Spring 2018: STA 6448
Advanced Probability and Inference II
Lecture 20

Yun Yang

- Structural covariance estimation
- High-dimensional linear regression

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

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

Recovery in the noiseless setting

- We begin by focusing on the noiseless model

$$y = X\theta^*, \quad \text{where } y \in \mathbb{R}^n,\ X \in \mathbb{R}^{n\times d},\ \theta^* \in \mathbb{R}^d.$$

- When $d \geq n$, the solution of θ^* is not unique.
- Our goal is to find the sparsest solution:

$$\min_{\theta\in\mathbb{R}^d} \|\theta\|_0 \qquad \text{such that } X\theta = y.$$

- Computationally infeasible when d is large.
- Convex relaxation:

$$\min_{\theta\in\mathbb{R}^d} \|\theta\|_1 \qquad \text{such that } X\theta = y.$$

- Can be formulated as a linear program, we call it the *basis pursuit linear program*.

Exact recovery and restricted nullspace

- Question: when is solving the basis pursuit linear program equivalent to solving the original ℓ_0-problem?
- For any subset $A \subset \{1, \ldots, d\}$, define the sub-vector $\theta_A = (\theta_j : j \in A)$.
- Let S denote the support of θ^*.
- Define the cone

$$\mathcal{C}(S) = \big\{\Delta \in \mathbb{R}^d : \|\Delta_{S^c}\|_1 \leq \|\Delta_S\|_1\big\}.$$

Definition

The matrix X satisfies the restricted nullspace property with respect to S if $\mathcal{C}(S) \cap \text{null}(X) = \{0\}$.

Exact recovery and restricted nullspace

Theorem

For any fixed subset S, the following two properties are equivalent:

1. *For any $\theta^* \in \mathbb{R}^d$ with support S, the basis pursuit linear program has unique solution $\theta = \theta^*$;*
2. *The matrix X satisfies the restricted nullspace property with respect to S.*

Sufficient conditions for restricted nullspace

The earliest sufficient conditions were based on the incoherence parameter of the design matrix:

$$\delta_{PI}(X) = \max_{j \neq k} \left| \frac{\langle X_j,\, X_k \rangle}{n} \right|.$$

Property

If the pairwise incoherence satisfies the bound

$$\delta_{PI}(X) \leq \frac{1}{3s},$$

then the restricted nullspace property holds for all subsets S of cardinality at most s.

This condition holds with high probability for sub-Gaussian random matrices with i.i.d. elements as long as $n = \Omega(s^2 \log d)$.

Restricted isometry property (RIP)

Definition

For each $s = 1, \ldots, d$, the restricted isometry constant of $X \in \mathbb{R}^{n \times d}$ of order s is the smallest quantity $\delta_s(X) > 0$ such that

$$\left\| \frac{X_S^T X_S}{n} - I_s \right\|_{\text{op}} \leq \delta_S(X) \quad \text{for all subsets } S \text{ of size at most } s.$$

- Connection to the incoherence parameter: If $X/\sqrt{n}$ has unit-norm columns, then $\delta_{PI}(X) = \delta_2(X)$.
- In general, we have for $s \geq 2$,

$$\delta_{PI}(X) \leq \delta_s(X) \leq s\, \delta_{PI}(X).$$

RIP and restricted nullspace

Property

If the RIP constant of order $2s$ satisfies $\delta_{2s} < 1/3$, then the *uniform restricted nullspace property* holds for any subset S of cardinality $|S| \leq s$.

- The RIP constants for sub-Gaussian random matrices with i.i.d. elements are well-controlled as long as $n = \Omega(s \log(d/s))$.
- Neither the pairwise incoherence condition nor the RIP condition are necessary conditions.
- Counter-example: $\Sigma = (1-\mu) I_d + \mu \mathbf{1}\mathbf{1}^T$ for $\mu \in (0,1)$.