Spring 2018: STA 6448
Advanced Probability and Inference II
Lecture 18

Yun Yang

- Random matrices and covariance estimation

Applications to covariance matrices

Corollary (Sample Covariance concentration)

Let X_i be i.i.d. zero-mean random vectors with covariance Σ, such that $\|x_i\|_2 \leq \sqrt{b}$ almost surely. Then for all $\delta > 0$,

$$\mathbb{P}\big[\|\widehat{\Sigma} - \Sigma\|_{op} \geq \delta\big] \leq 2d \exp\left(-\frac{n\delta^2}{2b(\|\Sigma\|_{op} + \delta)}\right).$$

Proof: Apply matrix Bernstein concentration inequality to $Q_i = x_i x_i^T - \Sigma$.

$$\|Q_i\|_{\mathrm{op}} \leq \|x_i\|_2^2 + \|\Sigma\|_{\mathrm{op}} \leq 2b.$$

Moreover,

$$\mathrm{Var}(Q_i) \leq \mathbb{E}[(x_i x_i^T)^2] \preceq b\Sigma.$$

Example: Random vectors uniform on sphere

x_i are chosen uniformly from the sphere $S^{d-1}(\sqrt{d})$, so that $\|x_i\|_2 = \sqrt{d}$.

By construction, $\mathbb{E}[x_i x_i^T] = \Sigma = I_d$, and $\|\Sigma\|_{\mathrm{op}} = 1$. Therefore,

$$\mathbb{P}\left[\|\widehat{\Sigma} - \Sigma\|_{\mathrm{op}} \geq \delta\right] \leq 2d \exp\left(-\frac{n\delta^2}{2d(1+\delta)}\right),$$

which implies the high probability bound

$$\|\widehat{\Sigma} - \Sigma\|_{\mathrm{op}} \lesssim \sqrt{\frac{d \log d}{n}} + \frac{d \log d}{n}.$$

This bound is off by a factor of $\log d$, since we can directly apply the matrix sub-Gaussian concentration inequality (x_i is sub-Gaussian with parameter c for some universal constant $c > 0$).

Example: “Spiked” random vectors

x_i is uniformly chosen from $\{\sqrt{d}e_1, \ldots, \sqrt{d}e_d\}$, where $e_j \in \mathbb{R}^d$ is the canonical basis vector with 1 in position j.

As before, we have $\|x_i\|_2 = \sqrt{d}$, and $\mathbb{E}[x_i x_i^T] = I_d$. Therefore, the same bound applies:

$$\|\widehat{\Sigma} - \Sigma\|_{\mathsf{op}} \lesssim \sqrt{\frac{d \log d}{n}} + \frac{d \log d}{n}.$$

This time, this bound is sharp (up to constant factors).

Structured covariance estimation: sparsity and thresholding

- Suppose Σ is known to be sparse, but the positions of non-zero entires are unknown.
- Motivates estimators based thresholding.
- Given a tuning parameter $\lambda > 0$, define the *hard thresholding operator* $T_\lambda : \mathbb{R} \to \mathbb{R}$ by

$$T_\lambda(u) = u\,\mathbb{I}[|u| > \lambda].$$

- For a matrix M, we define $T_\lambda(M)$ by applying T_λ to each element.
- We will study the property of the estimator $T_{\lambda_n}(\widehat{\Sigma})$, where $\lambda_n > 0$ is a suitably chosen parameter.

Sparsity and thresholding

- Let $A \in \mathbb{R}^{d\times d}$ denote the adjacency matrix, where $A_{ij} = \mathbb{I}(\Sigma_{ij} \neq 0)$.
- $\|A\|_{\text{op}}$ provides a measure of sparsity: if Σ has at most s non-zero entries per row, then $\|A\|_{\text{op}} \leq s$.

Theorem

x_i are independent zero-mean sub-Gaussian with parameter at most σ^2. If $n \geq \log d$, then for any $\delta > 0$ and $\lambda_n/\sigma^2 = 8\sqrt{\frac{\log d}{n}} + \delta$,

$$\mathbb{P}\left[\|T_{\lambda_n}(\widehat{\Sigma}) - \Sigma\|_{op} \geq 2\,\|A\|_{op}\lambda_n\right] \leq 8e^{-\frac{n}{16}\min\{\delta,\delta^2\}}.$$

Corollary

Suppose Σ has at most s non-zero entries per row, then

$$\mathbb{P}\left[\|T_{\lambda_n}(\widehat{\Sigma}) - \Sigma\|_{op}/\sigma^2 \geq 16s\sqrt{\frac{\log d}{n}} + 2\delta\right] \leq 8e^{-\frac{n}{16}\min\{\delta,\delta^2\}}.$$

Example: Sparsity and adjacency matrices

- In certain cases, the two bounds discussed before coincide.
- Consider any graph with maximum degree $s-1$ that contains a s-clique
- For any such graph, we have

$$\|A\|_{\text{op}} = s - 1.$$

- In general, the bound with $\|A\|_{\text{op}}$ can be substantially sharper.
- Consider a hub-and-spoke graph, in which one central node known as the hub is connected to s of the remaining $d-1$ node.
- For this graph, we have

$$\|A\|_{\text{op}} = \sqrt{s}.$$

Proof of the bound

Step one: for any $\lambda_n > 0$ such that $\|\widehat{\Sigma} - \Sigma\|_{\max} \leq \lambda_n$, we have

$$\|T_{\lambda_n}(\widehat{\Sigma}) - \Sigma\|_{\mathrm{op}} \leq 2\,\|A\|_{\mathrm{op}}\lambda_n.$$

In fact, this is implied by the (element-wise) relation

$$|T_{\lambda_n}(\widehat{\Sigma}) - \Sigma| \leq 2\lambda_n A.$$

Step two: Element-wise infinity norm concentration bound:

Lemma

Let $\widehat{\Delta} = T_{\lambda_n}(\widehat{\Sigma}) - \Sigma$, *then for all* $t \geq 0$,

$$\mathbb{P}\left[\|\widehat{\Delta}\|_{\max}/\sigma^2 \geq t\right] \leq 8e^{-\frac{n}{16}\min\{t,t^2\}+2\log d}.$$

Proof: Using the sub-exponential tail bound and a union bound argument.

Approximate sparsity

- In many cases, σ has many non-zero entries, but many of them are "near-zero".
- One way to measure that is through the ℓ_q-norm of each row.
- More precisely, given a parameter $q \in [0, 1]$, assume

$$\max_{j=1,\ldots,d} \sum_{k=1}^{d} |\Sigma_{jk}|^q \leq R_q.$$

Property

Under this ℓ_q-norm constraint, for any $\lambda_n > 0$ such that $\|\widehat{\Sigma} - \Sigma\|_{\max} \leq \lambda_n/2$, we have

$$\|T_{\lambda_n}(\widehat{\Sigma}) - \Sigma\|_{\mathrm{op}} \leq 2\,R_q \lambda_n^{1-q}.$$

Linear model: Formulation

- Observe a response vector $Y \in \mathbb{R}^n$, and a collection of covariates (vectors) $\{X_1, \ldots, X_d\}$
- Assume Y is linked with X_j via the linear model

$$Y = \sum_{j=1}^{d} X_j \theta_j^* + w = X\theta^* + w, \quad w \sim \mathcal{N}(0, \sigma^2 I_n).$$

- $X = (X_1, \ldots, X_d)$ is called the design matrix, and $\theta^* = (\theta_1^*, \ldots, \theta_d^*)^T$ is the unknown regression coefficient of interest.
- Scalarized form: for each index $i = 1, \ldots, n$,

$$y_i = \langle x_i, \theta^* \rangle + w_i,$$

where y_i, w_i are the ith component of y, w, and x_i^T is the ith row of X.

Sparse linear models in high dimensions

- We are interested in the high-dimensional regime where $d > n$
- The noiseless linear model is an under-determined linear system, and we need some form of low-dimensional structure
- A commonly made assumption is the hard sparsity assumption, meaning that the support set of θ^*,

  $$S(\theta^*) = \{j : \theta_j \neq 0\},$$

  has cardinality $s = |S|$ substantially smaller than d.
- A related milder assumption is the weak sparse assumption, where θ^* belongs to the ℓ_q-ball for some $q \in [0, 1]$,

  $$\mathbb{B}_q(R_q) = \big\{\theta \in \mathbb{R}^d : \sum_{i=1}^{d} |\theta_j|^q \leq R_q\big\}.$$

Applications of sparse linear models

Gaussian sequence model

Observations are of the form

$$y_i = \sqrt{n}\theta_i^* + w_i, \quad \text{for } i = 1, \ldots, n,$$

where $w_i \sim \mathcal{N}(0, \sigma^2)$ are i.i.d. noise variables.

Many non-parametric estimation problems can be reduced to an “equivalent" instance of the Gaussian sequence model.

Signal denoising in orthonormal bases

One observes corrupted samples $\widetilde{y}_i = \beta_i^* + \widetilde{w}_i$, where w_i are additive noises. Based on the observation vector $y \in \mathbb{R}^n$, the goal is to “denoise” the signal. Many classes of signals exhibit sparsity when transformed into an appropriate basis. Such transform can be represented as an orthogonal $\Psi \in \mathbb{R}^{d \times d}$, so that $\theta^* = \Psi^T \beta^*$ is expected to be sparse.

Applications of sparse linear models

Lifting and non-linear functions

Consider polynomial functions of degree k,

$$f_\theta(t) = \theta_1 + \theta_2 t + \cdots + \theta_{k+1} t^k.$$

Then polynomial regression $y_i = f_\theta(t_i) + w_i$ can be converted into an instance of the linear regression model.

More generally, we may consider lifting to linear combinations of some set of basis functions $\{\phi_1, \ldots, \phi_b\}$,

$$f_\theta(t) = \sum_{j=1}^{b} \theta_j \phi_j(t).$$

The same ideas also apply to multivariate functions.

Applications of sparse linear models

Signal compression in overcomplete bases

In the signal denoising example, we considered orthogonal transformations represented by the columns of an orthonormal matrix $\Psi \in \mathbb{R}^{d \times d}$. In many cases, it can be useful to consider an overcomplete set of basis functions, represented by the columns of a matrix $X \in \mathbb{R}^{n \times d}$ with $d > n$.

Signal compression can be performed by finding a vector $\theta \in \mathbb{R}^d$ such that $y = X\theta$. Since $d > n$, this equation may have multiple solutions, and the goal is to find the a sparse solution θ^* with $\|\theta^*\|_0 = s \ll n$ non-zeros.

Problems involving ℓ_0-constraints are computationally intractable. A popular relaxation is to seek a sparse solution by solving the basis pursuit program

$$\widehat{\theta} \in \operatorname{argmin} \ \|\theta\|_1, \quad \text{such that } y = X\theta.$$

Applications of sparse linear models

Compressed sensing

The classical approach to exploiting sparsity for signal compression is wasteful since it needs to compute the full vector $\theta = \Psi^T \beta^* \in \mathbb{R}^d$. This motivates compressed sensing, which is based on the combination of ℓ_1-relaxation with the random projection method.

The idea is to take $n \ll d$ random projections of β^*, each of the form $y_i = \langle x_i, \beta^* \rangle$, where $x_i \in \mathbb{R}^d$ is a random vector. Then, the problem of exact reconstruction amounts to finding a solution of the under-determined linear system $y = X\beta$ such that $\Psi^T \beta$ is as sparse as possible. The transformed ℓ_1-relaxation becomes

$$\min_{\theta \in \mathbb{R}^d} \|\theta\|_1, \quad \text{such that } y = \widetilde{X}\theta,$$

where $\widetilde{X} = X\Psi$ and the recovered signal is $\beta = \Psi^T \theta$.

Applications of sparse linear models

Selection of Gaussian graphical models

Any zero-mean Gaussian random vector $(Z_1, \ldots, Z_d)$ has a density of the form

$$p_\Theta(z_1, \ldots, z_d) = \frac{1}{\sqrt{(2\pi)^d \det(\Theta^{-1})}} \exp\Big(- \frac{1}{2} z^T \Theta z \Big),$$

where $\Theta \in \mathbb{R}^{d \times d}$ is the inverse covariance matrix, also known as the precision matrix. For many interesting models, the precision matrix is sparse, with relatively few non-zero entries.

This problem can be reduced to an instance of sparse linear regression. For a given index $s \in V := \{1, 2, \ldots, d\}$, suppose that we are interested in recovering its neighborhood, meaning the subset $\mathcal{N}(s) = \{t \in V \mid \Theta_{st} \neq 0\}$. We can perform variable selection in linear regression

$$Z_s = \langle Z_{-s}, \theta^* \rangle + w_s, \quad w_s \sim \mathcal{N}(0, \sigma_s^2).$$