10-755: Advanced Statistical Theory I **Fall 2017**

Lecture 14: October 18

Lecturer: Alessandro Rinaldo *Scribes: Ivan Stelmakh*

14.1 Persistence

14.1.1 Setup

In general linear regression model we observe the sequence of random variables

$$Z_1, \dots, Z_n \sim P$$

Here $Z_k = (X_k, Y_k) \in \mathbb{R}^{d+1}$ where $X_k \in R^d, Y_k \in \mathbb{R}$. The goal is to predict Y based on X for $(X, Y) \sim P$.

If we are interested in linear regression model, we want to compute β^* such that

$$\beta^* = \arg\min_{\beta \in \mathbb{R}^d} \left\{ \underbrace{\mathbb{E}\left[\left(Y - X^T\beta\right)^2\right]}_{R_P(\beta)} \right\}$$

Now suppose we are working in the following settings

- We have a sequence $\{\mathcal{P}_n\}$ of probability distributions for $Z = (X, Y)$ indexed by n where $Z \in \mathbb{R}^{d_n+1}$. For each n we observe n samples $Z_1, \dots, Z_n$ from some probability distribution $P \in \mathcal{P}_n$
- We have a sequence of sets $\{K_n\}$ where $K_n \subset \mathbb{R}^{d_n}$
- For each n we are interested in constrained least squares estimators

  $$\beta_n^* = \arg\min_{\beta \in K_n} \left\{ \mathbb{E}\left[\left(Y - X^T\beta\right)^2\right] \right\}$$

  Note, that $\beta_n^* = \beta_n^*(P)$ where $P \in \mathcal{P}_n$ is distribution of observed Z

Example Here are two examples of K_n

- $K_n = \left\{ \beta \in \mathbb{R}^{d_n} \middle| ||\beta||_1 \leq L_n \right\}$ - Lasso-type condition
- $K_n = \left\{ \beta \in \mathbb{R}^{d_n} \middle| ||\beta||_0 \leq C_n \right\}$ - Best subset-type condition

Definition 14.1 *Given a pair of sequences* $[\{\mathcal{P}_n\}, \{K_n\}]$*, a sequence of estimators* $\left\{\widehat{\beta}_n\right\}$ *is persistent if*

$$R_{P_n}(\widehat{\beta}_n) - R_{P_n}(\beta_n^*) \to^p 0$$

uniformly over $\{\mathcal{P}_n\}$*. Here* $\to^p$ *denotes convergence in probability.*

Let $K_n = \left\{\beta \in \mathbb{R}^{d_n} \middle| ||\beta||_1 \leq L_n\right\}$ - Lasso condition. This sequence of sets defines Lasso estimator

$$\widehat{\beta}_n = \arg\min_{\beta \in K_n} \left\{\frac{1}{n}\sum_{i=1}^{n}\left(Y_i - X_i^T\beta\right)^2\right\}$$

14.1.2 Persistence for Lasso

Theorem 14.2 *Under some growth condition on* d_n *and* L_n *Lasso estimator provides persistent sequence* $\left\{\widehat{\beta}_n\right\}$*. In other words, Lasso estimator is persistent.*

Proof:

For simplicity, we assume that Z is zero-mean random variable.

Let $\Sigma_n \in \mathbb{R}^{(d_n+1)\times(d_n+1)}$ - covariance matrix of Z. Let us also consider the estimator $\widehat{\Sigma}_n = \frac{1}{n}\sum_{i=1}^{n} Z_i Z_i^T$.

Now assume that $||\Sigma_n - \widehat{\Sigma}_n||_\infty = \max_{ij} |\Sigma_n^{(ij)} - \widehat{\Sigma}_n^{(ij)}| \leq E_n(\delta_n)$ with probability at least $1-\delta_n$. To maintain brevity, we do the following notational switch:

- $\beta \to \widetilde{\beta} = \begin{pmatrix} -\beta \\ 1 \end{pmatrix} \in \mathbb{R}^{d_n+1}$ so $Y - X^T\beta = Z^T\widetilde{\beta}$
- $L_n \to \widetilde{L}_n = L_n + 1$
- $K_n \to \widetilde{K}_n = \left\{\begin{pmatrix} -\beta \\ 1 \end{pmatrix} \middle| \beta \in K_n\right\}$
- $R_P(\beta) \to R_P(\widetilde{\beta}) = \mathbb{E}\left[\left(Z^T\widetilde{\beta}\right)^2\right]$

To proof the theorem, we need the following lemms

Lemma 14.3 *Uniformly over all* $P \in \{\mathcal{P}_n\}$

$$R_P(\widehat{\widetilde{\beta}}) \leq R_P(\widetilde{\beta}^*) + 2E_n(\delta_n)\widetilde{L}_n^2$$

with probabiliy at least $1-\delta_n$

Proof: Note that $R_P(\widetilde{\beta}) = \widetilde{\beta}^T\Sigma\widetilde{\beta}$ and $\widehat{R}_P(\widetilde{\beta}) = \widetilde{\beta}^T\widehat{\Sigma}\widetilde{\beta}$ where

$$\widehat{R}_P(\widetilde{\beta}) = \frac{1}{n}\sum_{i=1}^{n}\left(Z_i^T\widetilde{\beta}\right)^2$$

If $\widetilde{\beta} \in \tilde{K}_n$ then with probability at least $1 - \delta_n$

$$|R_p(\widetilde{\beta}) - \widehat{R}_P(\widetilde{\beta})| = |\widetilde{\beta}^T \left(\Sigma - \widehat{\Sigma}\right) \widetilde{\beta}| \leq ||\Sigma - \widehat{\Sigma}||_\infty ||\widetilde{\beta}||_1 \leq E_n(\delta_n)\widetilde{L}_n^2$$

Now we can derive that

$$R_P(\widehat{\widetilde{\beta}}) \leq^{(i)} \widehat{R}_P(\widehat{\widetilde{\beta}}) + E_n(\delta_n)\widetilde{L}_n^2 \leq^{(ii)} \widehat{R}_P(\widetilde{\beta}^*) + E_n(\delta_n)\widetilde{L}_n^2 \leq^{(iii)} R_P(\widehat{\widetilde{\beta}}) + 2E_n(\delta_n)\widetilde{L}_n^2$$

Here (i) and (iii) follows from the obtained bound on $|R_p(\widetilde{\beta}) - \widehat{R}(\widetilde{\beta})|$, (ii) follows from the fact that $\widehat{\widetilde{\beta}}$ minimizes $\widehat{R}(\widetilde{\beta})$ over all $\widetilde{\beta} \in \widetilde{K}_n$. This concludes the proof of the lemma.

■

The result of the lemma allows us to conclude that if $\delta_n \to 0$ and $E_n(\delta_n)\widetilde{L}_n^2 \to 0$ then the sequence of $\widehat{\beta}_n$ that corresponds to Lasso is persistent. ■

Comment: Under standard sub-gaussian conditions if $d_n \sim n^\alpha, \alpha \geq 0$

$$E_n(\delta_n) \sim \sqrt{\frac{\log d + \log n}{n}} \sim \sqrt{\frac{\log n}{n}}.$$

Then Lasso Estimator is persistent if

$$L_n = o\left[\left(\frac{n}{\log n}\right)^{1/4}\right]$$

14.1.3 Persistence for Best subset selection

If we consider the sequence of sets $K_n = \left\{\beta \in \mathbb{R}^d \middle| ||\beta||_0 \leq C_n\right\}$. If we assume that $\forall n : ||\beta_n^*||_0 \leq C$ where C is some universal constant that does not depend on n, then the rate of persistence for best subset selection least squares is

$$C_n = o\left(\sqrt{\frac{n}{\log n}}\right)$$

14.1.4 Further reading

The following papers are recommended

- A Distribution-Free Theory of Nonparametric Regression [GKKW02]
- Assumptionless consistency of the Lasso [C13]

14.2 Principal Component Analysis

14.2.1 Setup

$X \in \mathbb{R}^d$ - random vector with covariance matrix Σ which has eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_d \geq 0$. Each eigenvalue λ_i has associated eigenvector v_i such that $\Sigma v_i = \lambda_i v_i$. Given that, we can represent Σ as

$$\Sigma = \sum_{i=1}^{d} \lambda_i v_i v_i^T$$

The PCA is connected with direction of maximal variance of the distribution. The following figure represent samples from 2D Gaussian distribution with covariance matrix $\Sigma \neq I$. The variance is not uniform across all the directions and there is a direction along which the variance takes maximal value.

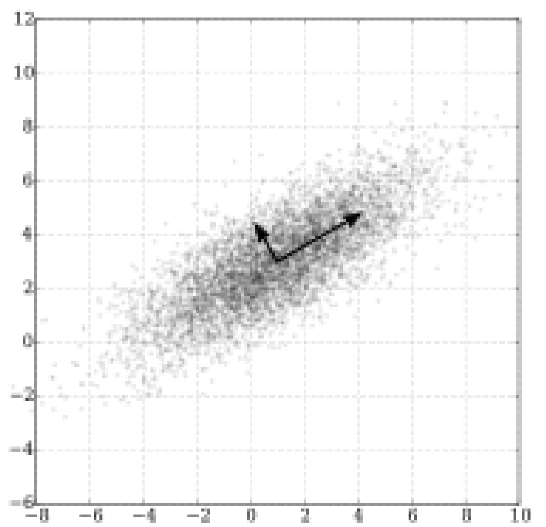

Figure 14.1: Samples from 2D Gaussia Distribution

Equivalently, v^* gives the direction of the maximal variance if

$$v^* \in \arg\max_{v \in S^{d-1}} \mathbb{V}\left[v^T X\right] = \arg\max_{v \in S^{d-1}} \left\{v^T \Sigma v\right\} = v_1$$

Here v_1 is the eigenvector associated with the largest eigenvalue λ_1

References

[GKKW02] L. Gyorfi, M. Kohler, A. Krzyzak and H.Walk, "A Distribution-Free Theory of Nonparametric Regression", *Springer Series in Statistics*, 2002

[C13] S. Chatterjee, "Assumptionless consistency of the Lasso", *arXiv:1303.5817*, 2013