20/07/2015. RAJESH SUNDARESAN

Yihong Wu, High dimensional statistics. Lec. 1.

Examples.

1) • Sparse linear regression (compressed sensing) $\beta \in \mathbb{R}^p$, some underlying vector

Observe: $y = A\beta + \text{noise}$ (A: $p \times m$)

If β is k-sparse, then we need $m \gtrsim k \log p$ samples

2) • Cov. matrix $X_1, \dots, X_n \overset{iid}{\sim} N(0, \Sigma)$; $\Sigma : p \times p$, psd.

$n \times p$ data matrix

Test: $H_0 : \Sigma = I_p$.

$H_1 : \Sigma \neq I_p$ or its variant $\|\Sigma - I_p\| \geq \varepsilon$.

3) • Species problem : Ball-urn model.

k balls in urn

Coloured, red, black, blue, etc. # colours unknown.

Can only draw from urn. Estimate # colours.

Ref: Fischer's work in the 1940s, butterflies in the Malayan islands.

I : Fundamentals

II : Unstructured problems – primarily for lower bounds.

III : Structured problems – sparsity / smoothness / shape

IV : Unstructured functional estimation.

- E.g., instead of estimating Σ completely, focus on the principal component.
- instead of a distribution, estimate entropy.

V : Computational issues and barriers.

- E.g., community detection.
- Illustrative relaxation of combinatorial procedures.

I: Fundamentals : Decision theoretic framework

Focus on minimax risk.

• Experimental $\{P_\theta, \theta \in \Theta\}$ prob distributions on alphabet $\mathcal{X}$.

Θ : parameter space. $\Longleftrightarrow$ (Model)

b) Data : $X_1, X_2, \dots X_n \overset{iid}{\sim} P_\theta$.

c) Estimator : $\hat{\theta} : \mathcal{X}^n \to \Theta$

$(x_1, \dots, x_n) \mapsto \hat{\theta}(x^n)$.

Loss function: $\ell : \Theta \times \Theta \to \mathbb{R}$

$(\theta, \hat{\theta}) \mapsto \ell(\theta, \hat{\theta})$, e.g. $\|\theta - \hat{\theta}\|$

$\inf_{\hat{\theta}(\cdot)} \sup_{\theta \in \Theta} \mathbb{E}_\theta \left[\ell(\hat{\theta}, \theta) \right]$: worst-case expected loss. $:= R_n(\Theta)$.

Questions: How does it depend on n, p, sparsity parameter, Θ.

Worst case Bayes: $\theta \sim \pi$.

$$\max_{\pi} \inf_{\hat{\theta}} \mathbb{E}_{\theta \sim \pi} \left[\ell(\hat{\theta}, \theta) \right] \leq \inf_{\hat{\theta}(\cdot)} \sup_{\theta \in \Theta} \mathbb{E}_\theta \left[\ell(\hat{\theta}, \theta) \right] = R_n(\Theta).$$

A convex program.

↳ Relax to randomized strategies $P_{\hat{\theta}|X^n}$.

Relaxed version is also a convex program.

Remarks: 1) All lower bounds go this way; pick a ~~smart~~ π smartly.

2) Inspired good procedures (Aggregation, exponential weighting)

3) But often, impossible to compute, esp. in high dimensions.

Objectives: 1) Classical asymptotics: Fixed p, large n, $\Theta \subset \mathbb{R}^p$.

How does $R_n(\Theta) \xrightarrow[n \to \infty]{} 0$; $\frac{1}{n} \times$ Const

↳ Asymp. normality $\mathrm{Var} = \frac{1}{n} \cdot$ Fisher information

In high dimensions: $p \gg n$ or $p, n \to \infty$.

So we want nonasymp. characterisations of $R_n(\Theta)$ with universal constant factors.

Minimax rate vs. minimax risk

$\psi_n(\Theta) \asymp R_n(\Theta)$

Notation: $X_n \gtrsim Y_n \iff \frac{X_n}{Y_n} \geq C_1$, where C_1 is some abs. constant

$Y_n \lesssim X_n \iff \frac{X_n}{Y_n} \geq C_2$

When both, we write $X_n \asymp Y_n$.

Objective: Characterise $R_n(\Theta)$ as $\psi_n(\Theta)$ as above.

Lecture 2

Remarks: Is it wise to focus on $\mathbb{E}_\theta[\ell(\ )]$? Why not high prob. statements?

* Qn: $\psi_n(\Theta)$ as a function of ℓ?
  Structure theorems on the class of ℓs?

Focus on: • Θ — normed linear space ← set of sparse cov. matrices, set of sparse vectors, set of distr.
$\ell(\theta, \hat{\theta})$, a norm of $\|\theta - \hat{\theta}\|$.

Sample complexity: Min sample size for which $\exists \hat{\theta}$ s.t. $\sup_\theta \mathbb{E}_\theta[\ell(\hat{\theta}, \theta)] \leq \varepsilon$
$=: n^*(\varepsilon)$.
(inverse of $n \longmapsto R_n(\Theta)$).

Qn: Is independence of sampling important?

Example: Scalar Gaussian Location Model.

$\theta \in \Theta = \mathbb{R}$
$X_i = \theta + Z_i$, $Z_i \overset{iid}{\sim} N(0,1)$

$$\inf_{\hat\theta} \sup_{\theta\in\Theta} \mathbb{E}_\theta\left[\|\hat\theta-\theta\|^2\right] \geq \sup_{\theta_0,\theta_1\in\Theta} \left[\frac{\|\theta_0-\theta_1\|^2}{8}\left(1-TV(P_{\theta_0},P_{\theta_1})\right)\right]$$

$$TV(P_{\theta_0},P_{\theta_1}) = TV\left(N(0,I_k), N\left(\begin{pmatrix}0\\0\\\|\theta_1-\theta_0\|\end{pmatrix}, I_k\right)\right)$$

$$\overset{\text{here}}{=} TV\left(N(0,1), N(\|\theta_1-\theta_0\|,1)\right) = 2Q\left(\tfrac{1}{2}\|\theta_1-\theta_0\|\right)$$

Thus $\inf_{\hat\theta}\sup_{\theta\in\Theta}[\dots] \geq \sup_{\theta_0,\theta_1\in\mathbb{R}^k} \frac{\|\theta_1-\theta_0\|^2}{8}\left(1-2Q\left(\frac{\|\theta_1-\theta_0\|}{2}\right)\right)$

$= \Theta(1)$. (k disappears)

How to fix this issue?

1) $\hat\theta : \mathbb{R}^k \to \mathbb{R}^k$ $\qquad \hat\theta_1 : \mathbb{R}^k \to \mathbb{R}$, $\dots$, $\hat\theta_k : \mathbb{R}^k \to \mathbb{R}$

$$\sum_{i=1}^k \inf_{\hat\theta_i}\sup_{\theta\in\Theta} \mathbb{E}_\theta\left[(\hat\theta_i-\theta_i)^2\right] \geq \sum_{i=1}^k \inf_{\hat\theta_i}\sup_{\theta_i} \mathbb{E}_{\theta_i}\left[(\hat\theta_i-\theta_i)^2\right]$$

$= k\cdot\Theta(1)$.

2) Many hypotheses. Fano's inequality

Given $\{P_1, P_2, \dots, P_M\}$ on $\mathcal{X}$.

Test $\hat J : \mathcal{X} \to [M] = \{1,2,\dots,M\}$

$$\max_{j\in[M]} P_j[\hat J \neq j] \geq \frac{1}{M}\sum_j P_j(\hat J \neq j) \geq 1 - \frac{I(U;X)+\log 2}{\log M}.$$

(U: distr. that is unif$([M])$.)

References:
1) A. Tsybakov, Introduction to nonparametric statistics.
2) Le Cam, Asymp. theory of ... statistics.
3) Le Cam & Yang. 2nd book.

Day 2 21/07/2015

Yihong Wu Lecture 3.

Recapitulation.

- $\{P_\theta : \theta \in \Theta\}$ model
  Typically $\ell(\theta, \hat\theta) = \|\theta - \hat\theta\|^2$
  $\inf_{\hat\theta} \sup_{\theta\in\Theta} E_\theta[\|\theta - \hat\theta\|^2]$.

Running example.
$P_\theta = N(\theta, \frac{I_k}{n})$; $X = \theta + \frac{Z}{\sqrt{n}}$
min-max risk is $\frac{k}{n}$.

$$R_n(\Theta) = \inf_{\hat\theta} \sup_{\theta\in\Theta} E_\theta[\|\theta - \hat\theta\|^2] = \max_{\pi} \inf_{\hat\theta} E_{\theta\sim\pi}[\|\theta - \hat\theta\|_2^2]$$

Pick θ_0, θ_1, and concentrate π on it.
Provides a l.b.
This is Le Cam's method.

$$R_n(\Theta) \gtrsim \sup_{\theta_0,\theta_1} \frac{\|\theta_1 - \theta_0\|^2}{8}\left(1 - TV(P_{\theta_0}, P_{\theta_1})\right)$$

$$= \Theta(1).$$

So, we need a better π.

- Two fixes
  a) Component-wise distances
  b) Many hypotheses and Fano inequality.

- Recall Fano inequality.
  $\{P_1, P_2, \dots, P_M\}$
  $P_{err,avg} = \frac{1}{M}\sum_{j=1}^M P_j(\hat J \neq j) \geq 1 - \frac{I(U;X) + \log 2}{\log M}$
  ($U \sim \mathrm{unif}[M]$.)

$$I(U;X) = D(P_{X|U} \| P_X | P_U) = \inf_Q D(P_{X|U} \| Q | P_U)$$

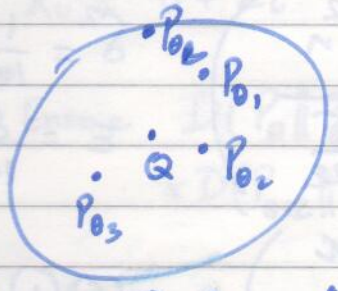
$$= \inf_Q \frac{1}{M}\sum_{j=1}^M D(P_j \| Q)$$

Note: $C = \max_{P_U} I(U;X) = \max_{P_U} \inf_Q D(P_{X|U} \| Q | P_U)$

$$\leq \inf_Q \max_{P_U} D(P_{X|U} \| Q | P_U) = \inf_Q \max_j D(P_j \| Q).$$

$\leq \max_{i,j} D(P_j \| P_i)$.

Thus $P_{err,avg} \geq 1 - \dfrac{\max_{i,j} D(P_j \| P_i) + \log 2}{\log M}$.



- Suppose $\theta_1, \theta_2, \ldots, \theta_M \in T \subset \Theta$ are such that $\|\theta_i - \theta_j\| \geq \varepsilon$.
- Estimator $\hat{J} = \arg\min_j \|\theta_j - \hat{\theta}\|^2$.

$\gtrsim \varepsilon^2 \left(1 - \dfrac{C(T) + \log 2}{\log M(T,\varepsilon)}\right)$, where $C(T) = \max_{U \in T} I(U;X)$.

$M(T,\varepsilon)$: packing number. Metric entropy (T will be defined soon as a convenient convex set)

- $T \subset (\Theta, d)$

ε-packing number of T : $M(T,\varepsilon)$ = max # (centres in T) of disjoint balls of radius ε that can be packed in T.

ε-covering number of T : $N(T,\varepsilon)$ = small number of balls of radius ε that cover T

- Fact: $M(T,\varepsilon) \geq N(T,2\varepsilon) \geq M(T,2\varepsilon)$.

$$N(T,2\varepsilon) \geq \frac{\mathrm{vol}(T)}{\mathrm{vol}(B(0,2\varepsilon))}$$

- Let $T = B(0,\delta)$, L^2-ball

$C(T)$ for the example is $\max_{\|\theta\| \leq \delta} I\left(\theta; \theta + \frac{1}{\sqrt{n}} \Gamma_k\right) \leq \max_{\mathbb{E}\|\theta\|^2 \leq \delta} I\left(\theta; \theta + \frac{1}{\sqrt{n}} \Gamma_k\right)$

$= \frac{k}{2} \log\left(1 + \frac{\delta^2 n}{k}\right) \leq \text{const.} \cdot \frac{1}{2} \delta^2 n.$

$$\log M(T,\varepsilon) \geq \log \frac{\mathrm{Vol}(T)}{\mathrm{Vol}(B(0,2\varepsilon))} = \log \frac{\delta^k}{(2\varepsilon)^k} = k \log \frac{\delta}{2\varepsilon}.$$

Now lower bound: $\varepsilon^2 \left(1 - \frac{\text{const.} \cdot \frac{1}{2}\delta^2 n}{k \log \frac{\delta}{2\varepsilon}}\right)$ $\quad \delta^2 = \frac{1}{100}\frac{k}{n}$, $\quad \varepsilon = \frac{\delta}{100}$

$$\varepsilon^2 \left(1 - \frac{\text{const.} \cdot \frac{k^2}{n^2} \cdot n}{k}\right) = O(1) \cdot \varepsilon^2 = O\left(\frac{k}{n}\right)$$

(controlled)

Lec 4.

1) What if noise is non-Gaussian?
2) What if loss is not mean-square.

1) $X_i = \theta + Z_i$ → non-Gaussian

Ensure $D(P_\theta \| P_{\theta'}) \simeq \frac{1}{2}\|\theta - \theta'\|^2$.

$\propto D(P_{\theta+Z} \| P_Z) \simeq \frac{1}{2}\|\theta\|^2$.

$$\frac{C(T)}{M(T,\varepsilon)} \leq \frac{\max_{\theta,\theta' \in T} D(P_\theta^{\otimes n} \| P_{\theta'}^{\otimes n})}{M(T,\varepsilon)} \simeq \frac{n\delta^2}{M(T,\varepsilon)}$$

2) Diff. loss fn. $\|\cdot\|$ norm on $\mathbb{R}^n$.

$$\|x\|_* = \sup_{\|y\| \leq 1} \langle x, y \rangle.$$

We will see $D\left(\frac{k}{\mathbb{E}[\|Z\|_*]}\right) \leq R_n \leq \frac{\mathbb{E}[\|K_n\|^2]}{n}$

↑ to show

$$\frac{1}{n}\left(\frac{k}{\mathbb{E}[\|Z\|_*]}\right)^2$$

Take $\|\cdot\|_1$. Then $\mathbb{E}[\|Z\|_\infty] \simeq \sqrt{\log n}$

Thus $\frac{k^2}{n \cdot \log n}$: off by a factor $\log n$.

Use the volume method:

$$\log M(B_2(\delta), \varepsilon) \geq \log \frac{\mathrm{Vol}(B_2(\delta))}{\mathrm{Vol}(B_{\|\cdot\|}(\varepsilon))}$$

Need upper bound on $\mathrm{Vol}(B_{\|\cdot\|}(\varepsilon))$.

Fact: Let $K \subset \mathbb{R}^d$ convex body with a nonempty interior.

Then $\mathrm{vol}(K) \geq \frac{1}{\mathbb{E}[\|Z\|_K]}$, where $Z \sim N(0, I_d)$

$\|x\|_K = \inf\{v \mid x \in vK\}$

$= \inf\left\{v : \frac{x}{v} \in K\right\}$

If $K = B_{\|\cdot\|}$, then $\|\cdot\|_K = \|\cdot\|$.

Fact: $\mathrm{Vol}(K) \leq \frac{w(K)}{\sqrt{d}}$, where $w(K) = \mathbb{E} \sup_{y \in K} \langle Z, y \rangle$, Gaussian width of K.

Thus $\mathrm{vol}(K)^{1/k} \leq \delta \mathrm{vol}(B_{\|\cdot\|}) \leq \frac{\delta}{\sqrt{k}} \mathbb{E}\left[\sup_{y \leq 1} \langle Z, y \rangle\right]$.

- Covariance matrix estimation.

  $X_1, X_2, \ldots, X_n \overset{iid}{\sim} N(0, \Sigma_{k \times k})$.

  Suff. statistic $S = \frac{1}{n} \sum_{i=1}^{n} X_i X_i' = \frac{1}{n} XX'$.

  (X_i': transpose)

  Is it minimax? Stein: minimax must satisfy a group invariance property. This S does not.

$$\inf_{\hat{\Sigma}} \sup_{\|\Sigma\|_{op} \leq \lambda} \mathbb{E}\left[\|\hat{\Sigma} - \Sigma\|^2\right] \asymp \left(\frac{k}{n} \wedge 1\right) \cdot \lambda^2 \cdot \|I_k\|^2$$

($\|\Sigma\|_{op} = \sigma_{max}(\Sigma)$)

as long as the norm is unitarily invariant.

Fact: S attains the upper bound to within constant factors.

- $D(N(0, \Sigma_0) \| N(0, \Sigma_1)) = \frac{1}{2} \mathrm{Tr}\left(\Sigma_0^{-1} \Sigma_1 - I\right) - \frac{1}{2} \log \frac{\det \Sigma_1}{\det \Sigma_0}$

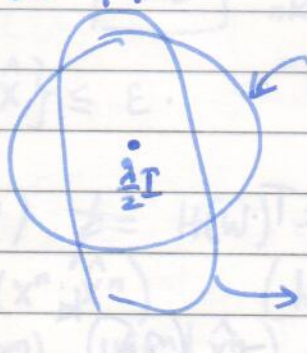
$$= \frac{1}{2} \sum_k \left(\sigma_i\left(\Sigma_0^{-1} \Sigma_1\right) - 1 - \log \sigma_i\left(\Sigma_0^{-1} \Sigma_1\right)\right)$$

Take $T = \left\{ \Sigma \;\middle|\; D(N(0,\Sigma) \| N(0,\Sigma_0)) \le \cdots \right\}$.
($N(0,\Sigma_0) \to \frac{1}{2}\lambda I_k$)

When $\sigma_i(\Sigma_0^{-1}\Sigma_1) \approx 1$, then $D(\cdot\|\cdot)$ is roughly quadratic and we get $\frac{1}{2}\sum_k (\sigma_i(\Sigma_0^{-1}\Sigma_1 - I))^2 = \frac{1}{2}\|\Sigma_0^{-1}\Sigma_1 - I\|_F^2$

$\le \frac{1}{2}\|\Sigma_0^{-1}\|_{op}^2 \cdot \|\Sigma_0 - \Sigma_1\|_F^2$.
($\|\Sigma_0^{-1}\|_{op} \to \frac{2}{\lambda} I_k$)

Use this to pick T:



Basic idea: $\frac{\text{Vol}(\text{KL nbd})}{\text{Vol}(\text{ball coming from loss fn})}$ ← model.

- DMS $P_X = P$.
  $X_1, X_2, \ldots, X_n$ iid P
  P: a distr. on $[k] = \{1, 2, \ldots, k\}$.

  $\inf_{\hat{P}} \sup_{P \in M_k} \mathbb{E}\, TV(P, \hat{P}) = R(n,k) \asymp \sqrt{\frac{k-1}{n}} \wedge 1$.

  Sufficient statistic empirical distr. $\frac{N(j|X^n)}{n}$, $j \in [k]$.
  ($\to \approx \frac{Bin(n, p_j)}{n}$.)

Focus on $k \ge 2$

$\mathbb{E}[TV(\tilde{P}, P)] \le \mathbb{E}[\|P - \tilde{P}\|_1]$

$= \sum_{j=1}^k |p_j - \hat{p}_j| \overset{CS}{\le} \sum_{j=1}^k \sqrt{\frac{p_j(1-p_j)}{n}} \le \frac{1}{\sqrt{n}} \sum_{j=1}^k \sqrt{p_j} \le \sqrt{\frac{k}{n}}$

Lower bound: $D(P\|Q) = \sum_{j=1}^{k} p_j \log \frac{p_j}{q_j} = \sum_{j=1}^{k} p_j \left(-\frac{q_j}{p_j} + 1 - \left(\frac{q_j}{p_j} - 1\right)^2 \right) + \dots$

$$\cong \sum_{j=1}^{k} p_j \left(\frac{q_j}{p_j} - 1\right)^2 + \dots$$

$Q = \left(\frac{1}{2(k-1)}, \dots, \frac{1}{2(k-1)}, \frac{1}{2} \right)$.

$T = \left\{ P \;\middle|\; |p_i - q_i| \le \delta \;\; \forall i = 1, \dots, k-1, \; p_k = 1 - \sum_{i=1}^{k-1} p_i \right\} \subset M_k$.
(M_k: prob-simplex.)

$\delta = \left(\frac{1}{k-1} \wedge \frac{1}{\sqrt{n(k-1)}} \right)$ const

$D(P\|Q) \le \delta^2 (k-1)^2$ follows for all $P \in T$.

$$\log M(T, \varepsilon) \ge \frac{\mathrm{Vol}(T)}{\mathrm{Vol}(B_{\|\cdot\|}(\varepsilon))} \ge \frac{(2\delta)^k}{\frac{(2\varepsilon)^{k-1}}{(k-1)!}} \ge \left(\frac{\delta(k-1)}{2\varepsilon e} \right)^{k-1}$$

($B_{\|\cdot\|}(\varepsilon)$: ℓ_1-distance)

Summary:

$$\inf_{\hat{P}} \sup_{P \in M_k} \mathbb{E}\left[D(P\|\hat{P}) \right] \asymp \frac{k-1}{n} \quad \text{(for ML)}.$$

Need to change the estimators to get better constants

add 1 estimator (Laplace)
add ½ estimator etc.
$\Big\}$ $\frac{k-1}{2n}$.

Discussion.

① Open questions: Does metric entropy relates bounds govern statistic nonse?

Ⓗ convex body $\subset \mathbb{R}^n$, $X = \theta + \frac{1}{\sqrt{n}} Z$

$\|\cdot\|^2$ loss fn.

How does $R_n(\text{Ⓗ})$ behave as a fn. of Ⓗ.

-11- Can say something if Ⓗ is characterised by sparsity.

Day 3

Lecture 5 Yihong Wu

- $R_n(\Theta) \geq \varepsilon^2\left(1 - \frac{C(T)+\log 2}{\log M(T,\varepsilon)}\right)$

A connection: Set ε to be pairwise $\min$ distance

$M(T,\varepsilon) = \max\{M \mid \exists \theta_1, \theta_2, \dots, \theta_M \text{ s.t. } \|\theta_i - \theta_j\| \geq \varepsilon \ \forall i,j\}$

We already saw

$M(T,\varepsilon) \geq N(T,\frac{\varepsilon}{2}) \geq \frac{Vol(T)}{Vol(B_{\|\cdot\|}(\frac{\varepsilon}{2}))}$

$T = B_2(1) \subset \mathbb{R}^k$.

$\left(1+\frac{2}{\varepsilon}\right)^k = \left(\frac{1+\varepsilon/2}{\varepsilon/2}\right)^k \geq M(T,\varepsilon) \geq N(T,\frac{\varepsilon}{2}) \geq \left(\frac{1}{\varepsilon/2}\right)^k$.

$X = \theta + Z$ where $Z \sim N(0, I_k)$, $\theta \in \mathbb{R}^k$

Criterion min. $\|\theta - \hat{\theta}\|^2$

Know: Min MSE $= \frac{k}{n} \times$ const.

$T = B_2(c_1\sqrt{k})$

$\varepsilon = c_2\sqrt{k}$.

$C(T) = \max_{\theta,\theta' \in T} D(P_\theta \| P_{\theta'}) = \max_{\theta,\theta' \in T} \|\theta - \theta'\|^2 = \text{const.}\ k$

$\log M(T,\varepsilon) = \text{const.}\ k$.

- Now, structure.

Denoising with sparsity.

$X = \theta + Z$, $Z \sim N(0, I_p)$, $\theta \in \mathbb{R}^p$.

But now $\theta \in \Theta = \{k\text{-sparse vectors}\} \subset \mathbb{R}^p$

$= \{\theta \in \mathbb{R}^p \mid \sum_{i=1}^p 1\{|\theta_i| \neq 0\} \leq k\} = \{\theta \in \mathbb{R}^p \mid \ell_0\text{-norm of } \theta \leq k\}$

$= B_0(k)$. $\|\theta\|_0$

Aside: $Z_1, Z_2, \dots, Z_p \overset{iid}{\sim} N(0,1)$

$\max_{1\le i\le p} Z_i = \sqrt{2\log p} + o_p(1)$: $P_r\{Z_i > a\} \doteq Q(a) \sim \frac{1}{a}\varphi(a) = \frac{\text{const.}}{a} e^{-a^2/2}$

$$P_r\left\{\max_{1\le i\le p} Z_i > \sqrt{2\log p}\right\} \le \frac{p}{\sqrt{2\log p}} e^{-\log p} = o(1).$$

Lower bound: $\inf_{\hat\theta} \sup_{\|\theta\|_0 \le k} \mathbb{E}_\theta \|\hat\theta - \theta\|_2^2 \gtrsim k\log\frac{ep}{k}$ ($\forall k \in [p]$, $\forall p \in \mathbb{N}$)

(because even if told where the nonzero locations are, need to estimate θ).

$\log\frac{ep}{k}$: extra for sparsity, locations.

- $T = B_2(\delta) \cap B_0(k)$, $C(T) \le \delta^2$.
- Choose (anticipating above result) $\varepsilon = \text{const.}\sqrt{k\log\left(\frac{ep}{k}\right)}$, $\delta = \text{const.}\cdot\varepsilon$.

$\Rightarrow C(T) \le k\log\frac{ep}{k}$.

- Goal: Find $\theta_1, \theta_2, \dots, \theta_M \in T$ s.t. $\|\theta_i - \theta_j\| \ge \varepsilon$

First, find $b_1, b_2, \dots, b_M \in \{0,1\}^p$ s.t.

a) $w_H(b_i) = k$

b) $d_H(b_i, b_j) \gtrsim \text{const.}\cdot k =: m$

Define: $\theta_i = b_i\sqrt{\frac{\delta}{k}}$, $i = 1, \dots, M$.

Now $\|\theta_i - \theta_j\|_2^2 = d_H(b_i, b_j)\cdot\frac{\delta}{k} \gtrsim \text{const.}\cdot\delta = \varepsilon$.

Number of such binary vectors: $M \gtrsim \frac{\binom{p}{k}}{|B_{\text{Hamming}}(m)|} = \frac{\binom{p}{k}}{1 + \binom{p}{1} + \binom{p}{2} + \cdots + \binom{p}{m}} = \frac{\binom{p}{k}}{\binom{p}{\le m}}$

Know: $\left(\frac{p}{k}\right)^k \le \binom{p}{k} \le \left(\frac{ep}{k}\right)^k$

Thus $M \gtrsim (p/k)^k / 2^{m\log\left(\frac{ep}{m}\right)} \Rightarrow \log M \gtrsim \frac{k\log p/k - m}{\text{const.}\cdot k - \log(ep/k)}$.

MLE: $X \sim P_\theta = N(\theta, I_p)$

$P_{x|\theta} \propto \exp\left(-\frac{1}{2}\|x-\theta\|_2^2\right)$

$\hat{\theta}_{MLE} = \underset{\text{s.t. } \|\theta\|_0 \le k}{\arg\min} \|x-\theta\|_2^2$ ⟶ Keep the largest k

Let $h = \hat{\theta}_{MLE} - \theta$

Observe that the ground truth is a feasible point.

Thus $\|x-\hat{\theta}\|_2^2 \le \|x-\theta\|_2^2$ (↳ Ground truth)

$\|z-h\|_2^2 \le \|x-\theta\|_2^2 = \|z\|^2$

$= \|z\|^2 + \|h\|^2 - 2\langle z, h\rangle$

$\Rightarrow \|h\|^2 \le 2\langle z, h\rangle = 2\|h\| \cdot \left\langle z, \frac{h}{\|h\|}\right\rangle$

$\Rightarrow \|h\| \le 2\left\langle z, \frac{h}{\|h\|}\right\rangle$

① $u = \frac{h}{\|h\|}$ has $\|u\| = 1$

② h is difference of 2 k-sparse vectors

So u is $2k$ sparse.

Thus $\|h\| \le 2 \cdot \sup_{u:\ \|u\|_0 \le 2k} \langle z, u\rangle =: B.$

- Can analyse this directly (chi-squared entries)
- Instead use metric entropy approach.

$u \in S^{p-1} \cap B_0(2k)$ ~~$\subseteq B_2(1) \cap B_0(2k)$~~

Let $\theta_1, \theta_2, \dots, \theta_N$ be an ε-net of ~~$B_2(1)$~~ $S^{p-1} \cap B_0(2k)$

$\forall u, \exists i$ s.t. $u = \theta_i + \varepsilon_i$, where $\|\varepsilon_i\| \le \varepsilon$.

$\langle z, u\rangle = \langle z, \theta_i\rangle + \langle z, \varepsilon_i\rangle$

$\langle z, \theta_i\rangle \sim N(0,1)$

ε_i is $4k$ sparse.

$\sup_{\|u\|_2=1,\ \|u\|_0 \le 4k} \langle z, u\rangle \le \sqrt{2}B$

Thus $B \leq \max_{i \in [N]} \langle Z, \theta_i \rangle + \varepsilon\sqrt{2}\, B$

Thus $B \lesssim \max_{i \in [N]} \langle Z, \theta_i \rangle \leq \sqrt{\log N}$.

But we know $\sqrt{\log\left[\left(1+\frac{2}{\varepsilon}\right)^k \cdot \binom{p}{k}\right]} \simeq \sqrt{k \log \frac{ep}{k}}$.

Remark: What if we don't know k?
We will have two more procedures that give $\sqrt{k \log p}$ instead of $\sqrt{k \log \frac{ep}{k}}$
But k need not be known.

Lecture 6.

Two procedures with simpler analyses.

1) $$\arg\min_{\text{s.t } \|X-\theta\|_\infty \le \tau = \sqrt{2\log p}} \|\theta\|_0 \quad = \quad \hat{\theta}$$

Solution: hard thresholding $\hat{\theta}_i = 1\{|x_i| > \tau\}$.

We will ~~do~~ analyse it differently.

w.h.p., $\mathbb{E}\|\hat{\theta}-\theta\|_2^2 \lesssim k \log p$

Also, ground truth θ is feasible w.h.p.

$$\left(\|\underbrace{Z+\theta}_{X} - \theta\|_\infty = \|Z\|_\infty \overset{\text{wlp}}{\le} \sqrt{2\log p} \right.$$
$$\left. \Rightarrow \theta \text{ is feasible wlp}\right)$$

$$\|\hat{\theta}\|_0 \le \|\theta\|_0 \le k$$

$\Rightarrow h = \hat{\theta}-\theta$ is at most $2k$-sparse.

So $\|X-\hat{\theta}\|_\infty \le \tau$

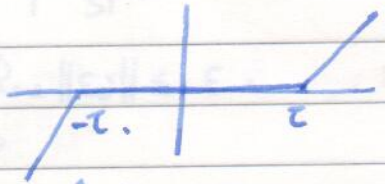
$\hookrightarrow = \|Z-h\|_\infty$. Since $\|Z\| \le \tau$, we have $\|h\|_\infty \le 2\tau$.

Now $\|h\|_2^2 \le \|h\|_\infty^2 \|h\|_0^2 \le 4\tau^2 \cdot 2k$ which is $k \log p$.

However, we can get to $k \log \frac{p}{k}$ by thresholding ~~each~~ the top level differently, the second level differently, and so on.

2) $$\arg\min_{\text{s.t. } \|X-\theta\|_\infty \le \tau = \sqrt{2\log p}} \|\theta\|_1 \quad \overset{\text{soft thresholding}}{\Longleftrightarrow} \quad \hat{\theta}$$

$$\hat{\theta}_i = (|x_i| - \tau)_+$$



We will analyse it differently, as before.

a) True θ is feasible.

b) $\hat{\theta}$ has lower $\|\theta\|_1$ norm. $\|\hat{\theta}\|_1 \le \|\theta\|$; $h = \hat{\theta}-\theta$

$h = h_J + h_{J^c}$, where $J = \mathrm{supp}(\theta)$, $|J| \le k$.

$\|h_{J^c}\|_1 = \|\theta\|_1 - \|\theta + h_J\|_1$ ✓

$\le \|\theta\|_1 - \|\theta\|_1 + \|h_J\|_1$

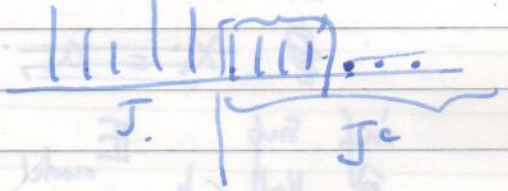
$\le \|h_J\|_1$.

$(\|\theta + h_J\|_1 \ge \|\theta\|_1 - \|h_J\|_1$

Step 3. h is well-approx. by a $2k$-sparse vector.

K = k largest indices of $|h_i|$, $i \notin J$.

$|J \cup K| \le 2k$.



Step 4: $\|h\|_2^2 \le \|h_{J\cup K}\|_2^2 \times 2$ → to be proved.

$\le \|h\|_\infty^2 \cdot \|h_{J\cup K}\|_0$ ($\|h_{J\cup K}\|_0 \le 2k$)

$= 2k \|h\|_\infty^2$

More than half of energy is in $2k$-sparse components

$\|h_{J^c}\|_1 \le \|h_J\|_1 \Rightarrow \|h_{J\cup K}\|_2 \ge \|h_{(J\cup K)^c}\|_2$

$\forall i \in (J\cup K)^c$, $|h_{(i)}| \le \frac{1}{i}\|h_{J^c}\|_1$

$\sum_{i=k+1}^{p} |h_{(i)}|^2 \le \|h_{J^c}\|_1^2 \cdot \sum_{i=k+1}^{p} \frac{1}{i^2} \le \|h_{J^c}\|_1^2 \cdot \frac{1}{k}$

$(i \in (J\cup K)^c)$

$\le \frac{1}{k}\|h_J\|_1^2 \le \|h_J\|_2^2 \le \|h_{J\cup K}\|_2^2$

(k-sparse; $\le$ by CS ineq.)

Sparse linear regression. (Linear inverse problem, compressed sensing)

$$X = A\theta + Z \qquad Z \sim N(0, I_n)$$

(X: $n\times 1$, A: $n\times p$, θ: $p\times 1$, Z: $n\times 1$)

A: Gaussian design matrix with $\overset{iid}{\text{entries}}$ $N(0, \frac{1}{n})$

Goal: Estimate θ based on X and A.

$a_i \to \boxed{\quad} \to \theta$; a_i = row of A.

~~$X_i = a_i \theta$~~

$$\inf_{\hat\theta} \sup_{\|\theta\|_0 \le k} \mathbb{E}_{\text{model}} \left[\|\hat\theta - \theta\|_2^2\right] \gtrsim k \log \frac{ep}{k}$$

is a lower bound even if we see θ in noise directly.

(model: over $P_{A,Z|\theta}$)

But, we can achieve $k \log \frac{ep}{k}$ inefficiently

can achieve $k \log p$ efficiently

when $n \gtrsim k \log \frac{ep}{k}$

Open: Why $n \gtrsim k \log\left(\frac{ep}{k}\right)$?

Lower bound:

$$D(P_{\theta_0} \| P_{\theta_1}) = D(P_{A,X|\theta_0} \| P_{A,X|\theta_1}) = \mathbb{E}_A \left[D(N(A\theta_0, I_n) \| N(A\theta_1, I_n)) \right]$$

$$= \frac{1}{2} \mathbb{E}\left[(\theta_0 - \theta_1)' A' A (\theta_0 - \theta_1)\right]$$

$$= \frac{1}{2} (\theta_0 - \theta_1)' \mathbb{E}[A'A] (\theta_0 - \theta_1)$$

$\mathbb{E}[A'A]$: $p\times p$ $\to$ ~~I_n~~ I_p.

$$= \frac{1}{2} \|\theta_0 - \theta_1\|_2^2$$

$\longrightarrow$ Same analysis as before carries over now

(It is as good as A is absent.)

Achievability MLE

$$\hat{\theta}_{MLE} = \arg\min_{\theta : \|\theta\|_0 \le k} \|X - A\theta\|_2^2$$

Discussion: 1) Connection between randomised versus the worst-case settings we are studying in these lectures.

2) Sequential approaches?

3) $\theta \overset{iid}{\sim} P_X = (1-\alpha)\delta_0 + \alpha N(0,1)$.

$\frac{n}{p} = \alpha$.

$n = \alpha p$ suffices.

Message passing style algorithms incorporating priors

Spatial coupling codes (entries of the matrix not necessarily iid

Enable successive cancellations).

Achieves information theoretic lower bound.

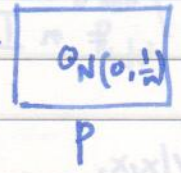
Day 4 23/07/2015

Yihong Wu Lec 7.

Recall: Sparse linear regression.

$X = A\theta + Z$ ($Z \sim N(0, I_n)$, $\theta \in B_0(k)$)

minimax MSE $\gtrsim k \log \frac{ep}{n}$



We will see that $R(k,p,n) \asymp k \log \frac{ep}{k}$, provided $n \gtrsim k \log \frac{ep}{k}$

$\hat{\theta}_{MLE} = \arg\min \|X - A\theta\|_2^2$
$\text{s.t.} \|\theta\|_0 \le k$

First $\|X - A\hat{\theta}\|_2^2 \le \|X - A\theta\|_2^2 \le \|Z\|_2^2$.

Write $h = \hat{\theta} - \theta$, then $\|Ah\|_2^2 \le -\langle Ah, Z\rangle$

$\le \|h\| \left\langle A\frac{h}{\|h\|}, Z \right\rangle$

Step 1. $\sup_{\substack{\|u\|_2 = 1 \\ \|u\|_0 \le 2k}} \langle Au, Z\rangle$.

Note that $\langle Au, Z\rangle = \langle A'Z, u\rangle$

We will show $\sup \langle Au, Z\rangle \lesssim_p \sqrt{k \log \frac{ep}{k}}$.

Step 2: $\|h\|_2^2 \asymp \|Ah\|_2^2$

$A'Z = \left(A' \frac{Z}{\|Z\|}\right) \cdot \|Z\| = \frac{\|Z\|_2}{\sqrt{n}} r$, $r \sim N(0, I_p)$ vector.

{ conditional law, but unconditional also yields $N(0, I_p)$.

$Ah = A_J \cdot (-)$ (2k columns)

$$\Pr\left\{ \sigma_{\min}\left(\sqrt{n}\, A_J \right) \text{ where } A \sim N\left(0, \tfrac{1}{n}\right) \text{ (m rows)} \leq 1 - \sqrt{\frac{m}{n}} - \frac{t}{\sqrt{n}} \right\} \leq e^{-t^2/2}$$

Since $\binom{p}{2k} \approx e^{2k \log \frac{ep}{k}}$.

Choose $t = \sqrt{k \log \binom{p}{2k}}$

Recall:

We already saw $A = I_p$ case.

$\min \|\theta\|_0$ s.t. $\|\theta - y\|_\infty \leq \tau = \sqrt{2 \log p}$, $\min \|\theta\|_1$ s.t. $\|\theta - y\|_\infty \leq \tau = \sqrt{2 \log p}$ } both were considered.

Now A is a fat matrix.

$\hat{\theta}_{DS} = \arg\min \|\theta\|_1$ (Dantzig-Selector)

s.t.

$$\|A'(X - A\theta)\|_\infty \leq \tau = \sqrt{2 \log p}.$$

Again, as before $h = \hat{\theta} - \theta$, $J = \operatorname{supp}(\theta)$.

Then

$$\|h_J\|_1 \leq \|h_{J^c}\|_1$$

We will soon see that θ is also feasible. (Indeed, $A'Z = \frac{\|k\|_2}{\sqrt{n}} N(0, I_n)$. Choose $\tau = \sqrt{2 \log p}$ so $\|\cdot\|_\infty$ ...)

So $\|h_{J \cup K}\|_2^2 \gtrsim \frac{1}{2} \|h\|_2^2$, as before, K = set of k largest magnitudes outside J.

- By linear algebra + ε-net argument

$$\|h\|_2^2 \lesssim \|Ah\|_2^2 \leq 4\varepsilon\sqrt{k}\, \|h\|_2^2.$$

↑ near isometry argument

We will argue only the second inequality:

$\|A'(X - A\theta)\|_\infty \leq \tau$, $\|A'(X - A\hat{\theta})\|_\infty \leq \tau$

$\Rightarrow \|A'Ah\| \leq 2\tau$ by triangle ineq

$$\|Ah\|^2 = \langle Ah, Ah\rangle \le \max_{\|w\|_1 \le 1} \langle A'Ah, h\rangle$$
$$\le \|A'Ah\|_\infty \cdot \|h\|_1$$
$$\le 2\tau \cdot 2\|h_J\|_1 \le 4\tau\sqrt{k}\,\|h_J\|_2$$
$$\le 4\tau\sqrt{k}\,\|h\|_2.$$

Functional Estimation. (E.g., Entropy estimation)

- $X = M + \frac{1}{\sqrt{n}} Z$ (Z: $p \times p$, with $N(0,1)$ components; estimate M.)
- $\inf_{\hat{M}} \sup_{M} \|\hat{M} - M\|_{op}^2 = \frac{1}{n}\mathbb{E}\left[\|Z\|_{op}^2\right]$ by Anderson's lemma
$$\asymp \frac{p}{n}.$$
- What if we want only $\|M\|_{op}$?
$$\left(\|\hat{M}\|_{op} - \|M\|_{op}\right) \le \|\hat{M} - M\|_{op}.$$
$\Rightarrow$ $\frac{p}{n}$ error is ach. via plug-in.

It will turn out, we will have $\asymp \frac{p}{n}$ here. So plug-in is best. No so for other functionals.

Thm: $\inf_{\hat{T}} \sup_{M} \mathbb{E}\left[\left|\hat{T} - \|M\|_{op}\right|\right] \asymp \frac{p}{n}$.

For lower bound, Le Cam's method works.

H_0: $\|M\|_{op} = 0$ $(M = 0)$

H_1: $\|M\|_{op} = \lambda$.

Try $\lambda = \sqrt{\frac{p}{n}}$.

For simplicity, let $n=1$ (scaling)

H_0: $M = 0$ $\quad (X = Z)$

H_1: $M = \lambda uu'$, (u uniform S^{p-1}). , $X = \lambda uu' + Z$

$TV(Z, \lambda uu' + Z) \leq \cdot$, is of interest since this arises in the LRT for the above test.

Let's bound $\sqrt{\chi^2(\lambda uu' + Z \| Z)}$ $\quad \left(\chi^2(P\|Q) = \mathrm{Var}_Q\left(\frac{P}{Q}\right) = \int \frac{(P-Q)^2}{Q}\right)$

Observations:

$$\chi^2(P_X \| P_Z) = \int \frac{p_X^2}{p_Z} - 1$$

$p_X(z) = \mathbb{E}_M(\phi(x-M))$: $\chi^2 = \int \frac{(\mathbb{E}\,\phi(x-M))^2}{\varphi(x)}\,dx - 1$

Now have $M, \tilde{M}$ iid $= \mathbb{E}\int \frac{\varphi(x-M)\varphi(x-\tilde{M})}{\varphi(x)} - 1$

$$= \mathbb{E}\left[\exp\langle M, \tilde{M}\rangle\right] - 1$$

$$\langle M, \tilde{M}\rangle = \langle (\lambda uu'), (\lambda \tilde{u}\tilde{u}') \rangle = \text{c. p. } \langle u, \tilde{u}\rangle^2$$

$$\langle u, \tilde{u}\rangle = \left\langle \frac{Z}{\|Z\|}, \tilde{u} \right\rangle = O_p(\sqrt{p})$$

Sparse M: Estimate $\|M\|_{op}$.

$k \sim p^{1/3}$ is a threshold.

Lec. 8

Estimating entropy.

$X_1, X_2, \dots, X_n \overset{iid}{\sim} p$ on $[k]$

Estimate $H(p)$.

samples $\frac{k}{\log k} f(\text{accuracy})$.

- If k is fixed, provably the best.
- If $k \sim n$, bias $= \mathbb{E}\left[\hat{H}_{\text{plug-in}}\right] - H(P) = -\mathbb{E}\left[D(\hat{P}\|P)\right] \le 0$.

Severely underbiased

Taylor expansion of bias $\mathbb{E}[\text{bias}] = -\frac{S(P)-1}{2n} + \frac{1}{12n^2}(\dots)$

$S(P)$: support. But we may not know it.

Paninsky $\exists n_k = o(k)$ s.t. $R^*(k, n_k) \xrightarrow[k\to\infty]{} 0$.

Valiant & Valiant: sharp scaling $n \asymp \frac{k}{\log k}$

Result. $n \gtrsim \frac{k}{\log k}$

$$R^*(k,n) \gtrsim \left(\frac{k}{n\log k}\right)^2 + \frac{\log^2 k}{n}$$

($\left(\frac{k}{n\log k}\right)^2$: bias^2; $\frac{\log^2 k}{n}$: variance)

$\lesssim \frac{k}{\log k}$ no consistent est. exists., i.e., $R^*(k,n) \gtrsim 1$

Compare with plugin: $\left(\frac{k}{n}\right)^2 + \dots$

Do unbiased estimators exist?

- $\text{Ber}(p)$ r.v.'s. Estimate $f(p)$.

  Unbiased est exists $\Leftrightarrow$ $f(p)$ = polynomial of degree $\le n$.

$$\mathbb{E}\left[\hat{f}(Y)\right] = \sum_{k=0}^{n} \hat{f}(k)\binom{n}{k} p^k (1-p)^{n-k},$$

- Use a large degree polynomial.
  Bias $\to 0$. But variance increases because # parameters is large.
  Find a sweet spot for the tradeoff.

- $h_j = \sum_{i=1}^{k} 1\{N_i = j\}$: histogram of histogram.

  $\hat{H} = \sum_{j=1}^{n} \alpha(j)\, h_j$, significantly faster than LP-based approaches.

- Impossibility result.
  Best polynomial approx.

$$\mathbb{E}_L(f, I) = \inf_{\deg(p) \le L} \sup_{x \in I} |f(x) - p(x)|$$

Dual: moment matching, upto L^{th} moments are same.

$$\mathcal{E}_L(f, I) = \sup \mathbb{E}[f(U) - f(U')]$$
$$\text{s.t. } U, U' \in I$$
$$\mathbb{E}[U^j] = \mathbb{E}[U'^j],\ j = 1, \dots, L$$

Fact:

$$\mathcal{E}_L(f, I) = 2\, \mathbb{E}_L(f, I)$$

Idea. (LeCam's idea)
Construct two ensembles of ~~entropy~~ distributions, one from H_0: Entropy ≤ 3 bits
other from H_1: Entropy ≥ 10 bits,
yet very difficult to distinguish between the two.

- One more example: How many butterflies, $S(P)$, support size estimation.

• Community detection in networks

Summary

Community scales linearly, sharp recovery via SDP
Sublinear , computationally hard via l.b.'s.

Two-equally sized communities, + −

$i \sim j$ indep. w.p. $p = a\frac{\log n}{n}$ if ++ or −−

$q = b\frac{\log n}{n}$ if +− or −+

- Assume $p > q$

MLE: $\max_{\sigma} \langle A, \sigma\sigma^T \rangle$

s.t. $\sigma_i \in \{+1, -1\}$, $i \in [n]$

$\sigma^T 1 = 0$

$\max_{Y} \langle A, Y \rangle$

s.t. $\operatorname{rank} Y = 1$

$Y_{ii} = 1$

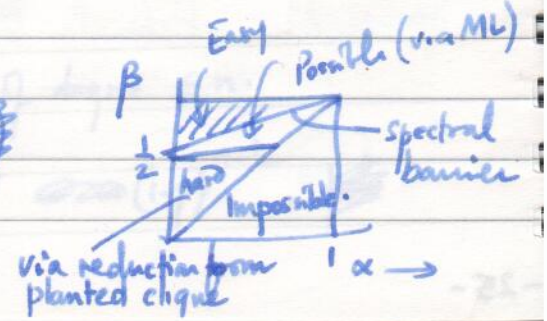
$\langle J, Y \rangle = 0$, J = all-1 matrix

Relax $\operatorname{rank} Y = 1$ to $Y \succeq 0$.
(Similar to Frieze–Jerrum 1995 for MAX BISECTION).

- Info theoretic: 1) If $(\sqrt{a} - \sqrt{b})^2 > 2$ → recovery in poly
< 2 → recovery is impossible.

2) SDP suffices in the recovery possible case.

- Finding a single community
(Planted clique problem). ~~Planted~~ Planted dense subgraph
$K = pn$: SDP achieves sharp threshold.



$K = n^p$

-26-
-24-

References

- Le Cam's book
- A. Tsybakov – Intro to nonparametric statistics.
  lower bound (2nd chapter)
- Ibragimov & Khasminski – Statistical decision theory
- Iain Johnstone – Gaussian seq. models.
  (On his web page).
- Inster & Suslina – Minmax ≥ quadratic.
- Quadratic fct estimates – papers