Tutorial Part I: Information theory meets machine learning

Emmanuel Abbe
UC Berkeley

Martin Wainwright
Princeton University

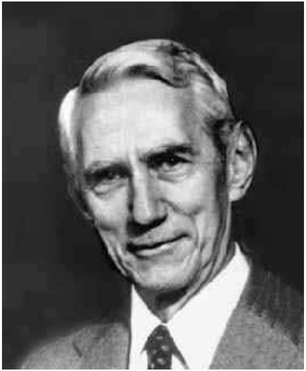
Introduction

Era of massive data sets

Fascinating problems at the interfaces between information theory and statistical machine learning.

1. Fundamental issues
   - Concentration of measure: high-dimensional problems are remarkably predictable
   - Curse of dimensionality: without structure, many problems are hopeless
   - Low-dimensional structure is essential
2. Machine learning brings in algorithmic components
   - Computational constraints are central
   - Memory constraints: need for distributed and decentralized procedures
   - Increasing importance of privacy

Historical perspective: info. theory and statistics



Claude Shannon

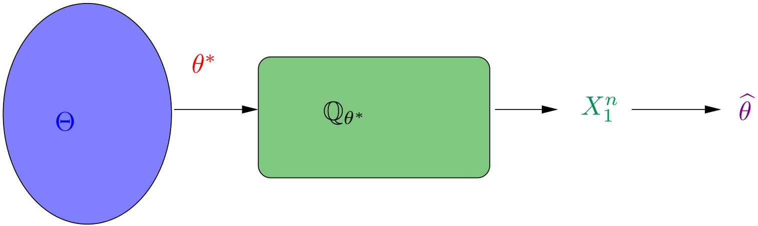
Andrey Kolmogorov

A rich intersection between information theory and statistics

1. Hypothesis testing, large deviations
2. Fisher information, Kullback-Leibler divergence
3. Metric entropy and Fano's inequality

Statistical estimation as channel coding

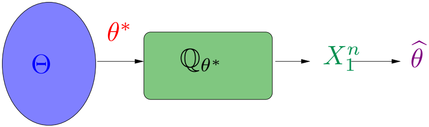
- Codebook: indexed family of probability distributions $\{\mathbb{Q}_\theta \mid \theta \in \Theta\}$
- Codeword: nature chooses some $\theta^* \in \Theta$



- Channel: user observes n i.i.d. draws $X_i \sim \mathbb{Q}_{\theta^*}$
- Decoding: estimator $X_1^n \mapsto \widehat{\theta}$ such that $\widehat{\theta} \approx \theta^*$

Statistical estimation as channel coding

- Codebook: indexed family of probability distributions $\{\mathbb{Q}_\theta \mid \theta \in \Theta\}$
- Codeword: nature chooses some $\theta^* \in \Theta$



- Channel: user observes n i.i.d. draws $X_i \sim \mathbb{Q}_{\theta^*}$
- Decoding: estimator $X_1^n \mapsto \widehat{\theta}$ such that $\widehat{\theta} \approx \theta^*$

Perspective dating back to Kolmogorov (1950s) with many variations:

- codebooks/codewords: graphs, vectors, matrices, functions, densities....
- channels: random graphs, regression models, elementwise probes of vectors/machines, random projections
- closeness $\widehat{\theta} \approx \theta^*$: exact/partial graph recovery in Hamming, ℓ_p-distances, $L^(\mathbb{Q})$-distances, sup-norm etc.

Machine learning: algorithmic issues to forefront!

1. Efficient algorithms are essential
   - only (low-order) polynomial-time methods can ever be implemented
   - trade-offs between computational complexity and performance?
   - fundamental barriers due to computational complexity?

2. Distributed procedures are often needed
   - many modern data sets: too large to stored on a single machine
   - need methods that operate separately on pieces of data
   - trade-offs between decentralization and performance?
   - fundamental barriers due to communication complexity?

3. Privacy and access issues
   - conflicts between individual privacy and benefits of aggregation?
   - principled information-theoretic formulations of such trade-offs?

Part I of tutorial: Three vignettes

1. Graphical model selection

2. Sparse principal component analysis

3. Structured non-parametric regression and minimax theory

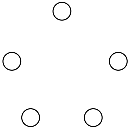
Vignette A: Graphical model selection

Simple motivating example: Epidemic modeling

Disease status of person s: $$x_s = \begin{cases} +1 & \text{if individual } s \text{ is infected} \\ -1 & \text{if individual } s \text{ is healthy} \end{cases}$$

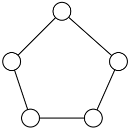
(1) Independent infection

$$\mathbb{Q}(x_1, \ldots, x_5) \propto \prod_{s=1}^{5} \exp(\theta_s x_s)$$



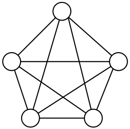
(2) Cycle-based infection

$$\mathbb{Q}(x_1, \ldots, x_5) \propto \prod_{s=1}^{5} \exp(\theta_s x_s) \prod_{(s,t) \in C} \exp(\theta_{st} x_s \, x_t)$$

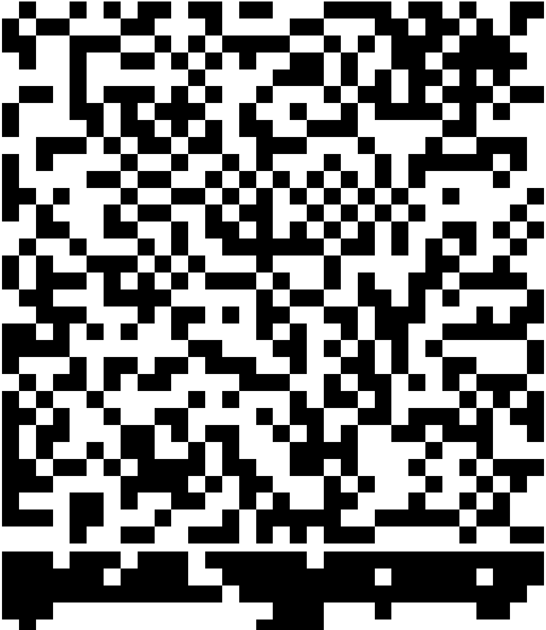


(3) Full clique infection

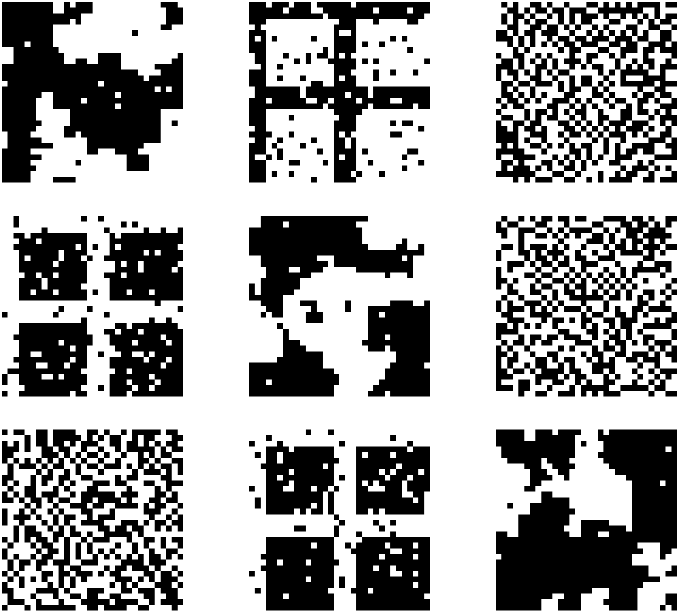
$$\mathbb{Q}(x_1, \ldots, x_5) \propto \prod_{s=1}^{5} \exp(\theta_s x_s) \prod_{s \neq t} \exp(\theta_{st} x_s x_t)$$



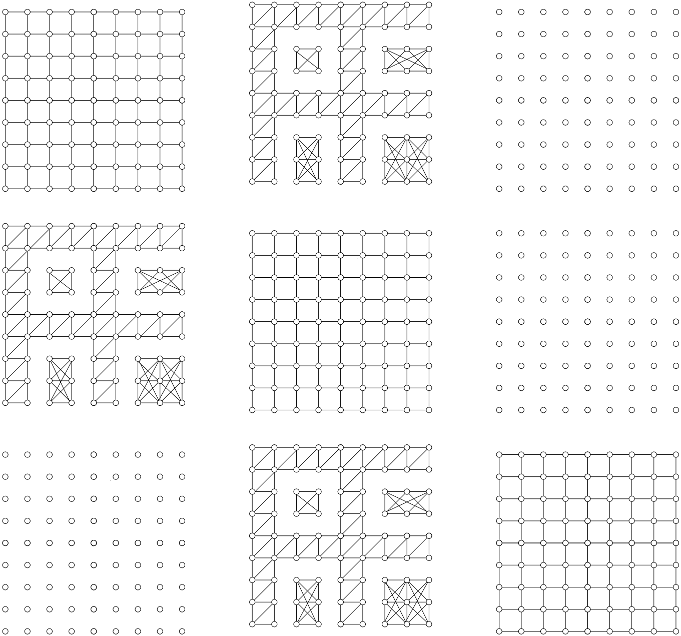
Possible epidemic patterns (on a square)



From epidemic patterns to graphs



Underlying graphs



Model selection for graphs

- drawn n i.i.d. samples from

$$\mathbb{Q}(x_1, \ldots, x_p; \Theta) \propto \; \exp\Big\{ \sum_{s \in V} \theta_s x_s + \sum_{(s,t) \in E} \theta_{st} x_s x_t \Big\}$$

- graph G and matrix $[\Theta]_{st} = \theta_{st}$ of edge weights are unknown

- data matrix $\mathbf{X}_1^n \in \{-1, +1\}^{n \times p}$

- want estimator $\mathbf{X}_1^n \mapsto \widehat{G}$ to minimize error probability

$$\underbrace{\mathbb{Q}^n \big[\widehat{G}(\mathbf{X}_1^n) \neq G\big]}_{\text{Prob. that estimated graph differs from truth}}$$

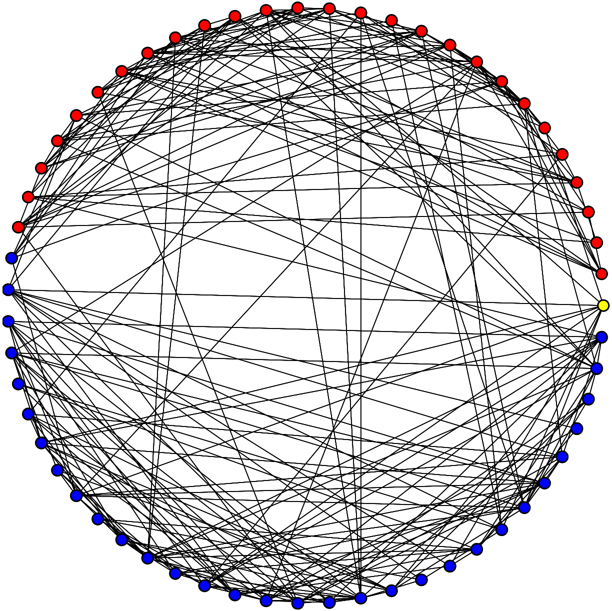
Channel decoding:

Think of graphs as codewords, and the graph family as a codebook.

Past/on-going work on graph selection

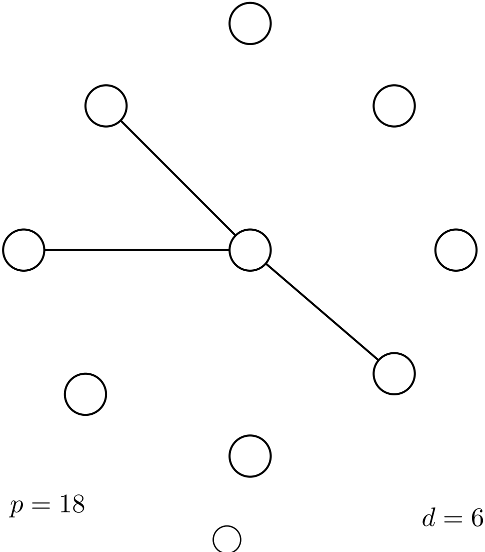
- exact polynomial-time solution on trees (Chow & Liu, 1967)
- testing for local conditional independence relationships (e.g., Spirtes et al, 2000; Kalisch & Buhlmann, 2008)
- pseudolikelihood and BIC criterion (Csiszar & Talata, 2006)
- pseudolikelihood and ℓ_1-regularized neighborhood regression
  - Gaussian case (Meinshausen & Buhlmann, 2006)
  - Binary case (Ravikumar, W. & Lafferty et al., 2006)
- various greedy and related methods: (Bresler et al., 2008; Bresler, 2015; Netrapalli et al., 2010; Anandkumar et al., 2013)
- lower bounds and inachievability results
  - information-theoretic bounds (Santhanam & W., 2012)
  - computational lower bounds (Dagum & Luby, 1993; Bresler et al., 2014)
  - phase transitions and performance of neighborhood regression (Bento & Montanari, 2009)

US Senate network (2004–2006 voting)



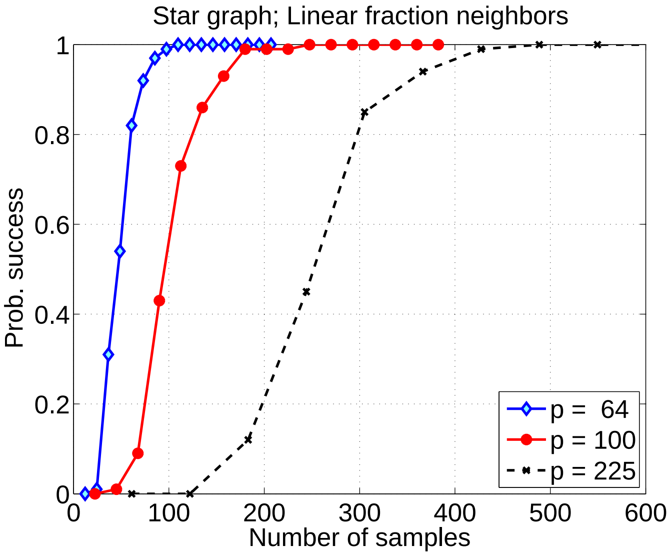
Experiments for sequences of star graphs

$p = 9$ $d = 3$



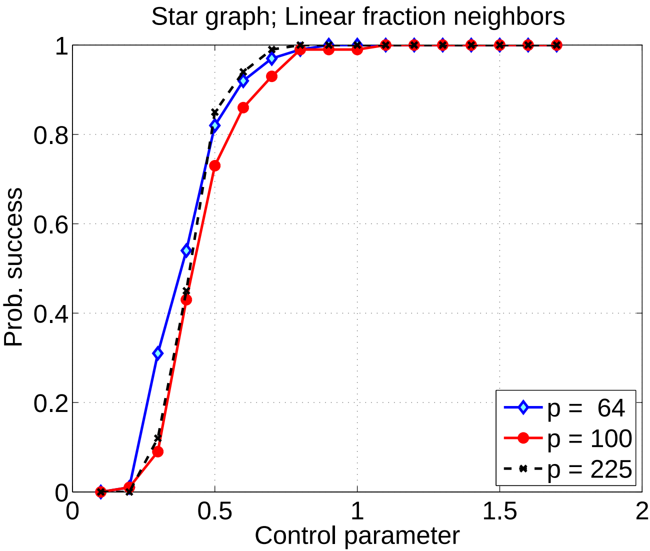
$p = 18$ $d = 6$

Empirical behavior: Unrescaled plots



Plots of success probability versus raw sample size n.

Empirical behavior: Appropriately rescaled



Plots of success probability versus rescaled sample size $\frac{n}{d^2 \log p}$

Some theory: Scaling law for graph selection

- graphs $G_{p,d}$ with p nodes and maximum degree d
- minimum absolute weight $\theta_{\min}$ on edges
- how many samples n needed to recover the unknown graph?

Theorem (Ravikumar, W. & Lafferty, 2010; Santhanam & W., 2012)

Achievable result: *Under regularity conditions, for a graph estimate* $\widehat{G}$ *produced by ℓ_1-regularized logistic regression:*

$$\underbrace{n > c_u \left(d^2 + 1/\theta_{\min}^2\right) \log p}_{\textit{Lower bound on sample size}} \quad \Longrightarrow \quad \underbrace{\mathbb{Q}[\widehat{G} \neq G] \to 0}_{\textit{Vanishing probability of error}}$$

Necessary condition: *For graph estimate* $\widetilde{G}$ *produced by any algorithm.*

$$\underbrace{n < c_\ell \left(d^2 + 1/\theta_{\min}^2\right) \log p}_{\textit{Upper bound on sample size}} \quad \Longrightarrow \quad \underbrace{\mathbb{Q}[\widetilde{G} \neq G] \geq 1/2}_{\textit{Constant probability of error}}$$

Information-theoretic analysis of graph selection

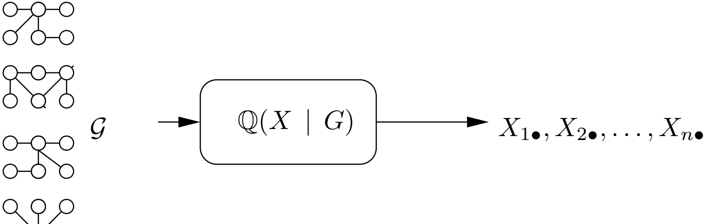
Question:

How to prove lower bounds on graph selection methods?

Answer:

Graph selection is an *unorthodox* channel coding problem.

- codewords/codebook: graph G in some graph class $\mathcal{G}$
- channel use: draw sample $X_{i\bullet} = (X_{i1}, \ldots, X_{ip}) \in \{-1, +1\}^p$ from the graph-structured distribution $\mathbb{Q}_G$
- decoding problem: use n samples $\{X_{1\bullet}, \ldots, X_{n\bullet}\}$ to correctly distinguish the "codeword"



$\mathcal{G}$

$\mathbb{Q}(X \mid G)$

$X_{1\bullet}, X_{2\bullet}, \ldots, X_{n\bullet}$

Proof sketch: Main ideas for necessary conditions

- based on assessing difficulty of graph selection over various sub-ensembles $\mathcal{G} \subseteq \mathcal{G}_{p,d}$

- choose $G \in \mathcal{G}$ u.a.r., and consider multi-way hypothesis testing problem based on the data $\mathbf{X}_1^n = \{X_{1\bullet}, \ldots, X_{n\bullet}\}$

- for any graph estimator $\psi : \mathcal{X}^n \to \mathcal{G}$, Fano's inequality implies that

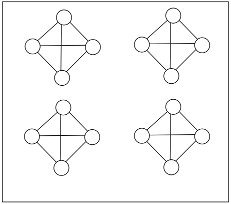
$$\mathbb{Q}[\psi(\mathbf{X}_1^n) \neq G] \geq 1 - \frac{I(\mathbf{X}_1^n; G)}{\log |\mathcal{G}|} - o(1)$$

  where $I(\mathbf{X}_1^n; G)$ is mutual information between observations $\mathbf{X}_1^n$ and randomly chosen graph G

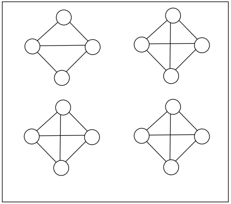
- remaining steps:

  1. Construct "difficult" sub-ensembles $\mathcal{G} \subseteq \mathcal{G}_{p,d}$
  2. Compute or lower bound the log cardinality $\log |\mathcal{G}|$.
  3. Upper bound the mutual information $I(\mathbf{X}_1^n; G)$.

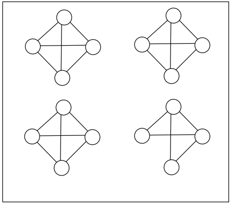
A “hard” d-clique ensemble



Base graph $\bar{G}$



Graph G^{uv}



Graph G^{st}

1. Divide the vertex set V into $\lfloor \frac{p}{d+1} \rfloor$ groups of size $d+1$, and form the base graph $\bar{G}$ by making a $(d+1)$-clique $\mathcal{C}$ within each group.
2. Form graph G^{uv} by deleting edge (u, v) from $\bar{G}$.
3. Consider testing problem over family of graph-structured distributions $\{\mathbb{Q}(\cdot\,; G^{st}),\ (s,t) \in \mathcal{C}\}$.

Why is this ensemble “hard”?

Kullback-Leibler divergence between pairs decays exponentially in d, unless minimum edge weight decays as $1/d$.

Vignette B: Sparse principal components analysis

Principal component analysis:

- widely-used method for {dimensionality reduction, data compression etc.}
- extracts top eigenvectors of sample covariance matrix $\widehat{\Sigma}$
- classical PCA in p dimensions: inconsistent unless $p/n \to 0$
- low-dimensional structure: many applications lead to sparse eigenvectors

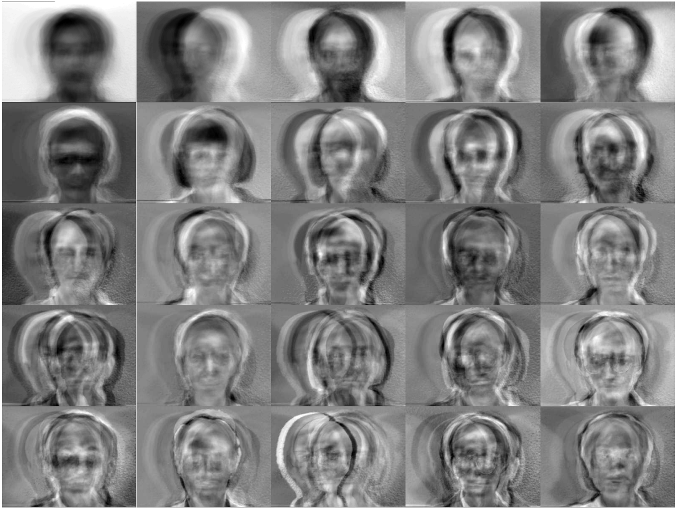
Population model: Rank-one spiked covariance matrix

$$\Sigma = \underbrace{\nu}_{\text{SNR parameter}} \quad \underbrace{\theta^*(\theta^*)^T}_{\text{rank-one spike}} \quad + I_p$$

Sampling model: Draw n i.i.d. zero-mean vectors with $\text{cov}(X_i) = \Sigma$, and form sample covariance matrix

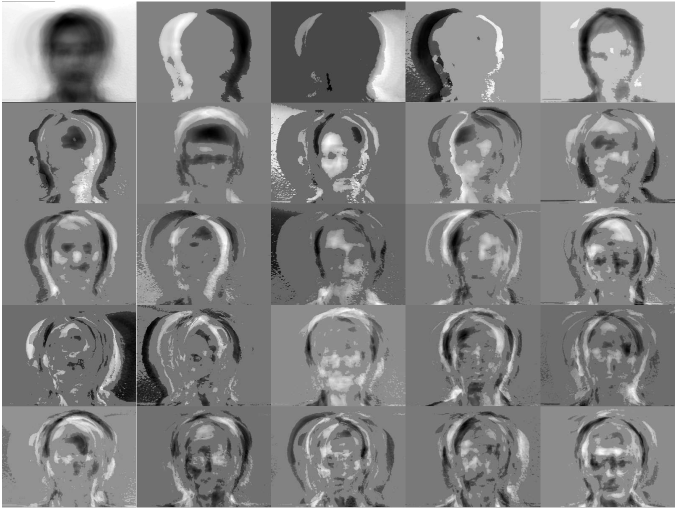
$$\widehat{\Sigma} := \quad \underbrace{\frac{1}{n}\sum_{i=1}^{n} X_i X_i^T}_{p\text{-dim. matrix, rank } \min\{n,p\}}$$

Example: PCA for face compression/recognition



First 25 standard principal components (estimated from data)

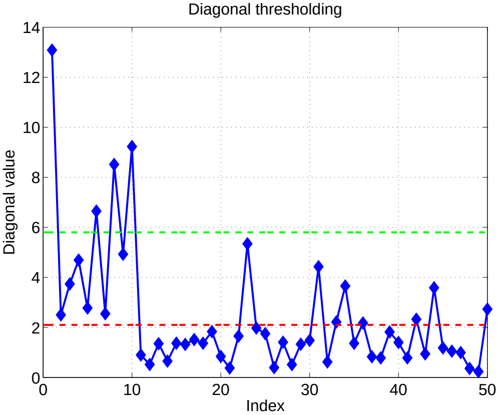
Example: PCA for face compression/recognition



First 25 sparse principal components (estimated from data)

Perhaps the simplest estimator....

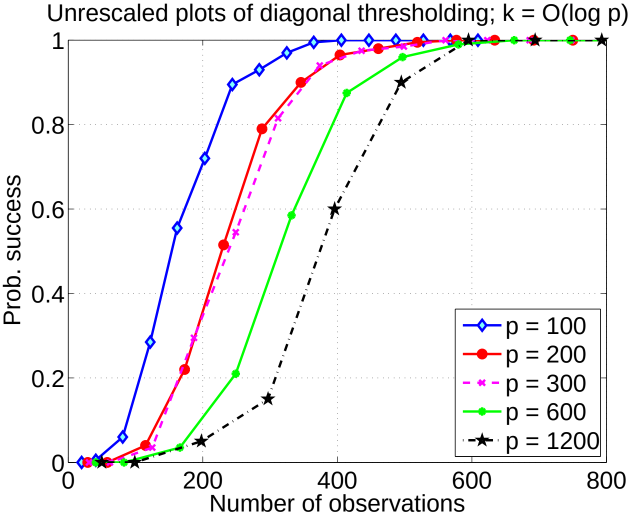
Diagonal thresholding: (Johnstone & Lu, 2008)



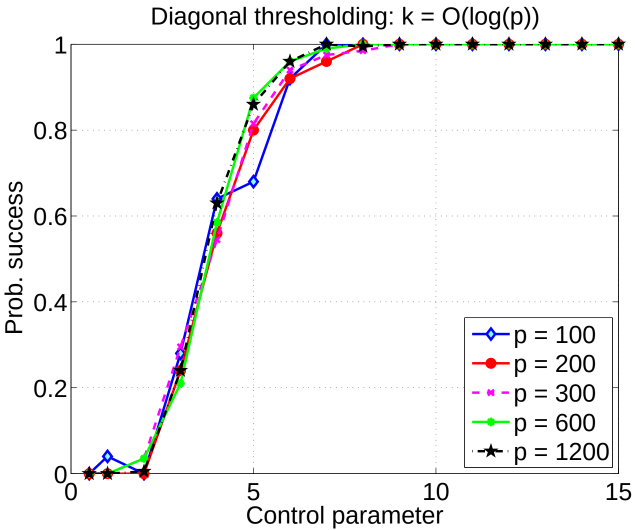
Given n i.i.d. samples X_i with zero mean, and with spiked covariance $\Sigma = \nu\theta^*(\theta^*)^T + I$:

1. Compute diagonal entries of sample covariance: $\widehat{\Sigma}_{jj} = \frac{1}{n}\sum_{i=1}^n X_{ij}^2$.
2. Apply threshold to vector $\{\widehat{\Sigma}_{jj}, j = 1, \ldots, p\}$

Diagonal thresholding: unrescaled plots



Diagonal thresholding: rescaled plots



Scaling is quadratic in sparsity: $\frac{n}{k^2 \log p}$

Diagonal thresholding and fundamental limit

Consider spiked covariance matrix

$$\Sigma = \underbrace{\nu}_{\text{Signal-to-noise}} \quad \underbrace{\theta^*(\theta^*)^T}_{\text{rank one spike}} \quad + I_{p \times p} \qquad \text{where} \quad \underbrace{\theta \in \mathbb{B}_0(k) \cap \mathbb{B}_2(1)}_{k\text{-sparse and unit norm}}$$

Theorem (Amini & W., 2009)

(a) *There are thresholds $\tau_\ell^{DT} \leq \tau_u^{DT}$ such that*

$$\underbrace{\frac{n}{k^2 \log p} \leq \tau_\ell^{DT}}_{\textit{DT fails w.h.p.}} \qquad \underbrace{\frac{n}{k^2 \log p} \geq \tau_u^{DT}}_{\textit{DT succeeds w.h.p.}}$$

(b) *For optimal method (exhaustive search):*

$$\underbrace{\frac{n}{k \log p} < \tau^{ES}}_{\textit{Fail w.h.p.}} \qquad \underbrace{\frac{n}{k \log p} > \tau^{ES}}_{\textit{Succeed w.h.p.}}.$$

One polynomial-time SDP relaxation

Recall Courant-Fisher variational formulation:

$$\theta^* = \arg\max_{\|z\|_2=1} \Big\{ z^T \underbrace{\Big(\nu\theta^*(\theta^*)^T + I_{p\times p}\Big)}_{\text{Population covariance } \Sigma} z \Big\}.$$

Equivalently, lifting to matrix variables $Z = zz^T$:

$$\theta\theta^T = \arg \max_{\substack{Z\in\mathbb{R}^{p\times p} \\ Z=Z^T, \quad Z\succeq 0}} \Big\{ \text{trace}(Z^T\Sigma) \Big\} \qquad \text{s.t. trace}(Z) = 1, \text{ and } \text{rank}(Z) = 1$$

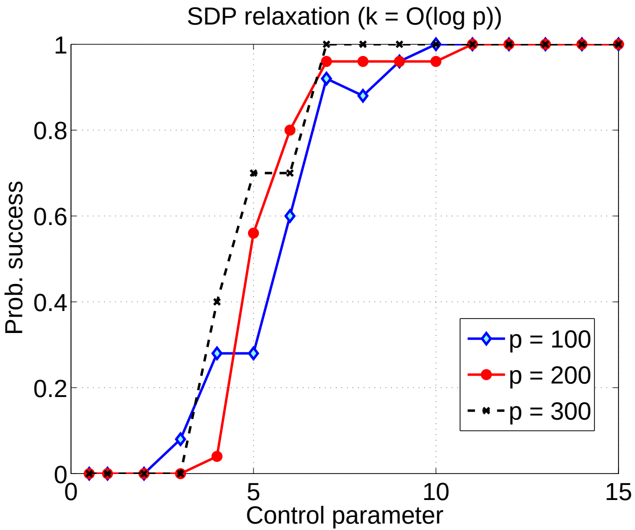
Dropping rank constraint yields a standard SDP relaxation:

$$\widehat{Z}^T = \arg \max_{\substack{Z\in\mathbb{R}^{p\times p} \\ Z=Z^T, \quad Z\succeq 0}} \Big\{ \text{trace}(Z^T\Sigma) \Big\} \qquad \text{s.t. trace}(Z) = 1.$$

In practice: (d'Aspremont et al., 2008)

- apply this relaxation using the sample covariance matrix $\widehat{\Sigma}$
- add the ℓ_1-constraint $\sum_{i,j=1}^{p} |Z_{ij}| \leq k^2$.

Phase transition for SDP: logarithmic sparsity



Scaling is linear in sparsity: $\frac{n}{k \log p}$

A natural question

Questions:

Can logarithmic sparsity or rank one condition be removed?

Computational lower bound for sparse PCA

Consider spiked covariance matrix

$$\Sigma = \underbrace{\nu}_{\text{signal-to-noise}} \underbrace{(\theta^*)(\theta^*)^T}_{\text{rank one spike}} + I_{p\times p} \quad \text{where} \quad \underbrace{\theta^* \in \mathbb{B}_0(k) \cap \mathbb{B}_2(1)}_{k\text{-sparse and unit norm}}$$

Sparse PCA detection problem:

$$\begin{array}{lll} \mathrm{H}_0 & \text{(no signal):} & X_i \sim \mathcal{D}(0, I_{p\times p}) \\ \mathrm{H}_1 & \text{(spiked signal):} & X_i \sim \mathcal{D}(0, \Sigma). \end{array}$$

Distribution $\mathcal{D}$ with sub-exponential tail behavior.

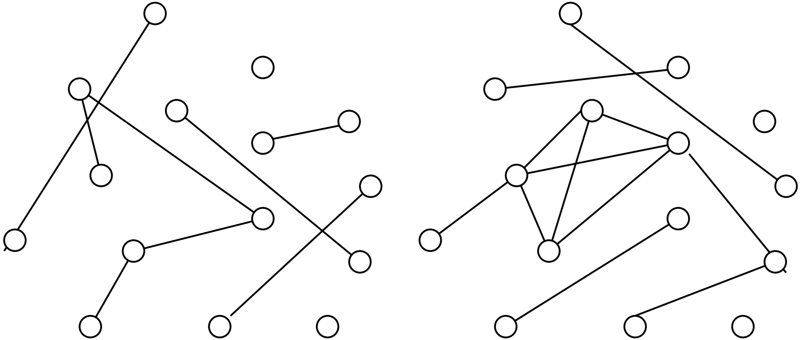
Theorem (Berthet & Rigollet, 2013)

Under average-case hardness of planted clique, polynomial-minimax level of detection ν_{POLY} *is given by*

$$\nu_{\text{POLY}} \asymp \frac{k^2 \log p}{n} \qquad \textit{for all sparsity } \log p \ll k \ll \sqrt{p}$$

Classical minimax level $\nu_{\text{OPT}} \asymp \frac{k \log p}{n}$.

Planted k-clique problem



Erdos-Renyi

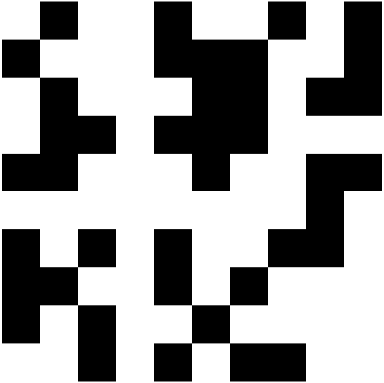
Planted k-clique

Binary hypothesis test based on observing random graph G on p-vertices:

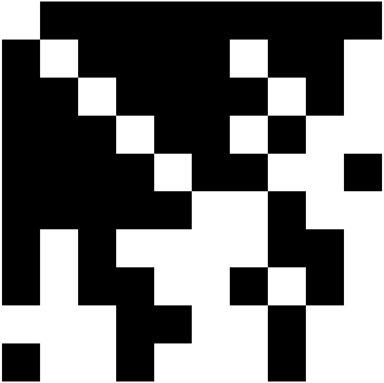
H_0 Erdos-Renyi, each edge randomly with prob. $1/2$

H_1 Planted k-clique, remaining edges random

Planted k-clique problem



Random entries



Planted $k \times k$ sub-matrix

Binary hypothesis test based on observing random binary matrix:

H_0 Random $\{0, 1\}$ matrix with $\mathrm{Ber}(1/2)$ on off-diagonal

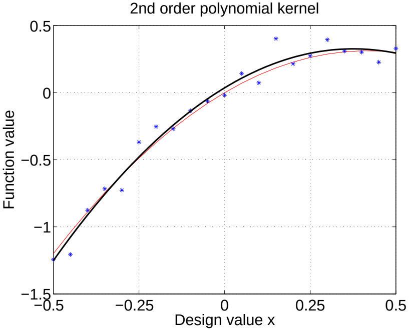
H_1 Planted $k \times k$ sub-matrix

Vignette C: (Structured) non-parametric regression

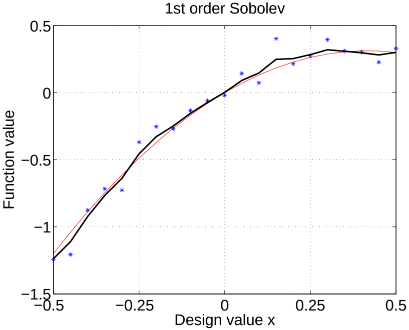
Goal: How to predict output from covariates?

- given covariates $(x_1, x_2, x_3, \ldots, x_p)$
- output variable y
- want to predict y based on $(x_1, \ldots, x_p)$

Examples: Medical Imaging; Geostatistics; Astronomy; Computational Biology .....



(a) Second-order polynomial fit



(b) Lipschitz function fit

High dimensions and sample complexity

Possible models:

- ordinary linear regression: $y = \underbrace{\sum_{j=1}^{p} \theta_j^* x_j}_{\langle \theta^*, x \rangle} + w$
- general non-parametric model: $y = f^*(x_1, x_2, \ldots, x_p) + w$.

Estimation accuracy: How well can f^* be estimated using n samples?

- linear models
  - without any structure: error $\delta^2 \asymp \underbrace{p/n}_{\text{linear in } p}$
  - with sparsity $k \ll p$: error $\delta^2 \asymp \underbrace{\left(k \log \frac{ep}{k}\right)/n}_{\text{logarithmic in } p}$
- non-parametric models: p-dimensional, smoothness α

Curse of dimensionality: Error $\delta^2 \asymp \underbrace{(1/n)^{\frac{2\alpha}{2\alpha+p}}}_{\text{Exponential slow-down}}$

Minimax risk and sample size

Consider a function class $\mathcal{F}$, and n i.i.d. samples from the model

$$y_i = f^*(x_i) + w_i, \qquad \text{where } f^* \text{ is some member of } \mathcal{F}.$$

For a given estimator $\{(x_i, y_i)\}_{i=1}^n \mapsto \widehat{f} \in \mathcal{F}$, *worst-case risk* in a metric ρ:

$$R^n_{\text{worst}}(\widehat{f}; \mathcal{F}) = \sup_{f^* \in \mathcal{F}} \mathbb{E}^n[\rho^2(\widehat{f}, f^*)].$$

Minimax risk

For a given sample size n, the minimax risk

$$\inf_{\widehat{f}} R^n_{\text{worst}}(\widehat{f}; \mathcal{F}) \;=\; \inf_{\widehat{f}} \sup_{f^* \in \mathcal{F}} \mathbb{E}^n\big[\rho^2(\widehat{f}, f^*)\big]$$

where the infimum is taken over all estimators.

How to measure “size” of function classes?

- A 2δ-packing of $\mathcal{F}$ in metric ρ is a collection $\{f^1, \ldots, f^M\} \subset \mathcal{F}$ such that

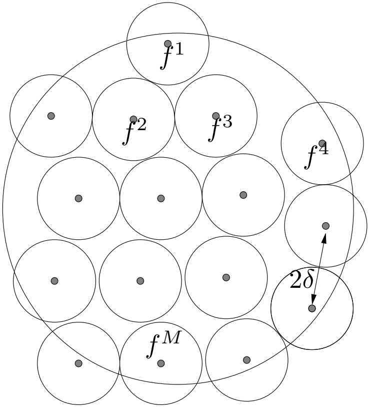
$$\rho\big(f^j, f^k\big) > 2\delta \qquad \text{for all } j \neq k.$$

- The packing number $M(2\delta)$ is the cardinality of the largest such set.

- Packing/covering entropy: emerged from Russian school in 1940s/1950s (Kolmogorov and collaborators)

- Central object in proving minimax lower bounds for nonparametric problems (e.g., Hasminskii & Ibragimov, 1978; Birge, 1983; Yu, 1997; Yang & Barron, 1999)

Packing and covering numbers

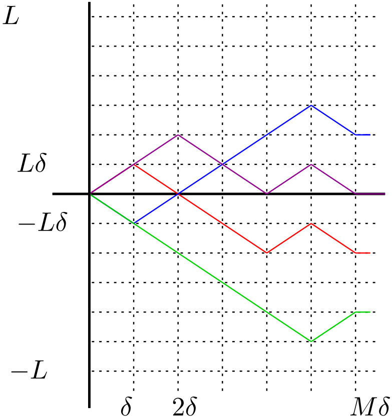


- A 2δ-packing of $\mathcal{F}$ in metric ρ is a collection $\{f^1, \ldots, f^M\} \subset \mathcal{F}$ such that

$$\rho\big(f^j, f^k\big) > 2\delta \qquad \text{for all } j \neq k.$$

- The packing number $M(2\delta)$ is the cardinality of the largest such set.

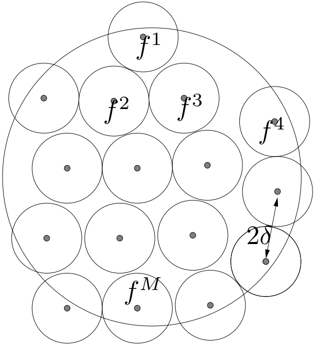
Example: Sup-norm packing for Lipschitz functions



- δ-packing set: functions $\{f^1, f^2, \ldots, f^M\}$ such that $\|f^j - f^k\|_2 > \delta$ for all $j \neq k$
- for L-Lipschitz functions in 1-dimension:

$$M(\delta) \asymp 2^{(L/\delta)}.$$

Standard reduction: from estimation to testing



- goal: to characterize the minimax risk for ρ-estimation over $\mathcal{F}$
- construct a 2δ-packing of $\mathcal{F}$:

  collection $\{f^1, \dots, f^M\}$ such that $\rho(f^j, f^k) > 2\delta$

- now form a M-component mixture distribution as follows:
  - draw packing index $V \in \{1, \dots M\}$ uniformly at random
  - conditioned on $V = j$, draw n i.i.d. samples $(X_i, Y_i) \sim \mathbb{Q}_{f^j}$

1. Claim: Any estimator $\widehat{f}$ such that $\rho(\widehat{f}, f^J) \leq \delta$ w.h.p. can be used to solve the M-ary testing problem.
2. Use standard techniques ({Assouad, Le Cam, Fano }) to lower bound the probability of error in the testing problem.

Minimax rate via metric entropy matching

- observe (x_i, y_i) pairs from model $y_i = f^*(x_i) + w_i$
- two possible norms

$$\|\widehat{f} - f^*\|_n^2 := \frac{1}{n}\sum_{i=1}^{n}\big(\widehat{f}(x_i) - f^*(x_i)\big)^2, \text{ or } \|\widehat{f} - f^*\|_2^2 = \mathbb{E}\big[(\widehat{f}(\widetilde{X}) - f^*(\widetilde{X}))^2\big].$$

Metric entropy master equation

For many regression problems, minimax rate $\delta_n > 0$ determined by solving the master equation

$$\log M\big(2\delta; \mathcal{F}, \|\cdot\|\big) \asymp n\delta^2.$$

- basic idea (with Hellinger metric) dates back to Le Cam (1973)
- elegant general version due to Yang and Barron (1999)

Example 1: Sparse linear regression

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

$$\theta^* \in \Theta(k,p) := \Big\{ \theta \in \mathbb{R}^p \mid \|\theta\|_0 \leq k, \quad \text{and} \quad \|\theta\|_2 \leq 1 \Big\}.$$

Gilbert-Varshamov: can construct a 2δ-separated set with

$$\log M(2\delta) \asymp k \log \big(\frac{ep}{k}\big) \qquad \text{elements}$$

Master equation and minimax rate

$$\log M(2\delta) \;\asymp\; n\delta^2 \quad \Longleftrightarrow \quad \delta^2 \asymp \frac{k \log \big(\frac{ep}{k}\big)}{n}.$$

Polynomial-time achievability:

- by ℓ_1-relaxations under restricted eigenvalue (RE) conditions (Candes & Tao, 2007; Bickel et al., 2009; Buhlmann & van de Geer, 2011)
- achieve minimax-optimal rates for ℓ_2-error (Raskutti, W., & Yu, 2011)
- ℓ_1-methods can be sub-optimal for prediction error $\|X(\widehat{\theta} - \theta^*)\|_2/\sqrt{n}$ (Zhang, W. & Jordan, 2014)

Example 2: α-smooth non-parametric regression

Observations $y_i = f^*(x_i) + w_i$, where

$$f^* \in \mathcal{F}(\alpha) = \Big\{ f : [0,1] \to \mathbb{R} \mid \text{ f is } \alpha\text{-times diff'ble, with } \textstyle\sum_{j=0}^{\alpha} \|f^{(j)}\|_\infty \le C \Big\}.$$

Classical results in approximation theory (e.g., Kolmogorov & Tikhomorov, 1959)

$$\log M(2\delta; \mathcal{F}) \asymp \big(1/\delta\big)^{1/\alpha}$$

Master equation and minimax rate

$$\log M(2\delta) \;\asymp\; n\delta^2 \quad \Longleftrightarrow \quad \delta^2 \asymp \big(1/n\big)^{\frac{2\alpha}{2\alpha+1}}.$$

Example 3: Sparse additive and α-smooth

Observations $y_i = f^*(x_i) + w_i$, where f^* satisfies constraints

Additively decomposable: $f^*(x_1, \ldots x_p) = \sum_{j=1}^{p} g_j^*(x_j)$

Sparse: At most k of $(g_1^*, \ldots, g^*p)$ are non-zero

Smooth: Each g_j^* belongs to α-smooth family $\mathcal{F}(\alpha)$

Combining previous results yields 2δ-packing with

$$\log M(2\delta; \mathcal{F}) \asymp k\big(1/\delta\big)^{1/\alpha} + k \log\big(\frac{ep}{k}\big)$$

Master equation and minimax rate

Solving $\log M(2\delta) \asymp n\delta^2$ yields

$$\delta^2 \asymp \underbrace{k\big(1/n\big)^{\frac{2\alpha}{2\alpha+1}}}_{k\text{-component estimation}} + \underbrace{\frac{k \log\big(\frac{ep}{k}\big)}{n}}_{\text{search complexity}}$$

Summary

To be provided during tutorial....