^{n} (y_i^{+} - x_i^{T} u)^2 + \lambda_n \sum_{j=1}^{p} \frac{|u_j|}{|\bar{\beta}_{j,n}^{+}|^{\gamma}}, \qquad (10)$$

where $y_i^{+} = x_i^{T} \breve{\beta}_n + \epsilon_i^{+}, i = 1, ..., n$, ϵ_i^{+}'s are bootstrapped from E and $\bar{\beta}_n^{+}$ is defined by replacing y_i's with y_i^{+}'s in the definition of $\bar{\beta}_n$

Bootsrapping the Adaptive Lasso estimator

Remarks:

- the penalty of the adaptive Lasso incorporates a built-in soft-thresholding for the zero parameters
- Hence, the Bootstrap procedure just described does not need an initial truncation as it is the case for the Lasso

Denote:

- $\breve{T}_n \equiv \sqrt{n}(\breve{\beta}_n - \beta)$ with distribution H_n, where $H_n(x) = P(\breve{T}_n \leq x), x \in R$
- $\breve{T}_n^+ \equiv \sqrt{n}(\breve{\beta}_n^+ - \breve{\beta}_n)$ with distribution H_n^+ conditional on the ϵ_i's

Bootsrapping the Adaptive Lasso estimator

Theorem 3:

Assume (C1), and (C3) hold and suppose

$$\frac{\lambda_n}{\sqrt{n}} \to 0 \quad \text{and} \quad \lambda_n n^{(\gamma-1)/2} \to \infty. \tag{11}$$

Then,

$$\mathscr{P}(\hat{H}_n, H_n) \xrightarrow{P} 0, \quad \text{as } n \to \infty, \tag{12}$$

Remarks:

- (12) can be used to construct valid confidence intervals for β
- a corollary of the form of corollary 1 can be formulated for the adaptive Lasso residual bootstrap method

Bootsrapping the Adaptive Lasso estimator

- estimation of the MSE of the adaptive Lasso estimator can be difficult
- adaptive Lasso residual Bootstrap method provides a consistent estimator of the MSE matrix of the scaled adaptie Lasso estimator $\breve{\beta}_n$ given by $MSE[\breve{T}_n] \equiv nE[(\breve{\beta}_n - \beta)(\breve{\beta}_n - \beta)^T]$

Corollary 3

Assume (C1) and (C2) hold, Then:

$$MSE_*(\breve{T}_n^+) - MSE(\breve{T}_n) \xrightarrow{P} 0, \quad \text{as } n \to n. \tag{13}$$

Data-based choice of the regularization parameter

The optimal regularization parameter

Remarks:

- it can be shown that the distribution of T_n depends on λ_n only through λ_0
- note that $MSE(\hat{\beta}_n)$ can be expressed as $n^{-1}E\|T_n\|^2$ and that $nMSE(\hat{\beta}_n)$ converges to the MSE of the limiting random variable T_∞
- The effect of the penalization by λ_n on the overall accuracy of $\hat{\beta}_n$ is reflected by its MSE
- for what comes next, consider the natural reparametrization $\lambda_n = \lambda_0 n^{1/2}, \lambda_0 \in [0, \infty)$

Data-based choice of the regularization parameter

Then define the **optimal penalization** parameter as

$$\lambda_0^{opt} \equiv \phi(\lambda_0) \tag{14}$$

where $\phi(\lambda_0) = E\| T_\infty \|^2$.

- Thus, choosing $\lambda_0 = \lambda_0^{opt}$ yelds a Lasso estimator that minimizes the *MSE* in large samples.