Paper Review: Semi-parametric efficiency bounds and efficient estimation for high-dimensional models

Sara van de Geer, Jana Janková

ETH Zürich

January 26, 2016

Overview

1. Problem Setup
2. Preliminary Results
3. Main Results
   - Lower bounds for the linear model
   - An asymptotically efficient estimator in the linear model
4. Le Cam's Lemma

Problem Setup

Consider the linear model:

$$Y = X\beta_0 + \varepsilon, p > n, \tag{1}$$

where $X \in \mathbb{R}^{n \times p}$, $Y \in \mathbb{R}^n$, $\varepsilon \in \mathbb{R}^n$ with $\mathbb{E}\varepsilon_i = 0$, and ε_i's independent.
Questions:

1. statistical inference of β_0, e.g. confidence intervals, hypothesis tests? "de-sparsifying" or "de-biasing", van de Geer (2014); Zhang and Zhang (2014)
2. optimality properties of these de-biased estimators? lower bounds on the variance.

De-sparsifying Lasso Estimator

Lasso Estimator:

$$\widehat{\beta} = \text{argmin}_{\beta \in \mathbb{R}^p} \left\{ \|Y - X\beta\|_2^2/n + \lambda \|\beta\|_1 \right\} \tag{2}$$

1. $\widehat{\beta}$ does not have a tractable limiting distribution since it introduces bias by shrinking all coefficients towards zero.
2. De-sparsified estimator $\widehat{b}$: uses $\widehat{\beta}$ as an initial estimator and implements a bias correction step.

Goal

1. show that $\widehat{b}$ is **asymptotically unbiased**, and achieve the **lower bound**,

   • i.e. the de-biased estimator is the best among all asymptotically unbiased estimators: thus in this sense asymptotically efficient.

2. show that the de-sparsified estimator converges **locally uniformly** to the limiting normal distribution with zero mean and the smallest possible variance.

Strong oracle inequalities for the Lasso I

Theorem

Assume the linear model in (1) with $\epsilon \sim \mathcal{N}(0, \sigma_\epsilon^2 I)$, where $\sigma_\epsilon^2 = \mathcal{O}(1)$. Suppose that $X_i \sim \mathcal{N}(0, \Sigma_0)$ are independent for $i = 1, \ldots, n$, where $\|\Sigma_0\|_\infty = \mathcal{O}(1)$ and $\Lambda_{\min}(\Sigma_0) \geq L > 0$ for a universal constant L. Suppose that $\|\beta_0\|_2 = \mathcal{O}(1)$, $\|\beta_0\|_0 \leq s$ and $s \log p/n = o(1)$. Let $k \in \{1, 2, \ldots\}$ be fixed and let $\tau > 0$ fixed be such that $p^{-\tau/2} = \mathcal{O}((s\lambda^2)^{k/2})$. Consider the Lasso $\hat\beta$ defined in (2) with tuning parameter $\lambda \geq c\tau\sqrt{\log p/n}$, where c is a sufficiently large universal constant. Then there exist universal constants C_1, C_2 such that

$$(\mathbb{E}\|\hat\beta - \beta_0\|_1^k)^{1/k} \leq C_1 s\lambda.$$

Moreover, for any $\nu > 0$ it holds with probability at least $1 - 1/\nu^k$

$$\|\hat\beta - \beta_0\|_1 \leq \nu C_1 s\lambda.$$

Strong oracle inequalities for the Lasso II

- Taking $k = 1$, under the conditions of Theorem 1 we obtain

$$\mathbb{E}\|\hat{\beta} - \beta_0\|_1 \leq C_1 s\lambda.$$

Local uniform asymptotic unbiasedness I

Definition

Let $a \in \mathbb{R}^p$ and let $0 < \delta_n \downarrow 0$. We call T_n a **strongly asymptotically unbiased** estimator of $g(\theta_0)$ at θ_0 **in the direction** a with rate δ_n if for $m_n := n/\delta_n$ and for $\theta := \theta_0 + a/\sqrt{m_n}$ and for $\theta := \theta_0$ it holds that

$$\sqrt{m_n}(\mathbb{E}_\theta T_n - g(\theta)) = o(1).$$

Definition

We say that T_n is **strongly asymptotically unbiased** for estimation of $g(\theta)$ if for all $\theta \in \Theta$ and $a \in \Theta$ it holds that

$$\sqrt{n}\left(\mathbb{E}_{\theta+a/\sqrt{n}} T_n - g\left(\theta + \frac{a}{\sqrt{n}}\right)\right) = o(1).$$

Local uniform asymptotic unbiasedness II

- The first definition assumes unbiasedness only along a particular direction
- In particular, we consider shrinking neighbourhoods of θ_0 of size $1/\sqrt{n}$, where we require the bias to vanish at a rate $1/\sqrt{n}$. Note that if $\sqrt{n}(\mathbb{E}_\theta(T_n) - g(\theta)) = o(1)$, then one may take e.g. $\delta_n := \sqrt{n}(\mathbb{E}_\theta(T_n) - g(\theta))$.
- It is particularly useful when recognizing the concept of a worst possible sub-direction
- The second definition assumes unbiasedness in every direction within the considered sparse model.

Lower bounds for the linear model

Assume that X is a random $n \times p$ matrix independent of ϵ with independent rows $X_i \sim \mathcal{N}(0, \Sigma_0)$ for $i = 1, \ldots, n$. We assume the inverse covariance matrix $\Theta_0 := \Sigma_0^{-1}$ exists.

Theorem

Let $a \in \mathbb{R}^p$ be such that $a^T \Sigma_0 a = 1$. Suppose that T_n is a strongly asymptotically unbiased estimator of $g(\beta_0)$ at β_0 in the direction a with rate δ_n. Assume moreover that for some $\dot{g}(\beta_0) \in \mathbb{R}^p$ and for $m_n = n/\delta_n$

$$\sqrt{m_n}\,(g(\beta_0 + a/\sqrt{m_n}) - g(\beta_0)) = a^T \dot{g}(\beta_0) + o(1). \tag{3}$$

Then

$$n\mathrm{var}(T_n) \geq [a^T \dot{g}(\beta_0)]^2 - o(1).$$

Corollary

The lower bound $[a^T \dot{g}(\beta_0)]^2$ *is maximized at the value*

$$a_0 := \Theta_0 \dot{g}(\beta_0) / \sqrt{\dot{g}(\beta_0)^T \Theta_0 \dot{g}(\beta_0)}.$$

Hence under the conditions of Theorem 4, we get

$$nvar(T_n) \geq \dot{g}(\beta_0)^T \Theta_0 \dot{g}(\beta_0) - o(1).$$

Definition

Let g be differentiable at β_0 with derivative $\dot{g}(\beta_0)$. We call

$$c_0 := \Theta_0 \dot{g}(\beta_0) / \dot{g}(\beta_0)^T \Theta_0 \dot{g}(\beta_0)$$

the worst possible sub-direction for estimating $g(\beta_0)$.

Assume that T_n is strongly asymptotically unbiased in all directions $a \in \mathcal{B}$, where $\mathcal{B} := \{\beta \in \mathbb{R}^p : \|\beta\|_0 \leq s, \|\beta\|_2 = \mathcal{O}(1)\}$.

Corollary

Let T_n be a strongly asymptotically unbiased estimator of $g(\beta_0)$, and for all $\beta_0 \in \mathcal{B}, a \in \mathcal{B}$ it holds

$$\sqrt{n}\left(g(\beta_0 + a/\sqrt{n}) - g(\beta_0)\right) = a^T \dot{g}(\beta_0) + o(1).$$

Suppose that $\Theta_0 \dot{g}(\beta_0)/\sqrt{\dot{g}(\beta_0)^T \Theta_0 \dot{g}(\beta_0)} \in \mathcal{B}$ for all $\beta_0 \in \mathcal{B}$ and suppose that $\Lambda_{\max}(\Sigma_0) = \mathcal{O}(1)$. Then it holds

$$n\mathrm{var}_{\beta_0}(T_n) \geq \dot{g}(\beta_0)^T \Theta_0 \dot{g}(\beta_0) - o(1).$$

Construction of the de-biased Lasso estimator I

- Consider X_i are iid rows with mean zero and covariance matrix Σ_0. Assume the inverse covariance matrix $\Theta_0 := \Sigma_0^{-1}$ exists.
- Consider the Lasso defined in (2) with $\lambda \asymp \sqrt{\log p/n}$.
- Let $\hat{\Theta}_j$ be an estimate of Θ_j^0 be obtained *nodewise regression*.
- Denote by X_{-j} the $n \times (p-1)$ matrix obtained by removing the j-th column from X.

For $j = 1, \ldots, p$, let

$$\hat{\gamma}_j := \arg \min_{\gamma \in \mathbb{R}^{p-1}} \|X_j - X_{-j}\gamma\|_2^2/n + 2\lambda_j \|\gamma\|_1, \tag{4}$$

$$\hat{\tau}_j^2 := \|X_j - X_{-j}\hat{\gamma}_j\|_2^2/n,$$

$$\hat{\Theta}_{Lasso,j} := (-\hat{\gamma}_{j,1}, \ldots, -\hat{\gamma}_{j,j-1}, 1, -\hat{\gamma}_{j,j+1}, \ldots, -\hat{\gamma}_{j,p})/\hat{\tau}_j^2, \tag{5}$$

Construction of the de-biased Lasso estimator II

where $\lambda_j \asymp \lambda \asymp \sqrt{\log p/n}$ for $j = 1, \ldots, p$. The necessary Karush-Kuhn-Tucker conditions corresponding to the nodewise regression (obtained by replacing derivatives by sub-differentials) imply the condition $\|\hat{\Sigma}\hat{\Theta}_j - e_j\|_\infty = O_P(\lambda)$.
Define the **de-biased** Lasso introduced in van de Geer (2014) by

$$\hat{b} := \hat{\beta} + \hat{\Theta}^T X^T (Y - X\hat{\beta})/n, \tag{6}$$

Estimation of β_j^0

- $\mathcal{B} := \{\beta \in \mathbb{R}^p : \|\beta_0\|_0 = \mathcal{O}(s), \|\beta_0\|_2 = \mathcal{O}(1)\}$
- Condition (A1) $1/\Lambda_{\min}(\Sigma_0) = \mathcal{O}(1)$ and $\Lambda_{\max}(\Sigma_0) = \mathcal{O}(1)$.

Lemma

Suppose that condition (A1) is satisfied and suppose that $s \log p/n = o(1)$. *Let* $\hat{\beta}$ *be the Lasso estimator defined in* (2) *with a sufficiently large tuning parameter of order* $\sqrt{\log p/n}$. *Then for every* $\beta_0 \in \mathcal{B}$

$$\mathbb{E}_{\beta_0} \|\hat{\beta} - \beta_0\|_1 = \mathcal{O}(s\lambda).$$

Strongly asymptotic unbiasedness of $\widehat{b}_j$

Lemma

Suppose that condition (A1) is satisfied and suppose that $s = o\left(\frac{\sqrt{n}}{\log p}\right)$.
Let $\hat{b}_j$ *be defined as in* (6) *with* $\hat{\Theta}_j$ *satisfying* $\|\hat{\Sigma}\hat{\Theta}_j - e_j\|_\infty \leq \lambda_j$. *Then for every* $\beta_0 \in \mathcal{B}$

$$\sqrt{n}\mathbb{E}_{\beta_0}(\hat{b}_j - \beta_j^0) = o(1).$$

Theorem

Suppose that condition (A1) is satisfied, $s = o\left(\frac{\sqrt{n}}{\log p}\right)$ *and that* $\|\Theta_j^0\|_0 = \mathcal{O}(s)$*. Let* $\hat{\Theta}_{Lasso,j}$ *be obtained using the nodewise regression as in* (5)*. Then* $\hat{b}_j$ *defined in* (6) *using the nodewise regression is strongly asymptotically unbiased and for any strongly asymptotically unbiased estimator* T *of* β_j^0 *it holds for all* $\beta_0 \in \mathcal{B}$

$$var(T) \geq var(\hat{b}_j) = \frac{\Theta_{jj}^0 + o(1)}{n}.$$

Locally uniform convergence I
Motivation

Classical examples of superefficiency: Hodges' Estimator.
$P_\theta = \{N(\theta, 1), \theta \in \Theta\}$
Let $T_n = \bar{X}_n$, where $X_1, \cdots, X_n$ iid $N(\theta, 1)$.

$$S_n = \begin{cases} T_n & \text{if } |T_n| \geq n^{-1/4} \\ 0 & \text{if } |T_n| < n^{-1/4}. \end{cases} \tag{7}$$

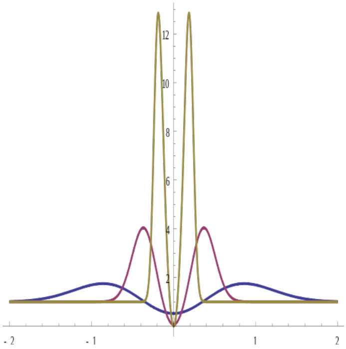
Asymptotics for T_n:

$$\sqrt{n}(T_n - \theta) \xrightarrow{d} N(\theta, 1)$$

Asymptotics for S_n:

$$\begin{cases} \sqrt{n}(S_n - \theta) \xrightarrow{d} N(\theta, 1), \text{if } \theta \neq 0 \\ r_n(S_n - \theta) \xrightarrow{d} 0, \text{if } \theta = 0 \end{cases}$$

for any sequence r_n including $r_n = \sqrt{n}$.
S_n is said to be "superefficient" at $\theta = 0$. However, consider the $\mathbb{E}(S_n - \theta)^2$,



Locally uniform convergence III

Motivation

Pointwise convergence is insufficient for asymptotic efficiency and that we in fact need uniform convergence on shrinking neighbourhoods.

1. Consider the model $\mathcal{P} := \{P_\theta : \theta \in \Theta\}$, where

$$\Theta := \{\theta \in \mathbb{R}^p : \|\theta\|_0 \leq s, \|\theta\|_2 = \mathcal{O}(1)\}.$$

2. The de-sparsified estimator T_n is asymptotically linear:

$$T_n - g(\theta) = \frac{1}{n}\sum_{i=1}^{n} l_\theta(X_i) + o_P(n^{-1/2}),$$

where $\mathbb{E}_\theta l_\theta = 0$ and $\mathbb{E} l_\theta^2 < \infty$.

Locally uniform convergence IV

Motivation

For asymptotically linear estimators, one has the asymptotic variance $V_\theta := \mathbb{E} l_\theta^2$. Consider the following condition for every $h \in \Theta$

$$P_\theta(l_\theta h^T s_\theta) - h^T \dot{g}(\beta) = 0.$$

If the condition is satisfied, then the Cauchy-Schwarz inequality implies

$$(h^T \dot{g}(\theta))^2 = (P_\theta l_\theta h^T s_\theta)^2 \leq \mathrm{var}(l_\theta)\mathrm{var}(h^T s_\theta) = V_\theta h^T I_\theta h.$$

Hence

$$V_\theta \geq \max_{h \in \Theta} (h^T \dot{g}(\theta))^2 / h^T I_\theta h. \tag{8}$$

Assuming that $I_\theta^{-1} \dot{g}(\theta) \in \Theta$, the right-hand side of (8) is maximized at $I_\theta^{-1} \dot{g}(\theta)$. Hence we obtain the following lower bound on the asymptotic variance

$$V_\theta \geq \dot{g}(\theta)^T I_\theta^{-1} \dot{g}(\theta).$$

Locally uniform convergence V

Motivation

1. Under the conditions of the central limit theorem, asymptotic linearity implies that

$$\sqrt{n}(T_n - g(\theta))/V_\theta^{1/2} \overset{\theta}{\rightsquigarrow} \mathcal{N}(0,1)$$

for every θ.

2. For every $h \in \Theta$ and every $\theta \in \Theta$ it holds that

$$\frac{\sqrt{n}(T_n - g(\theta + h/\sqrt{n}))}{V_\theta^{1/2}} \overset{\theta + h/\sqrt{n}}{\rightsquigarrow} \mathcal{N}(0,1).$$

Theorem I

Assume

- Let $g : \mathbb{R}^p \to \mathbb{R}$ satisfy

$$\sqrt{n}(g(\theta + h/\sqrt{n}) - g(\theta)) = h^T \dot{g}(\theta) + o(1).$$

- Suppose that for all $\theta \in \Theta$

$$T_n - g(\theta) = \frac{1}{n} \sum_{i=1}^{n} l_\theta(X_i) + o_{P_\theta}(n^{-1/2}),$$

where $P_\theta l_\theta = 0$ and $V_\theta := P_\theta l_\theta^2 < \infty$.

Theorem II

- Suppose that $V_\theta = \mathcal{O}(1)$ and $1/V_\theta = \mathcal{O}(1)$. Let s_θ be the score function, let $I_\theta := \mathbb{E}s_\theta s_\theta^T$ and assume that

$$\|\frac{1}{n}\sum_{i=1}^n \dot{s}_\theta(X_i) + I_\theta\|_\infty = \mathcal{O}_P(\lambda),$$

  where λ is such that $s\lambda = o(1)$.
- Assume further that $\Lambda_{\max}(I_\theta) = \mathcal{O}(1)$.

Theorem

Then for every $h \in \Theta$ it holds that

$$\frac{\sqrt{n}(T_n - g(\theta + h/\sqrt{n})) - (P_\theta(I_\theta h^T s_\theta) - h^T \dot{g}(\theta))}{V_\theta^{1/2}} \overset{\theta + h/\sqrt{n}}{\rightsquigarrow} \mathcal{N}(0,1)$$

Sara van de Geer. On the uniform convergence of empirical norms and inner products, with application to causal inference. *Electron. J. Statist.*, 8(1):543–574, 2014. doi: 10.1214/14-EJS894.

Cun-Hui Zhang and Stephanie S. Zhang. Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1): 217–242, 2014. ISSN 1467-9868. doi: 10.1111/rssb.12026.

Thank you!