Estimating High-dimensional Matrices: Information-Theoretic Limits and Computational Barriers

Yihong Wu

Department of ECE
University of Illinois at Urbana-Champaign

yihongwu@illinois.edu

Joint work with Zongming Ma (Penn)

October 21, 2013

High dimensionality of contemporary datasets

Field	Data
Biomedical Research	microarray, ECG, fMRI, ...
Signal Processing	array sensor data, face recognition, hyper-spectral data, ...
Finance	asset returns, ...
$\vdots$	$\vdots$

- Growth of data outpaced by increasing number of features
- Statistical inference on massive datasets can be very costly

Estimation large matrices

Estimating a $p \times p$ matrix based on n observations

- Inference of mean structure
  - Image denoising
  - Multi-task learning
  - Matrix completion
- Inference of covariance structure
  - Covariance matrix estimation
  - Gaussian graphical models
  - PCA

Three challenges

- High dimensionality of data
  - large p, but comparable or smaller n
  - intrinsic low dimensionality of the signal

Three challenges

- High dimensionality of data
  - large p, but comparable or smaller n
  - intrinsic low dimensionality of the signal
- Non-quadratic losses (particularly those dealing with eigenvalues)
  - operator norm [Bickel-Levina '08, Cai-Zhou-Zhang '10, ...]
  - nuclear norm [Rhode-Tsybakov '11, ...]

Three challenges

- High dimensionality of data
  - large p, but comparable or smaller n
  - intrinsic low dimensionality of the signal
- Non-quadratic losses (particularly those dealing with eigenvalues)
  - operator norm [Bickel-Levina '08, Cai-Zhou-Zhang '10, ...]
  - nuclear norm [Rhode-Tsybakov '11, ...]
- computationally efficient and provably optimal algorithms (?)

This talk

Objectives

- non-asymptotic understanding of the decision-theoretic fundamental limit
    - information theory
    - convex geometry

This talk

Objectives

- non-asymptotic understanding of the decision-theoretic fundamental limit
  - information theory
  - convex geometry
- impact of complexity contraint on statistical optimality

Decision-theoretic setup

Main ingredients

- Observation: $X_1, \ldots, X_n \overset{iid}{\sim} P_\theta, \theta \in \Theta$
- Estimator $\hat{\theta}$
- Loss $L(\theta, \hat{\theta})$

Minimax risk and minimax rate

- Minimax risk: worst-case expected loss

$$R_n(\Theta) \triangleq \inf_{\hat{\theta}} \sup_{\theta \in \Theta} \mathbb{E}_\theta[L(\theta, \hat{\theta})]$$

Minimax risk and minimax rate

- Minimax risk: worst-case expected loss

$$R_n(\Theta) \triangleq \inf_{\hat{\theta}} \sup_{\theta \in \Theta} \mathbb{E}_\theta[L(\theta, \hat{\theta})]$$

- Minimax rate: non-asymptotic characterization of minimax risk

$$R_n(\Theta) \asymp \Psi_n(\Theta)$$

Particularly useful in high dimensions

[Notation:

$$X \asymp Y \Leftrightarrow c < X/Y < C$$

for universal constants c, C. Example: $a + b \asymp a \vee b$]

Minimax theorem

- Minimax = Worst-case Bayesian

$$\inf_{\hat{\theta}} \sup_{\theta \in \Theta} \mathbb{E}_\theta[L(\theta, \hat{\theta})] = \underbrace{\sup_{\theta \sim P}}_{\text{prior}} \overbrace{\inf_{\hat{\theta}} \mathbb{E}[L(\theta, \hat{\theta})]}^{\text{Bayesian risk}}$$

Important element missing...

The minimax risk

$$\inf_{\hat{\theta}} \sup_{\theta \in \Theta} \mathbb{E}_\theta[L(\theta, \hat{\theta})]$$

does not incoorporate complexity constraint

Example a: Gaussian mean model

Warm-up: scalar case

$$X_i = \theta + Z_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\theta, 1), i = 1, \ldots, n$$

Example a: Gaussian mean model

Warm-up: scalar case

$$X_i = \theta + Z_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\theta, 1), i = 1, \ldots, n$$

Then

$$\inf_{\widehat{\theta}} \sup_{\theta \in \mathbb{R}} \mathbb{E}|\widehat{\theta} - \theta|^2 = \frac{1}{n}$$

Example a: Gaussian mean model

Warm-up: scalar case

$$X_i = \theta + Z_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\theta, 1), i = 1, \ldots, n$$

Then

$$\inf_{\widehat{\theta}} \sup_{\theta \in \mathbb{R}} \mathbb{E}|\widehat{\theta} - \theta|^2 = \frac{1}{n}$$

Proof

- Upper bound: estimate by $\bar{X}$ (sufficient statistics/maximal likehood)

Example a: Gaussian mean model

Warm-up: scalar case

$$X_i = \theta + Z_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\theta, 1), i = 1, \dots, n$$

Then

$$\inf_{\widehat{\theta}} \sup_{\theta \in \mathbb{R}} \mathbb{E}|\widehat{\theta} - \theta|^2 = \frac{1}{n}$$

Proof

- Upper bound: estimate by $\bar{X}$ (sufficient statistics/maximal likehood)
- Lower bound: Consider the prior $\theta \sim \mathcal{N}(0, \sigma^2) \Rightarrow \text{MMSE} = \frac{\sigma^2}{1+\sigma^2 n} \xrightarrow{\sigma \to \infty} \frac{1}{n}$

Example a: Gaussian mean model

Warm-up: vector case

$$X_i = \theta + Z_i \in \mathcal{N}(\theta, \mathbf{I}_k), i = 1, \ldots, n$$

Then $\bar{X}$ is sufficient statistic and minimax:

$$\inf_{\widehat{\theta}} \sup_{\theta \in \mathbb{R}^k} \mathbb{E}\|\widehat{\theta} - \theta\|_2^2 = \frac{k}{n}$$

Example a: Gaussian mean model

Lemma (Anderson's lemma)

Any bowl-shaped (i.e., with symmetric convex sublevel sets, e.g., norm) $\rho : \mathbb{R}^n \to \mathbb{R}_+$,

$$\operatorname*{argmin}_{\theta \in \mathbb{R}^n} \mathbb{E}\rho(\theta + Z) = 0.$$

Example a: Gaussian mean model

Lemma (Anderson's lemma)

Any bowl-shaped (i.e., with symmetric convex sublevel sets, e.g., norm) $\rho : \mathbb{R}^n \to \mathbb{R}_+$,

$$\operatorname*{argmin}_{\theta \in \mathbb{R}^n} \mathbb{E}\rho(\theta + Z) = 0.$$

Proof: Brunn-Minkowski inequality

Example a: Gaussian mean model

Lemma (Anderson's lemma)

Any bowl-shaped (i.e., with symmetric convex sublevel sets, e.g., norm) $\rho : \mathbb{R}^n \to \mathbb{R}_+$,

$$\operatorname*{argmin}_{\theta \in \mathbb{R}^n} \mathbb{E}\rho(\theta + Z) = 0.$$

Proof: Brunn-Minkowski inequality

$$X_i = \theta + Z_i \in \mathcal{N}(\theta, \mathbf{I}), i = 1, \ldots, n$$

Then for any norm $\| \cdot \|$,

$$\inf_{\widehat{\theta}} \sup_{\theta \in \mathbb{R}^k} \mathbb{E}\|\widehat{\theta} - \theta\|^2 = \frac{\mathbb{E}\|Z\|^2}{n}$$

Ref: Le Cam (86)

Example a: Gaussian mean model

Noisy observation of a $k \times k$ matrix

$$X_i = M + Z_i \in \mathbb{R}^{k\times k}, i = 1, \ldots, n$$

Then

$$\inf_{\widehat{M}} \sup_{M \in \mathbb{R}^{k\times k}} \mathbb{E}\|\widehat{M} - M\|^2 = \frac{\|Z\|^2}{n}$$

Example a: Gaussian mean model

Noisy observation of a $k \times k$ matrix

$$X_i = M + Z_i \in \mathbb{R}^{k\times k}, i = 1, \ldots, n$$

Then

$$\inf_{\widehat{M}} \sup_{M\in\mathbb{R}^{k\times k}} \mathbb{E}\|\widehat{M} - M\|^2 = \frac{\|Z\|^2}{n}$$

What if the noise is non-Gaussian?

$$\inf_{\widehat{M}} \sup_{M\in\mathbb{R}^{k\times k}} \mathbb{E}\|\widehat{M} - M\|^2 = ?$$

Example a: Gaussian mean model

Noisy observation of a $k \times k$ matrix

$$X_i = M + Z_i \in \mathbb{R}^{k \times k}, i = 1, \ldots, n$$

Then

$$\inf_{\widehat{M}} \sup_{M \in \mathbb{R}^{k \times k}} \mathbb{E}\|\widehat{M} - M\|^2 = \frac{\|Z\|^2}{n}$$

What if the noise is non-Gaussian?

$$\inf_{\widehat{M}} \sup_{M \in \mathbb{R}^{k \times k}} \mathbb{E}\|\widehat{M} - M\|^2 \asymp ?$$

Example b: Covariance matrix estimation

Observe

$$X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \Sigma_{k\times k})$$

Example b: Covariance matrix estimation

Observe

$$X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \Sigma_{k\times k})$$

What we know

Sample covariance matrix $S = \frac{1}{n}\sum_{i=1}^{n} X_i X_i'$ is

- maximal likelihood & sufficient statistic
- NOT minimax under KL loss for $n \geq k$ [Stein '54, Eaton '70, ...]

Example b: Covariance matrix estimation

Observe

$$X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \Sigma_{k\times k})$$

What we know

Sample covariance matrix $S = \frac{1}{n}\sum_{i=1}^{n} X_i X_i'$ is

- maximal likelihood & sufficient statistic
- NOT minimax under KL loss for $n \geq k$ [Stein '54, Eaton '70, ...]

Little is known about

- minimaxity in high dimensions $k > n$?
- norm losses
- rate minimaxity of S?

Basic problem: Mean model

$$X = M + \frac{1}{\sqrt{n}} Z$$

Theorem

Suppose $Z_{ij} \overset{\text{i.i.d.}}{\sim} P$ *with zero mean and finite fourth moment, s.t.* $D(P(\cdot + \theta) \| P) \lesssim \theta^2$. *Then for all unitarily invariant* $\|\cdot\|$ *and all* k, n,

$$\inf_{\widehat{M}} \sup_{M \in \mathbb{R}^{k \times k}} \mathbb{E}\|\widehat{M} - M\|^2 \asymp \frac{1}{n} k \|\mathbf{I}\|^2$$

Basic problem: Mean model

$$X = M + \frac{1}{\sqrt{n}} Z$$

Theorem

Suppose $Z_{ij} \overset{i.i.d.}{\sim} P$ *with zero mean and finite fourth moment, s.t.* $D(P(\cdot + \theta)\|P) \lesssim \theta^2$. *Then for all unitarily invariant* $\|\cdot\|$ *and all* k, n,

$$\inf_{\widehat{M}} \sup_{M \in \mathbb{R}^{k \times k}} \mathbb{E}\|\widehat{M} - M\|^2 \asymp \frac{1}{n} k \|\mathbf{I}\|^2$$

Message

There is nothing significantly better than ML if

1. prior knowledge of the signal is absent
2. loss function is sufficiently symmetric

Preliminaries on convex geometry

Dual norm

- Dual norm of $\|\cdot\|$:

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

Dual norm

- Dual norm of $\|\cdot\|$:

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

- Example

$$(\ell_p \text{ norm})_* = \ell_q \text{ norm}, \quad \frac{1}{p} + \frac{1}{q} = 1$$

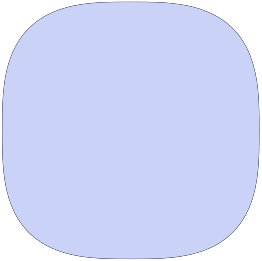
Convex body

- Polar of a convex body K:

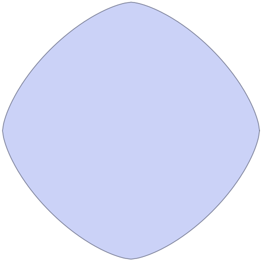
$$K^\circ \triangleq \left\{ y \in \mathbb{R}^d : \sup_{x \in K} \langle x, y \rangle \leq 1 \right\},$$

- Support function

$$\|x\|_K = \inf\{r > 0 : x \in rK\}$$



ℓ_p-ball



ℓ_q-ball

Convex body and polar

$$\|\cdot\| \xrightarrow{\text{unit ball}} B_{\|\cdot\|}$$

Convex body and polar

$$
\begin{array}{ccc}
\|\cdot\| & \xrightarrow{\text{unit ball}} & B_{\|\cdot\|} \\
& & \Big\downarrow \text{polar}
\end{array}
$$

Convex body and polar

$$\begin{array}{ccc} \|\cdot\| & \xrightarrow{\text{unit ball}} & B_{\|\cdot\|} \\ & & \Big\downarrow \text{polar} \\ \|\cdot\|_* & \xleftarrow{\text{support function}} & (B_{\|\cdot\|})_* \end{array}$$

Convex body and polar

$$\begin{array}{ccc} \|\cdot\| & \xrightarrow{\text{unit ball}} & B_{\|\cdot\|} \\ \Big\downarrow \text{dual} & & \Big\downarrow \text{polar} \\ \|\cdot\|_* & \xleftarrow{\text{support function}} & (B_{\|\cdot\|})_* \end{array}$$

Unitarily invariant norm

- A matrix norm $\|\cdot\|$ is unitarily invariant if

$$\|A\| = \|UAV\|, \quad \forall U, V: \text{ orthogonal}$$

Unitarily invariant norm

- A matrix norm $\|\cdot\|$ is unitarily invariant if

$$\|A\| = \|UAV\|, \quad \forall U, V: \text{ orthogonal}$$

- Representation theorem [von Neumann '37]

$$\|A\| = \tau(\underbrace{\sigma(A)}_{\text{singular values}})$$

τ: symmetric gauge function (i.e., a vector norm invariant to permutation and sign flips).

Unitarily invariant norm

- A matrix norm $\|\cdot\|$ is unitarily invariant if

$$\|A\| = \|UAV\|, \quad \forall U, V: \text{ orthogonal}$$

- Representation theorem [von Neumann '37]

$$\|A\|_\tau = \tau(\underbrace{\sigma(A)}_{\substack{\text{singular}\\ \text{values}}})$$

τ: symmetric gauge function (i.e., a vector norm invariant to permutation and sign flips).

Unitarily invariant norm

- A matrix norm $\|\cdot\|$ is unitarily invariant if

$$\|A\| = \|UAV\|, \quad \forall U, V: \text{ orthogonal}$$

- Representation theorem [von Neumann '37]

$$\|A\|_\tau = \tau(\underbrace{\sigma(A)}_{\text{singular values}})$$

  τ: symmetric gauge function (i.e., a vector norm invariant to permutation and sign flips).

- dual of $\|\cdot\|_\tau$ is $\|\cdot\|_{\tau_*}$

Examples

- Schatten norms:

$$\|\mathbf{A}\|_{\mathrm{S}_q} = \left(\sum_{i=1}^{k\wedge s} \sigma_i^q(\mathbf{A})\right)^{1/q}.$$

  - $q = 1$: nuclear norm
  - $q = 2$: Frobenius norm
  - $q = \infty$: spectral norm,

Examples

- Schatten norms:

$$\|\mathbf{A}\|_{\mathrm{S}_q} = \left(\sum_{i=1}^{k \wedge s} \sigma_i^q(\mathbf{A}) \right)^{1/q}.$$

  - $q = 1$: nuclear norm
  - $q = 2$: Frobenius norm
  - $q = \infty$: spectral norm,

- Ky Fan norms

$$\|\mathbf{A}\|_{(\ell)} = \sum_{i=1}^{\ell} \sigma_i(\mathbf{A}),$$

Volume: Upper estimates

Urysohn's Inequality

Let K be a symmetric convex body in $\mathbb{R}^d$. Then

$$\left(\frac{\mathrm{vol}(K)}{\mathrm{vol}(B_2)}\right)^{\frac{1}{d}} \le \frac{1}{\sqrt{d}}\,\mathbb{E}\sup_{y\in K}\langle G, y\rangle,$$

where $G \sim N(0, \mathbf{I})$ is standard Gaussian.

Volume: Upper estimates

Urysohn's Inequality

Let K be a symmetric convex body in $\mathbb{R}^d$. Then

$$\left(\frac{\mathrm{vol}(K)}{\mathrm{vol}(B_2)}\right)^{\frac{1}{d}} \leq \frac{1}{\sqrt{d}} \boxed{\mathbb{E} \sup_{y \in K} \langle G, y \rangle},$$

where $G \sim N(0, \mathbf{I})$ is standard Gaussian.

Gaussian width

Note: $\mathrm{vol}(B_2)^{\frac{1}{d}} \asymp \frac{1}{\sqrt{d}}$

Aside: Brunn-Minkowski $\Rightarrow$ Urysohn

- Brunn-Minkowski: $K \mapsto \mathrm{vol}(K)^{\frac{1}{d}}$ is concave

Ref: Pisier (99)

Aside: Brunn-Minkowski $\Rightarrow$ Urysohn

- Brunn-Minkowski: $K \mapsto \mathrm{vol}(K)^{\frac{1}{d}}$ is concave
- Random rotation $T \in O(d)$ + Jensen's inequality:

$$\mathrm{vol}(K)^{\frac{1}{d}} \geq \mathrm{vol}(\underbrace{\mathbb{E}T(K)}_{\text{ball:}B_2(\lambda)})^{\frac{1}{d}} = \lambda \mathrm{vol}(B_2)^{\frac{1}{d}}$$

Ref: Pisier (99)

Aside: Brunn-Minkowski $\Rightarrow$ Urysohn

- Brunn-Minkowski: $K \mapsto \mathrm{vol}(K)^{\frac{1}{d}}$ is concave
- Random rotation $T \in O(d)$ + Jensen's inequality:

$$\mathrm{vol}(K)^{\frac{1}{d}} \geq \mathrm{vol}(\underbrace{\mathbb{E}T(K)}_{\text{ball:}B_2(\lambda)})^{\frac{1}{d}} = \lambda \mathrm{vol}(B_2)^{\frac{1}{d}}$$

- Radius = mean width:

$$\lambda = \mathbb{E} \sup_{y \in K} \langle \theta, y \rangle ,$$

θ : uniform over sphere.

Ref: Pisier (99)

Mahler volume

[Bourgain-Milman '86, Kupenberg '08]

$$\frac{1}{2} \leq \left(\frac{\mathrm{vol}(K)\mathrm{vol}(K^\circ)}{\mathrm{vol}(B_2)^2} \right)^{\frac{1}{d}} \leq 1$$

Volume: Lower estimates

Inverse Santaló's inequality

$\exists$ universal constant c_0, s.t. for any symmetric convex body $K \subset \mathbb{R}^d$,

$$\mathrm{vol}(K)^{\frac{1}{d}} \geq \frac{c_0}{\mathbb{E}\|G\|_K}.$$

Proof

Vector result: general norm

Theorem

Consider

$$Y = \theta + Z \in \mathbb{R}^d.$$

Then

$$\frac{d^2}{(\mathbb{E}\|Z\|_*)^2} \lesssim \inf_{\hat{\theta}} \sup_{\theta \in \mathbb{R}^d} \mathbb{E}_\theta \|\hat{\theta} - \theta\|^2 \le \mathbb{E}\|Z\|^2,$$

where $\|\cdot\|_*$ *is the dual norm of* $\|\cdot\|$.

Vector result: general norm

Theorem

Consider

$$Y = \theta + Z \in \mathbb{R}^d.$$

Then

$$\frac{d^2}{(\mathbb{E}\|Z\|_*)^2} \lesssim \inf_{\hat{\theta}} \sup_{\theta \in \mathbb{R}^d} \mathbb{E}_\theta \|\hat{\theta} - \theta\|^2 \leq \mathbb{E}\|Z\|^2,$$

where $\|\cdot\|_*$ *is the dual norm of* $\|\cdot\|$.

Particularizing to matrix problem and $\|\cdot\|_\tau$, using random matrix theory,

$$\inf_{\widehat{M}} \sup_{M \in \mathbb{R}^{k\times k}} \mathbb{E}\|\widehat{M} - M\|_\tau^2 \asymp \frac{k}{n} \tau^2(\mathbf{1})$$

Information-theoretic determination of minimax rate

- [Ibragimov-Has'minskii '81]
- [Birgé '83]
- [Haussler-Opper '97]
- [Yang-Barron '99]
- ...

Reduction to multiple hypothesis testing

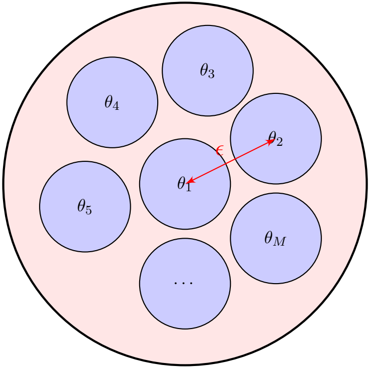
Intuition

testing is “easier” than estimation

Reduction to multiple hypothesis testing

Intuition

testing is “easier” than estimation



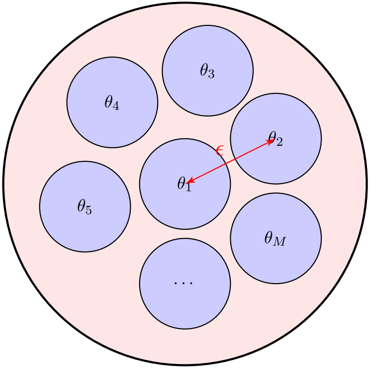
Find $\{\theta_1, \ldots \theta_M\} \subset \Theta$, s.t.

- $\|\theta_i - \theta_j\| \geq \epsilon, \forall i \neq j$.
- Based on data, any test fails w.p. 0.1 for the M-ary HT problem $H_i : \theta = \theta_i$.

Reduction to multiple hypothesis testing

Intuition

testing is "easier" than estimation



Find $\{\theta_1, \ldots \theta_M\} \subset \Theta$, s.t.

- $\|\theta_i - \theta_j\| \geq \epsilon, \forall i \neq j$.
- Based on data, any test fails w.p. 0.1 for the M-ary HT problem $H_i : \theta = \theta_i$.

Obtain a lower bound $\gtrsim \epsilon^2$

Fano's inequality

$$p_{\mathrm{e}} \geq 1 - \frac{I(\theta; X) + \log 2}{\log M}.$$

Fano's inequality

$$p_{\mathrm{e}} \geq 1 - \frac{I(\theta; X) + \log 2}{\log M}.$$

Let $\{\theta_1, \ldots \theta_M\}$ be a maximal ϵ-packing for Euclidean ball $B_2(\delta)$

- Upper bound I: Information radius $\leq$ diameter

$$I(\theta; X) = \inf_Q \frac{1}{M} \sum_{i=1}^{M} D(P_{\theta_i} || Q) \leq \frac{1}{2}\delta^2$$

Fano's inequality

$$p_e \geq 1 - \frac{I(\theta; X) + \log 2}{\log M}.$$

Let $\{\theta_1, \dots \theta_M\}$ be a maximal ϵ-packing for Euclidean ball $B_2(\delta)$

- Upper bound I: Information radius $\leq$ diameter

$$I(\theta; X) = \inf_Q \frac{1}{M} \sum_{i=1}^{M} D(P_{\theta_i} || Q) \leq \frac{1}{2}\delta^2$$

- Lower bound M: Gilbert-Varshamov

$$\begin{aligned} M &\geq N(\epsilon) && \text{[maximality]} \\ &\geq \frac{\mathrm{vol}(B_2(\delta))}{\mathrm{vol}(B_{\|\cdot\|}(\epsilon))} && \text{[union bound]} \\ &\geq \left(\frac{\delta\sqrt{d}}{\epsilon\,\mathbb{E}\|Z\|_*} \right)^d && \text{[Urysohn]} \end{aligned}$$

Volume ratio

What enters into the lower bound is

$$\frac{\mathrm{vol}(\text{KL neighborhood})}{\mathrm{vol}(\text{norm ball})}$$

model

loss

- Kullback-Leibler radius v.s. volume of K
- Extend far beyond normal mean problem (covariance model, exponential family)

Covariance matrix estimation

Theorem

Observe

$$X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \Sigma_{k \times k})$$

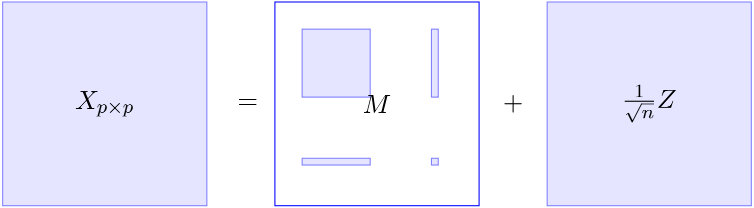
Then

$$\inf_{\hat{\Sigma}} \sup_{\|\Sigma\|_{\mathrm{op}} \leq \lambda} \mathbb{E}\|\hat{\Sigma} - \Sigma\|_\tau^2 \asymp \left(\frac{k}{n} \wedge 1\right) \lambda^2 \tau^2(\mathbf{1}).$$

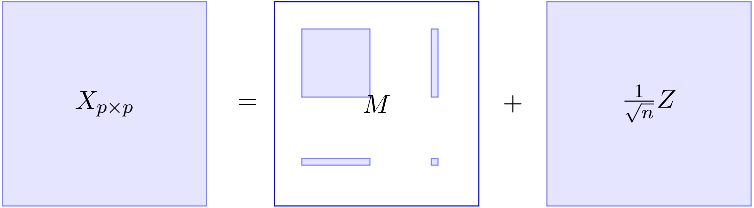
Sample covariance matrix is minimax rate-optimal.

Structured problems

Denoising under submatrix sparsity



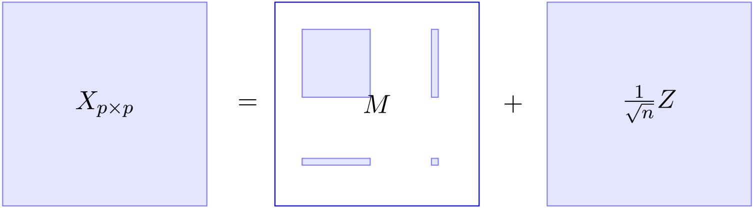
Denoising under submatrix sparsity



Connections

- Biclustering of microarray data [Sun-Nobel '13]
- Group sparsity [Lounici et al, '11]
- Community detection [Arias-Castro-Verzelen '13]
- Sparse PCA and rank detection [Johnstone-Lu '09]

Denoising under submatrix sparsity



Intuitions

- Ambient dimension $p \times p \to$ intrinsic dimension $k \times k$,
- Rule of thumb:
  can achieve the risk on the reduced dimension + some penalty

Minimax rate

Theorem

$$\inf_{\widehat{M}} \sup_{M} \mathbb{E}\|\widehat{M} - M\|^2 \asymp \underbrace{\frac{k}{n}\tau^2(\mathbf{1})}_{\textit{oracle risk}} + \underbrace{\mathrm{Lip}(\tau)^2 \frac{k}{n} \log \frac{ep}{k}}_{\textit{excess risk}}$$

where $\mathrm{Lip}(\tau) \triangleq \sup_{x \neq 0} \frac{\tau(x)}{\|x\|_2}$

- Oracle risk: Risk of an oracle estimator knowing $\mathrm{supp}(M)$
- Excess risk: Risk due to combinatorial uncertainty about $\mathrm{supp}(M)$

Minimax rate

Theorem

$$\inf_{\widehat{M}} \sup_{M} \mathbb{E}\|\widehat{M} - M\|^2 \asymp \underbrace{\frac{k}{n}\tau^2(\mathbf{1})}_{\textit{oracle risk}} + \underbrace{\mathrm{Lip}(\tau)^2 \frac{k}{n} \log \frac{ep}{k}}_{\textit{excess risk}}$$

where $\mathrm{Lip}(\tau) \triangleq \sup_{x \neq 0} \frac{\tau(x)}{\|x\|_2}$

- Oracle risk: Risk of an oracle estimator knowing $\mathrm{supp}(M)$
- Excess risk: Risk due to combinatorial uncertainty about $\mathrm{supp}(M)$

Dependence on the norm is fundamentally different

Remarks

- Oracle risk lower bound: done
- Excess risk lower bound: probabilistic construction of packing based on $\mathrm{Lip}(\tau)$
- Achievability: combinatorial procedure – complexity $\binom{p}{k}^4$

Remarks

- Oracle risk lower bound: done
- Excess risk lower bound: probabilistic construction of packing based on $\mathrm{Lip}(\tau)$
- Achievability: combinatorial procedure – complexity $\binom{p}{k}^4$

Question

- Is it achievable computationally efficient procedures ?
- Can we mimic the combinatorial algorithms (e.g., convex relaxation)?

Computational barrier

Triumphs in vector problems

- Denoising under vector sparsity:
  - $y = \theta + z$, $\theta \in \mathbb{R}^p$ is k-spare.
  - minimax rate $= k \log \frac{ep}{k}$
  - optimal procedure: entrywise thresholding – linear complexity

Triumphs in vector problems

- Denoising under vector sparsity:
  - $y = \theta + z$, $\theta \in \mathbb{R}^p$ is k-spare.
  - minimax rate $= k \log \frac{ep}{k}$
  - optimal procedure: entrywise thresholding – linear complexity
- Sparse linear regression/Compressed sensing
  - $y = A\theta + z$.
  - optimal procedure via convex prgramming: LASSO, Dantzig selector, etc. – polynomial complexity

Denoising with submatrix sparsity: Schatten norm

$$\inf_{\widehat{M}} \sup_{M} \mathbb{E}\|\widehat{M} - M\|_{\mathrm{S}_q}^2 \asymp \tilde{\Theta}\left(\frac{k^{2/q+1}}{n}\right)$$

Denoising with submatrix sparsity: Schatten norm

$$\inf_{\widehat{M}} \sup_{M} \mathbb{E}\|\widehat{M} - M\|_{\mathrm{S}_q}^2 \asymp \tilde{\Theta}\left(\frac{k^{2/q+1}}{n}\right)$$

- $1 \le q \le 2$: minimax rate is attained in linear time (entrywise thresholding)
- $2 < q \le \infty$: no efficient procedure can attain the minimax rates

Denoising with submatrix sparsity: Schatten norm

$$\inf_{\widehat{M}} \sup_{M} \mathbb{E}\|\widehat{M} - M\|_{\mathrm{S}_q}^2 \asymp \tilde{\Theta}\left(\frac{k^{2/q+1}}{n}\right)$$

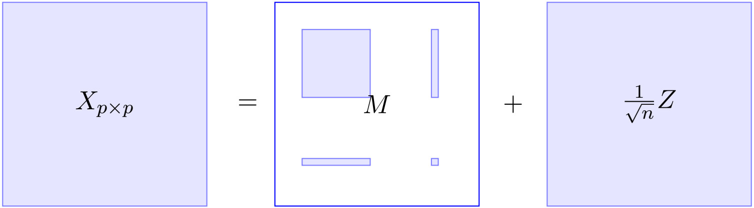
- $1 \leq q \leq 2$: minimax rate is attained in linear time (entrywise thresholding)
- $2 < q \leq \infty$: no efficient procedure can attain the minimax rates

Punchline

- no computationally efficient algorithm can harness the two-dimensional structure
- the best one can do is to treat it one-dimensionally (entrywise thresholding)

Complexity-constrained statistical inference

Submatrix detection/Biclustering/Community detection



An intriging question...

S. Balakrishnan, M. Kolar, A. Rinaldo, A. Singh, and L. Wasserman. "Statistical and computational tradeoffs in biclustering." In *NIPS 2011 Workshop on Computational Trade-offs in Statistical Learning*, 2011.

The biclustering problem highlights the tradeoff between computational complexity and statistical efficiency. The most significant open question with respect to our work is: "Is there a computationally efficient algorithm that achieves the minimax rate for all tuples (n, k, μ)?."

While we conjecture that the biclustering problem is computationally hard, the structure and randomness pose significant obstacles to the direct application of reductions to show hardness. In the biclustering problem we are given a particular *structured, random* (and *not* arbitrary, worst-case) instance of a known NP-hard problem. Showing that even these seemingly benign instances are not significantly easier than the worst-case instances is an important direction for future work.

Our work also highlights an important shortcoming of minimax analysis with regards to computational tractability. Ideally rather than being defined as an infimum over *all* estimators we would like to be able to define the minimax rate over a smaller class of all *efficiently* computable estimators, and develop tools to study this restricted minimax rate. Formalizing this notion is also an important direction of future work.

Clique problem

κ-CLIQUE: determine a N-vertex graph has a clique of size κ

- NP-complete
- Worst-case complexity $N^{\Theta(\kappa)}$ assuming $\mathrm{P} \neq \mathrm{NP}$.
- What about the average-case complexity?

Planted Clique problem

$$H_0 : \mathcal{G}(N, 1/2), \quad \text{versus} \quad H_1 : \mathcal{G}(N, 1/2, \kappa),$$

- statistically impossible if $\kappa = o(\log N)$
- greedy algorithm works if $\kappa = \Omega(\sqrt{N \log N})$
- spectral methods works if $\kappa = \Omega(\sqrt{N})$ [Alon-Krivelevich-Sudakov '98]

Planted Clique problem

$$H_0 : \mathcal{G}(N, 1/2), \quad \text{versus} \quad H_1 : \mathcal{G}(N, 1/2, \kappa),$$

Intermediate regime: $\log N \ll \kappa \ll \sqrt{N}$

- average-case hardness proved under some computational model [Rossman '10, Feldman et al. '13]
- widely believed to have high complexity

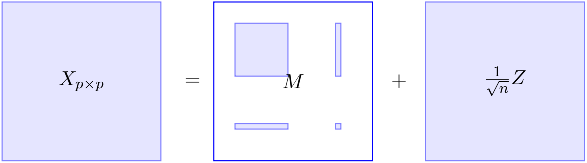
Planted Clique problem

$$H_0 : \mathcal{G}(N, 1/2), \quad \text{versus} \quad H_1 : \mathcal{G}(N, 1/2, \kappa),$$

Intermediate regime: $\log N \ll \kappa \ll \sqrt{N}$

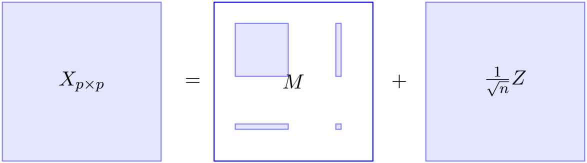
- average-case hardness proved under some computational model [Rossman '10, Feldman et al. '13]
- widely believed to have high complexity
- many hardness results assuming Planted Clique hardness
  - cryptography [Juels-Peinado '00]
  - independence testing [Alon et al. '07]
  - approximating Nash equilibrium [Hazan-Krauthgamer '11]
  - Certifying restricted isometry property [Koiran-Zouzias, '12]
  - Detecting sparse principle components [Berthet-Rigollet '13]
  - Sparse + low-rank matrix decomposition [Chen '13]

Submatrix detection



$$H_0 : M = 0, \quad \text{versus} \quad H_1 : \frac{1}{k^2}\sum_{ij} M_{ij} \geq \lambda.$$

Submatrix detection



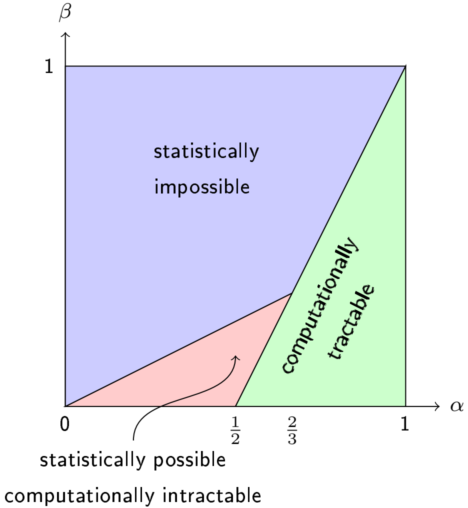
$$H_0 : M = 0, \quad \text{versus} \quad H_1 : \frac{1}{k^2}\sum_{ij} M_{ij} \geq \lambda.$$

- Assuming Planted Clique hardness

$$\underbrace{\lambda^*}_{\text{minimax}} = \Theta\left(\sqrt{\frac{1}{k}\log\frac{ep}{k}} \wedge \frac{p}{k^2}\right) \leq \underbrace{\lambda^\sharp}_{\text{computable}} = \tilde{\Theta}\left(1 \wedge \frac{p}{k^2}\right).$$

- For λ below $\lambda^\sharp$: can be reduced from Planted Clique in randomized polynomial time

Detectable region: $k = p^\alpha$ and $\lambda = p^{-\beta}$



Wait a minute..

- Test $\phi : \mathbb{R}^{p \times p} \to \{0, 1\}$ – computational complexity is ill-defined
- ~~flops~~
- Statistical computing system has finite precision

A Paradigm

We need a proxy:

- preseves the statistical difficulty of the original experiment
- computational complexity of inference procedures is well-defined

Asymptotic equivalence of discretized models

$$X = M + Z_{p\times p} \xrightarrow{\text{quantization}} [X]_t = 2^{-t}\lfloor 2^t X \rfloor.$$

Theorem

Le Cam distance satisfies:

$$\Delta(\{P_X : M \in \mathbb{R}^{p\times p}\}, \{P_{[X]_t} : M \in \mathbb{R}^{p\times p}\}) \leq p2^{-t/3+2}.$$

Then $t \geq (3+\epsilon)\log_2 p \overset{p\to\infty}{\Longrightarrow}$ *asymptotic equivalence*

Concluding remarks

- Unstructured problems
  - rate minimaxity of ML estimator
  - determining minimax rates using convex geometry and information theory

Concluding remarks

- Unstructured problems
  - rate minimaxity of ML estimator
  - determining minimax rates using convex geometry and information theory
- Exploit extra structure (intrinsic low-dimensionality)

Concluding remarks

- Unstructured problems
  - rate minimaxity of ML estimator
  - determining minimax rates using convex geometry and information theory
- Exploit extra structure (intrinsic low-dimensionality)
  - Some matrix problems elude efficient algorithms
  - Hardness depends on the model and loss function (Why?)
  - trade-off between computational efficiency and statistical optimality

Concluding remarks

- Unstructured problems
  - rate minimaxity of ML estimator
  - determining minimax rates using convex geometry and information theory
- Exploit extra structure (intrinsic low-dimensionality)
  - Some matrix problems elude efficient algorithms
  - Hardness depends on the model and loss function (Why?)
  - trade-off between computational efficiency and statistical optimality

References

- Z. Ma & W. (2013). *Volume Ratio, Sparsity, and Minimaxity under Unitarily Invariant Norms*. `arXiv:1306.3609`.
- Z. Ma & W. (2013). *Computational Barriers in Minimax Submatrix Detection*. `arXiv:1309.5914`.