Lasso Guarantees for High Time Dimensional Time Series Estimation under Mixing Conditions

Ambuj Tewari

Department of Statistics, and
Department of EECS,
University of Michigan, Ann Arbor

July 10, 2017

(based on joint work with Kam Chung Wong and Zifan Li)

Outline

Outline

Setup

- Consider a stochastic process of pairs $(X_t, Y_t)_{t=1}^{\infty}$ where $X_t \in \mathbb{R}^p$, $Y_t \in \mathbb{R}^q$
- We will be interested in time series prediction
- For a time series $(Z_t)_{t=1}^{\infty}$, we might be interested in predicting $Y_t = Z_t$ using $X_t = (Z_{t-d}, \ldots, Z_{t-1})$
- Cannot assume that the pairs (X_t, Y_t) are iid

Lasso

- Assume sequence $(X_t, Y_t)_{t=1}^T$ is strictly stationary and centered
- Best linear predictor of Y_t in terms of X_t

$$\Theta^\star = \arg\min_{\Theta \in \mathbb{R}^{p\times q}} \mathbb{E}[\| Y_t - \Theta' X_t\|_2^2].$$

- Collect the X_ts and Y_ts together in two matices:

$$\mathbf{Y} = (Y_1, Y_2, \dots, Y_T)' \in \mathbb{R}^{T\times q}$$

$$\mathbf{X} = (X_1, X_2, \dots, X_T)' \in \mathbb{R}^{T\times p}$$

- Lasso estimator $\widehat{\Theta} \in \mathbb{R}^{p\times q}$

$$\widehat{\Theta} = \arg\min_{\Theta \in \mathbb{R}^{p\times q}} \frac{1}{T} \| \mathrm{vec}(\mathbf{Y} - \mathbf{X}\Theta)\|_2^2 + \lambda_T \|\mathrm{vec}(\Theta)\|_1$$

Master Theorem - Informal

- (Lower) Restricted Eigenvalue (RE) condition: The empirical covariance matrix $\mathbf{X}'\mathbf{X}/T$ has "curvature" in a restricted set of directions
- Deviation Bound (DB) condition: The correlation between "noise" $\mathbf{W}$ and predictors $\mathbf{X}$ is small

$$\mathbf{W} = \mathbf{Y} - \mathbf{X}\Theta^\star$$

- Lasso Master Theorem: Sparsity assumption on $\Theta^\star$ + RE + DB implies bounds for Lasso

RE and DB Conditions

Lower Restricted Eigenvalue

$\Gamma \in \mathbb{R}^{p \times p}$ satisfies a lower RE with curvature $\alpha > 0$ and tolerance $\tau(T, p) > 0$ if

$$\forall v \in \mathbb{R}^p, \;\; v'\Gamma v \geq \alpha \, \|v\|_2^2 - \tau(T, p) \, \|v\|_1^2 \, .$$

Deviation Bound

$\mathbf{X}'\mathbf{W}$ satisfies the DB condition if there exists a deterministic multiplier function $\mathbb{Q}(\mathbf{X}, \mathbf{W}, \Theta^\star)$ and a rate of decay function $\mathbb{R}(p, q, T)$ such that,

$$\frac{1}{T} |||\mathbf{X}'\mathbf{W}|||_\infty \leq \mathbb{Q}(\mathbf{X}, \mathbf{W}, \Theta^\star)\mathbb{R}(p, q, T).$$

Master Theorem - Formal

Theorem (Lasso Estimation and Prediction Errors)

Suppose

1. $\Theta^\star$ *is s-sparse*
2. $\hat{\Gamma} := \mathbf{X}'\mathbf{X}/T$ *satisfies lower RE*(α, τ) *with* $\alpha \geq 32 s\tau$
3. $\mathbf{X}'\mathbf{W}$ *satisfies DB*

Then, for any $\lambda_T \geq 4\mathbb{Q}(\mathbf{X}, \mathbf{W}, \Theta^\star)\mathbb{R}(p, q, T)$,

$$\left\| \widehat{\Theta} - \Theta^\star \right\|_F \leq 4\sqrt{s}\lambda_T/\alpha,$$

$$\left|\left|\left| (\widehat{\Theta} - \Theta^\star)'\hat{\Gamma}(\widehat{\Theta} - \Theta^\star) \right|\right|\right|_F^2 \leq \frac{32\lambda_T^2 s}{\alpha}$$

Outline

RE via Concentration

- Consider a fixed vector $v \in \mathbb{R}^p$ and let $\Sigma_X = \mathbb{E}[X_t X_t^T]$
- Use concentration inequality to show

$$\frac{v'\mathbf{X}'\mathbf{X}v}{T} - v'\Sigma_X v = \frac{1}{T}\sum_{t=1}^{T}(X_t'v)^2 - \mathbb{E}[(X_t'v)^2]$$

  is sufficiently small
- Take union bound over sparse v

DB via Concentration

- Note that

$$\left|\left|\left|\mathbf{X}'\mathbf{W}\right|\right|\right|_\infty = \max_{1\le i\le p, 1\le j\le q} \left|[\mathbf{X}'\mathbf{W}]_{i,j}\right| = \max_{1\le i\le p, 1\le j\le q} \left|(\mathbf{X}_{:i})'\mathbf{W}_{:j}\right|$$

- At the population level, there is no correlation between $\mathbf{W}$ and $\mathbf{X}$

$$\mathbb{E}(\mathbf{X}_{:i})'(\mathbf{Y}-\mathbf{X}\Theta^\star) = 0, \forall i \ \Rightarrow \mathbb{E}(\mathbf{X}_{:i})'\mathbf{W}_{:j} = 0, \forall i, j$$

- Fix i, j and write

$$\begin{aligned}\left|(\mathbf{X}_{:i})'\mathbf{W}_{:j}\right| &= \left|(\mathbf{X}_{:i})'\mathbf{W}_{:j} - \mathbb{E}[(\mathbf{X}_{:i})'\mathbf{W}_{:j}]\right| \\ &\le \frac{1}{2}\left|\|\mathbf{X}_{:i}+\mathbf{W}_{:j}\|^2 - \mathbb{E}[\|\mathbf{X}_{:i}+\mathbf{W}_{:j}\|^2]\right| \\ &\quad + \frac{1}{2}\left|\|\mathbf{X}_{:i}\|^2 - \mathbb{E}[\|\mathbf{X}_{:i}\|^2]\right| + \frac{1}{2}\left|\|\mathbf{W}_{:j}\|^2 - \mathbb{E}[\|\mathbf{W}_{:j}\|^2]\right|\end{aligned}$$

Concentration for Subexponential, Independent Case

Theorem (Bernstein's Inequality)

Let $\xi_1, \cdots, \xi_T$ be independent centered sub-exponential random variables, and $K = \max_i \|\xi_i\|_{\psi_1}$. Then for every $a = (a_1, \cdots, a_T) \in \mathbb{R}^T$ and every $t \geq 0$, we have

$$\mathbb{P}\left\{\left|\sum_{i=1}^{T} a_i \xi_i\right| \geq t\right\} \leq 2\exp\left[-C_B \min\left(\frac{t^2}{K^2 \|a\|_2^2}, \frac{t}{K\|a\|_\infty}\right)\right]$$

where $C_B > 0$ is an absolute constant.

Results for Independent, Subgaussian Case

- Bernstein's concentration inequality will allow us to prove Lasso guarantees
- But it requires independence and subexponential tails
- This means independence and subgaussian tails for the original stochastic process
- Fact: A random variable is subgaussian iff its square is subexponential

Towards Handling Dependence and Heavy Tails

- Time series applications require the ability to deal with dependence as well as heavier tails
- We need ways to quantify dependence and heavy tailed behavior
- Then we need concentration inequalities that hold under weaker conditions
- Next, we quantify dependence using mixing coefficients
- Also, we quantify tail behavior using the notion of subweibull random variables

Outline

β-Mixing

- There are several notions of mixing: α-, β-, ρ-, ϕ-
- Let's focus on β-mixing
- Define the coefficient of dependence

$$\beta(X, X') = \|\mathbb{P}_X \otimes \mathbb{P}_{X'} - \mathbb{P}_{X,X'}\|_{TV}$$

- Given a stationary process X_t, define

$$\beta(\ell) = \beta(X_{-\infty,t}, X_{t+\ell,\infty})$$

- Geometrically beta mixing: Assume $\beta(\ell) \leq 2\exp(-c\ell^{\gamma_1})$

Subweibull Random Variables and Vectors

- We say a r.v. ξ is subweibull(γ_2) if there exists K s.t.

$$P(|\xi| \geq t) \leq 2 \exp(-(t/K)^{\gamma_2})$$

- subweibull(2) = subgaussian, subweibull(1) = subexponential
- For $\gamma_2 < 1$, subweibull r.v. is heavy tailed (m.g.f. doesn't exist)
- A random vector $\vec{\xi}$ is subweibull(γ_2) if $u'\vec{\xi}$ is subweibull(γ_2) for all unit vectors u (with a common K)

Subweibull Equivalent Definitions

Theorem (Wong and T., 2017)

Then the following statements are equivalent for every $\gamma_2 > 0$. The constants K_1, K_2, K_3 differ from each other at most a constant depending only on γ_2.

1. *The tails of ξ satisfies*

$$\mathbb{P}(|\xi| > t) \leq 2 \exp\left\{-(t/K_1)^{\gamma_2}\right\}, \ \forall t \geq 0.$$

2. *The moments of ξ satisfy,*

$$\|\xi\|_p := (\mathbb{E}|\xi|^p)^{1/p} \leq K_2 p^{1/\gamma_2}, \ \forall p \geq 1.$$

3. *The moment generating function of $|\xi|^{\gamma_2}$ is finite at some point; i.e., $\mathbb{E}\left[\exp\left(|\xi| / K_3\right)^{\gamma_2}\right] \leq 2$*

The Difficulty Landscape

- Suppose X_t and Y_t are geometrically β-mixing with exponent γ_1
- Also suppose they're both subweibull(γ_2)
- The pair $(\gamma_1, \gamma_2) \in \mathbb{R}_+$ quantifies the difficulty of the problem
- Easy regime: $\gamma_1 \to \infty$ (independence), $\gamma_2 \to \infty$ (a.s. bounded)
- Hard regime: $\gamma_1 \to 0$, $\gamma_2 \to 0$
- E.g., independent, subgaussian case corresponds to $\gamma_1 = \infty, \gamma_2 = 2$

Outline

How to Cover the Entire Landscape

- Case I: we first handle the subgaussian case with $\gamma_1 = 1$
- Case II: then we handle the case $1/\gamma_1 + 2/\gamma_2 > 1$
- Together, these two cases handle all γ_1, γ_2 pairs

$$\{(\gamma_1, \gamma_2) : \gamma_1 \geq 1, \gamma_2 \geq 2\} \cup \{(\gamma_1, \gamma_2) : 1/\gamma_1 + 2/\gamma_2 > 1\} = \mathbb{R}_+$$

- Concentration inequality for Case I: extension of Bernstein's inequality to β-mixing processes via blocking
- Concentration inequality for Case II: Merlevede, Peligrad, Rio (2011)

The Blocking Technique

- Create blocks from a given β-mixing process X_t

$$X_1, X_2, \ldots, X_B \qquad X_{B+1}, X_{B+2}, \ldots, X_{2B} \qquad \ldots$$

- Look at, say, even, blocks – they're separated by B time steps
- Yu's (1994) lemma allows us to create independent blocks

$$\tilde{X}_1, \tilde{X}_2, \ldots, \tilde{X}_B \qquad \tilde{X}_{B+1}, \tilde{X}_{B+2}, \ldots, \tilde{X}_{2B} \qquad \ldots$$

- At the same time, for any bounded h,

$$\mathbb{E}[h(\text{even blocks of } X)] \approx \mathbb{E}[h(\text{even blocks of } \tilde{X})]$$

Case I Result

- Case I: $\gamma_1 = 1, \gamma_2 = 2$
- Let $\epsilon > 0$. For $T \geq T_0(\epsilon)$, w.h.p.

$$\left\| \widehat{\Theta} - \Theta^{\star} \right\|_F \leq C \, \frac{K^2}{\lambda_{\min}(\Sigma_X)} \sqrt{\frac{s \log(pq)}{T^{1-\epsilon}}}$$

  for some universal constant C
- Rate "almost" $\sqrt{s \log(pq)/T}$
- However, $T_0(\epsilon)$ blows up as $\epsilon \to 0$
- K is the subgaussian constant of X_t, Y_t

Concentration Inequality for Case II

- Let $(\xi_i)_{i=1}^T$ be a stationary sequence of zero mean subweibull(γ_2) (with constant K) r.v.
- β-mixing coefficients $\beta(\ell) \leq 2\exp(-c\ell^{\gamma_1})$

Theorem (Wong, T., 2017 based on Merlevede et al., 2011)

Let $\frac{1}{\gamma} = \frac{1}{\gamma_1} + \frac{1}{\gamma_2}$. *Then for* $\gamma < 1$, $T > 4$, *and any* $t > 1/T$,

$$\mathbb{P}\left\{\left|\frac{\sum_{i=1}^T \xi_i}{T}\right| > t\right\} \leq T\exp\left\{-\frac{(tT)^\gamma}{K^\gamma C_1}\right\} + \exp\left\{-\frac{t^2 T}{K^2 C_2}\right\}$$

where the constants C_1, C_2 *depend only on* γ_1, γ_2 *and* c.

Case II Result

- Case II: $1/\gamma = 1/\gamma_1 + 2/\gamma_2$, $\gamma < 1$
- For $T \geq T_0(\gamma)$, w.h.p.

$$\left\| \widehat{\Theta} - \Theta^{\star} \right\|_F \leq C \, \frac{K^2}{\lambda_{\min}(\Sigma_X)} \sqrt{\frac{s \log(pq)}{T}}$$

  constant C that depends only on γ_1, γ_2, c
- Sample size threshold $T_0(\gamma)$ blows up as γ approaches 0 or 1
- K is the subweibull(γ_2) constant of X_t, Y_t

Summary

- Need to quantify dependence and tail behavior to extend Lasso results to time series
- We used β-mixing and subweibull exponents to do this
- Extended Lasso guarantees to cover the full range of possibilities for the 2 exponents
- Key ingredients are new concentration inequalities

Future Work

- Weaken β-mixing assumption (already have results for Gaussian processes under α-mixing)
- Weaken subweibull assumption to allow even heavier tails
- Discrete time series: hard to establish mixing conditions for these
- Lower bounds (and hopefully matching upper bounds)

Thank You!

References

- Kam Chung Wong, Zifan Li, Ambuj Tewari. Lasso Guarantees for Time Series Estimation Under Subgaussian Tails and β-Mixing. arXiv:1602.04265v3
  Case I results
- Kam Chung Wong, Ambuj Tewari. Lasso Guarantees for β-Mixing Heavy Tailed Time Series. to be soon uploaded to arXiv
  Case II results