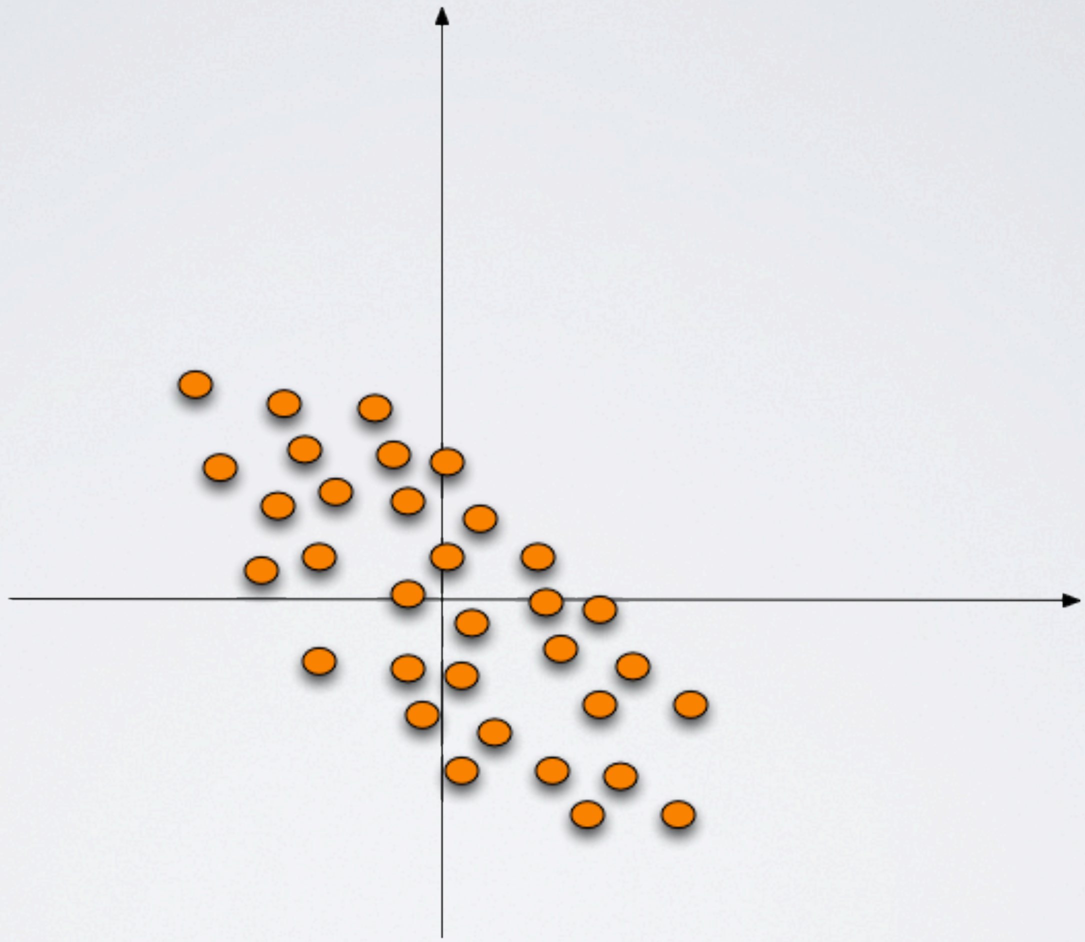
OPTIMAL DETECTION OF SPARSE PRINCIPAL COMPONENTS

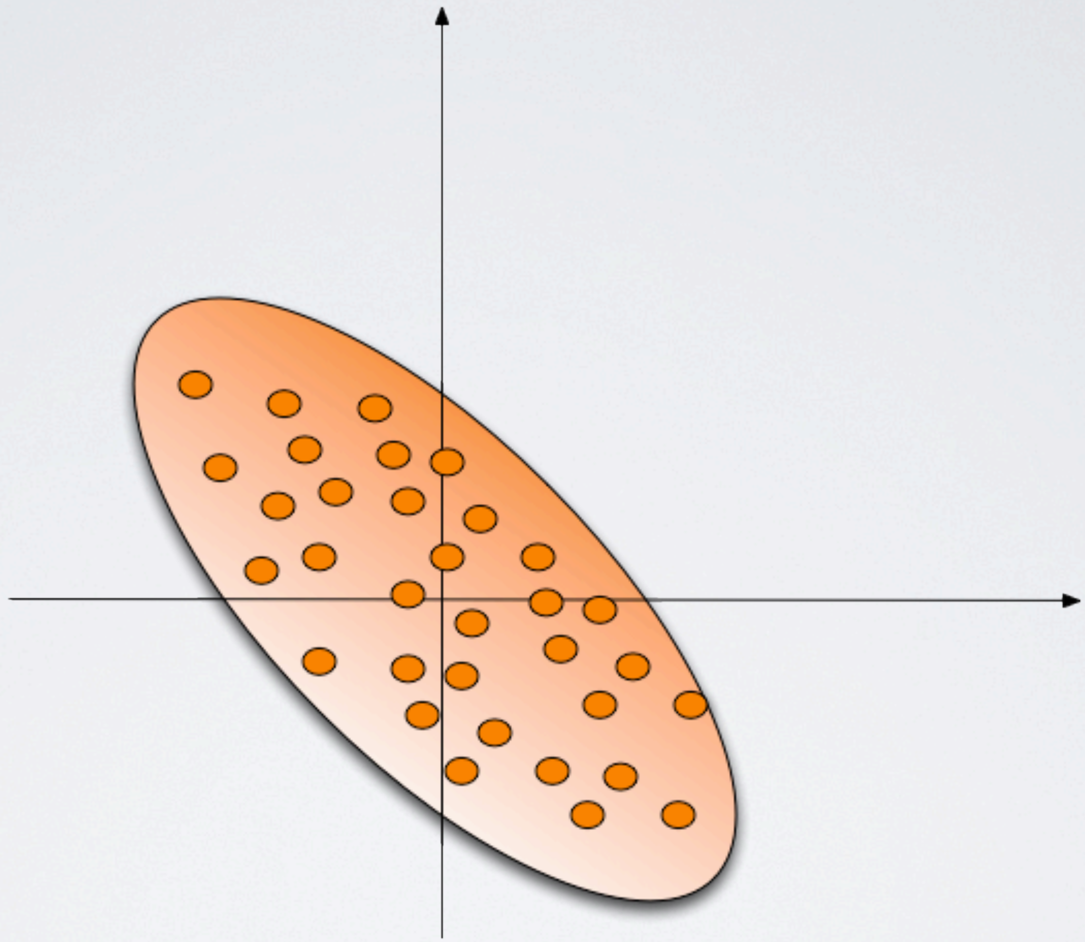
Philippe Rigollet (joint with Quentin Berthet)

High dimensional data



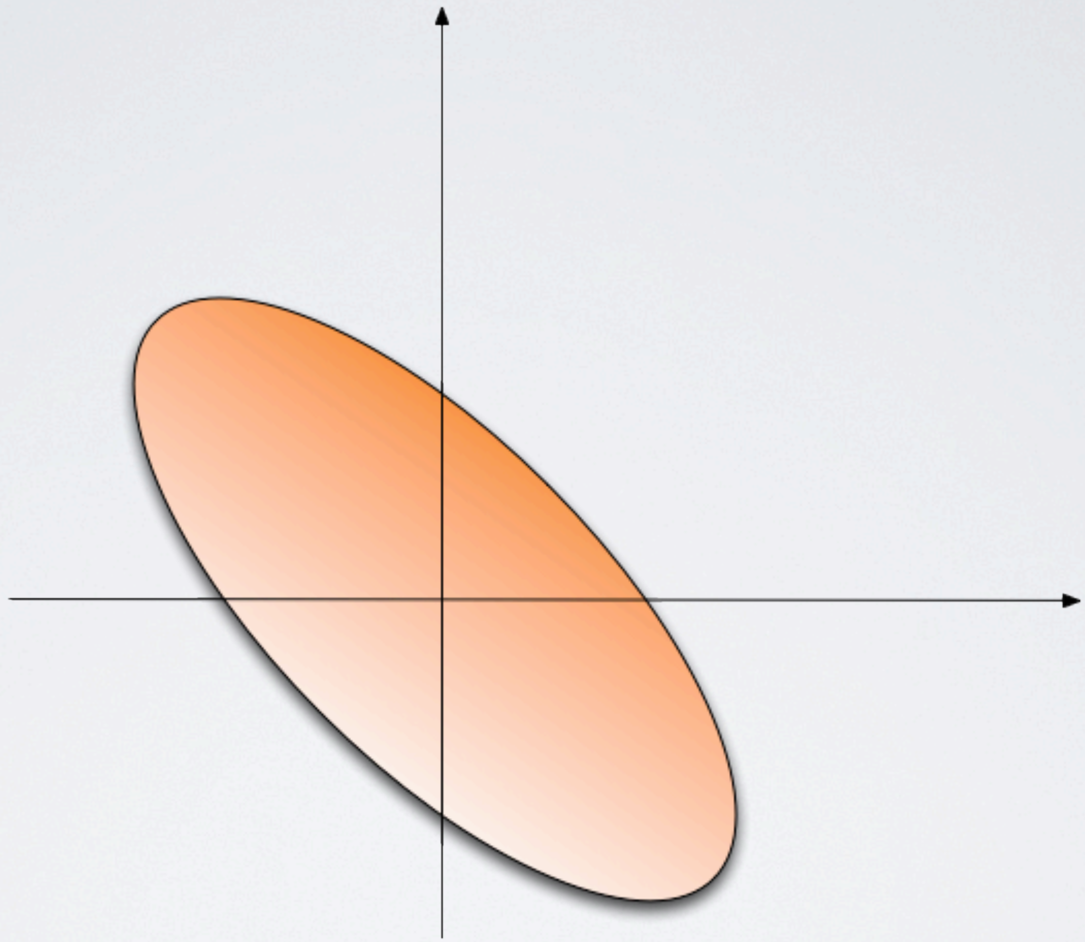
Cloud of point in $\mathbb{R}^p$

High dimensional data



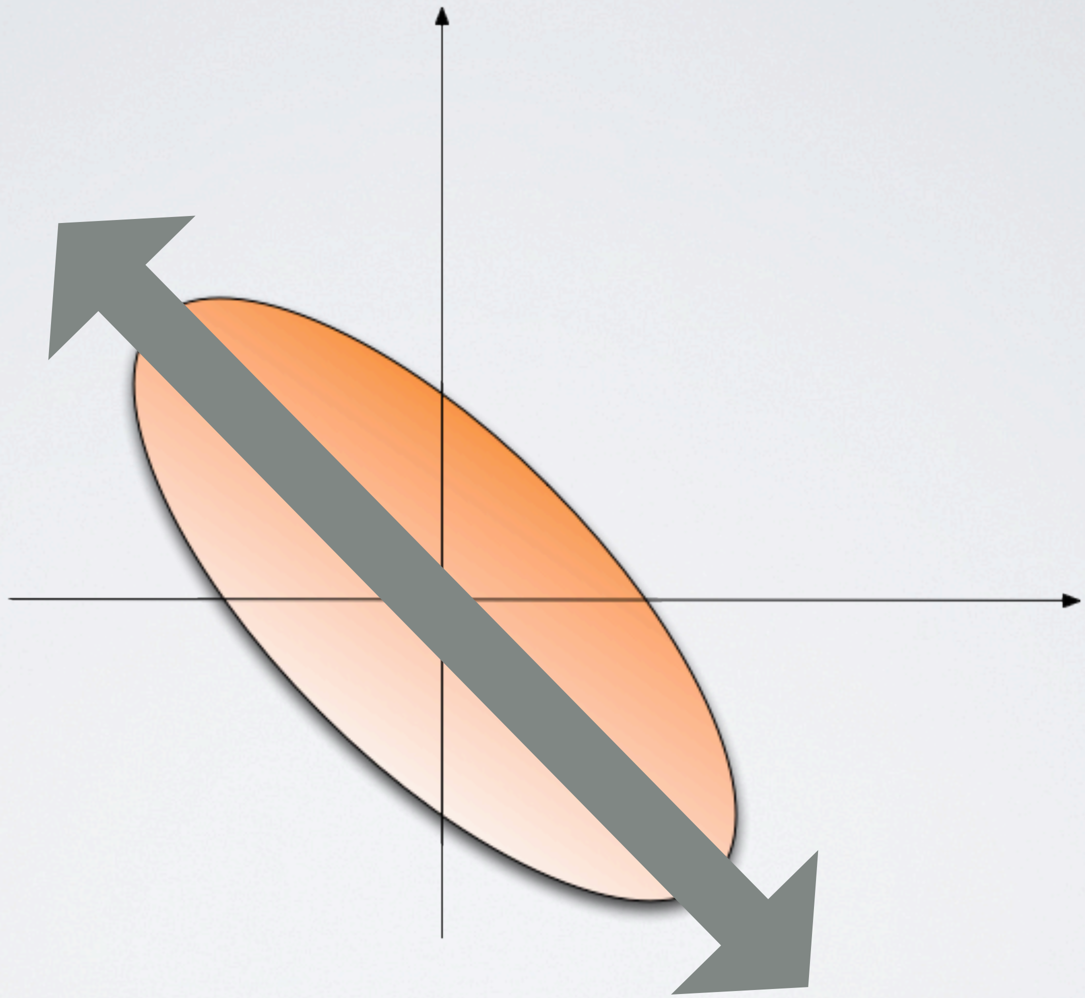
Cloud of point in $\mathbb{R}^p$

High dimensional data



Cloud of n points in $\mathbb{R}^p$

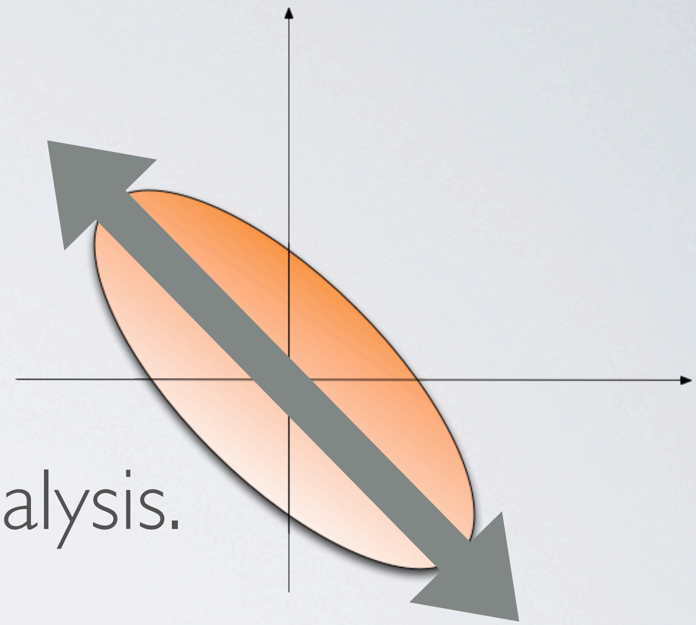
Principal component



Principal component = direction of largest variance

Principal component analysis (PCA)

- Tool for dimension reduction
- Spectrum of covariance matrix
- Main tool for exploratory data analysis.



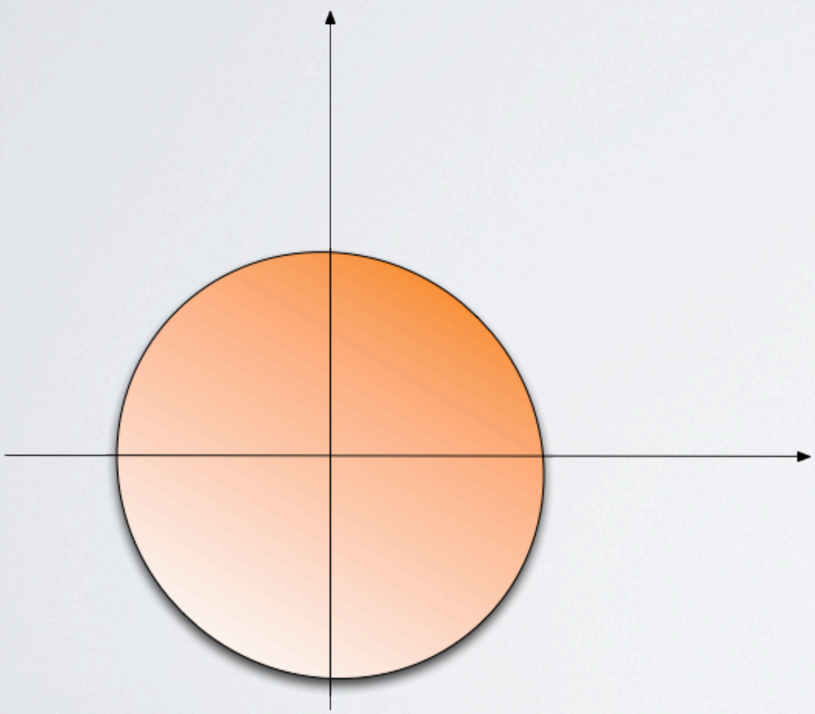
We study only the first principal component

This talk: high-dimensional $p \gg n$, finite sample framework.

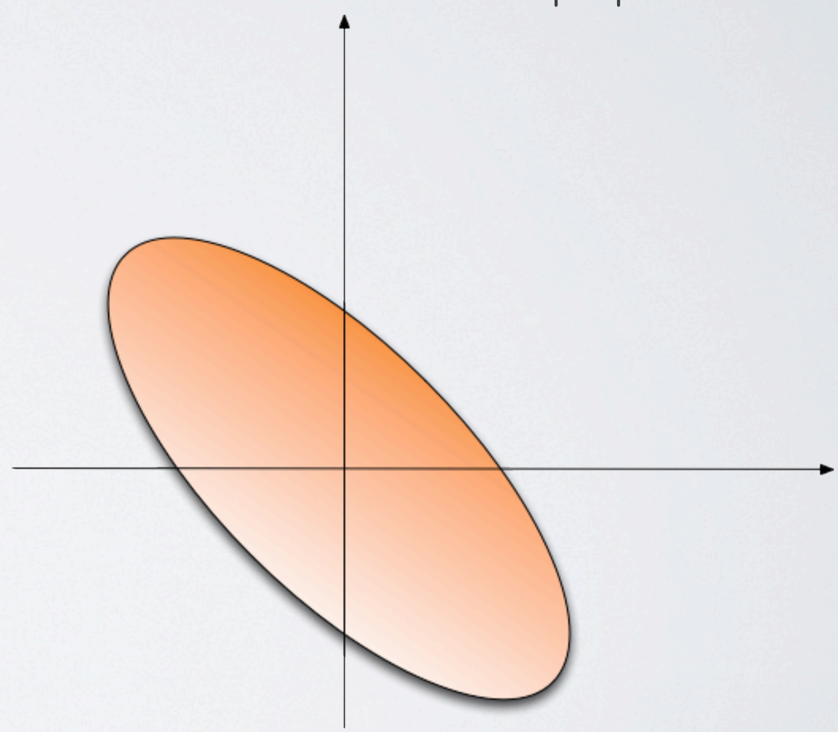
Testing for sphericity under rank-one alternative

$$H_0 \, : \, \Sigma = I_p$$

$$H_1 \, : \, \Sigma = I_p + \theta v v^\top$$
$$|v|_2 = 1$$



Isotropic



Principal component

The model

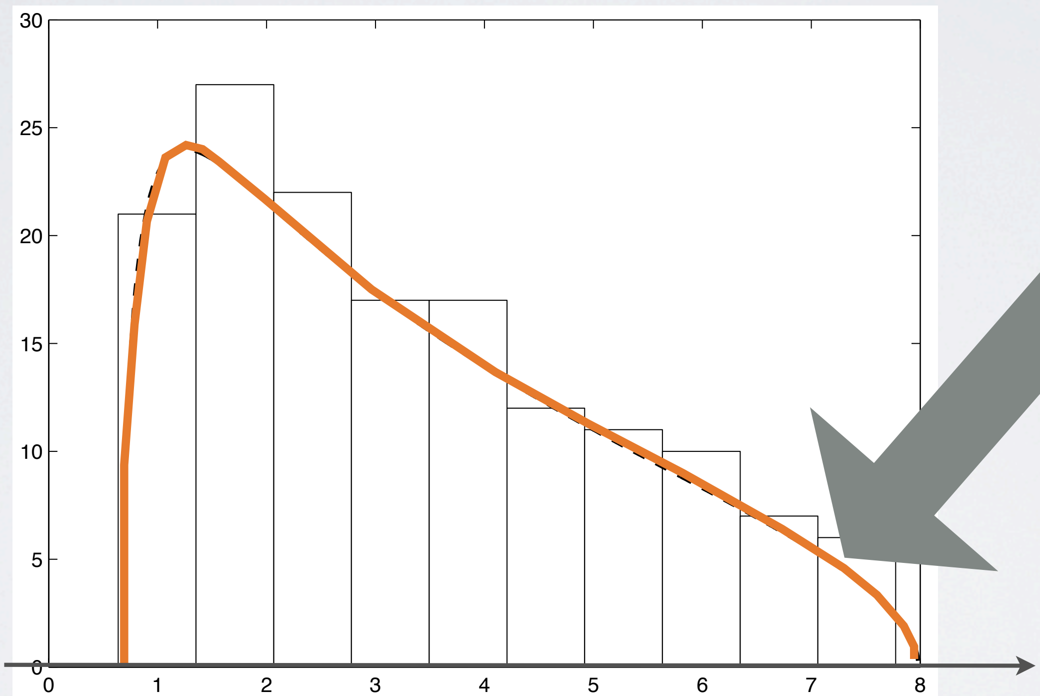
- Observations: i.i.d. $X_1, \ldots, X_n \sim \mathcal{N}_p(0, \Sigma)$
- Estimator: empirical covariance matrix

$$\hat{\Sigma} = \frac{1}{n} \sum_{i=1}^{n} X_i X_i^\top$$

If $n \gg p$ it is a consistent estimator.
If $n \simeq cp$ it is inconsistent (Nadler, Paul, Onatski, ...)
eigenvectors are orthogonal

Empirical spectrum under the null

$$H_0 : \Sigma = I_p$$



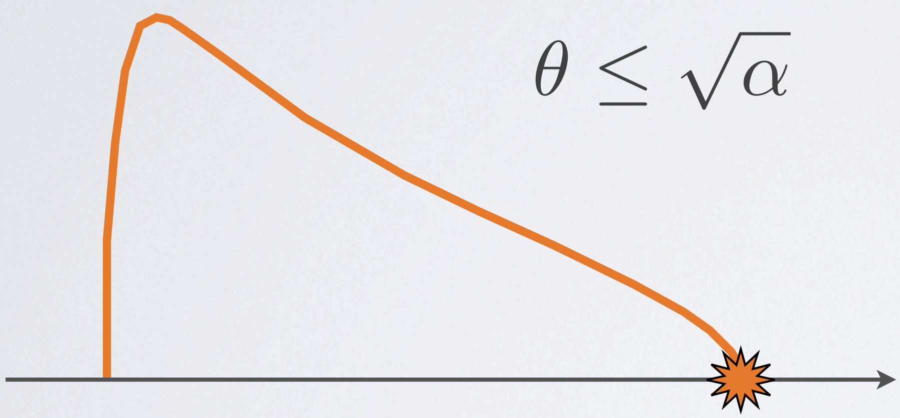
Marcenko-Pastur distribution

Spectrum of $\hat{\Sigma}$

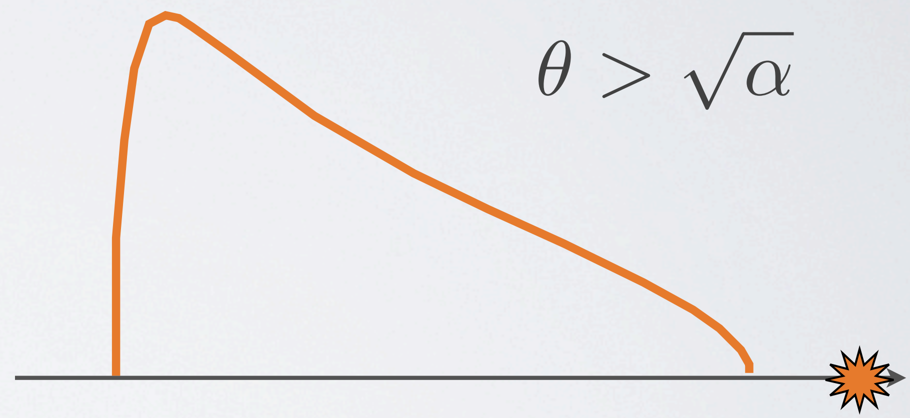
Empirical spectrum under the alternative

$$H_1 \ : \ \Sigma = I_p + \theta v v^\top \qquad |v|_2 = 1$$

The **BBP** (Baik, Ben Arous, Péché) transition $\dfrac{p}{n} \to \alpha > 0$



Indistinguishable from the null



detection possible if

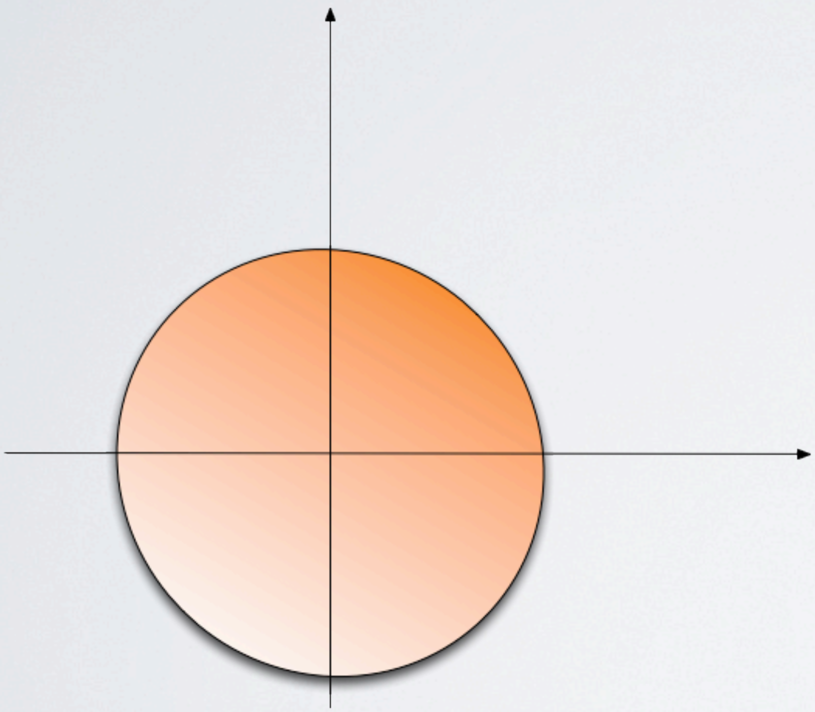
$$\theta > \sqrt{\frac{p}{n}}$$

very strong signal!

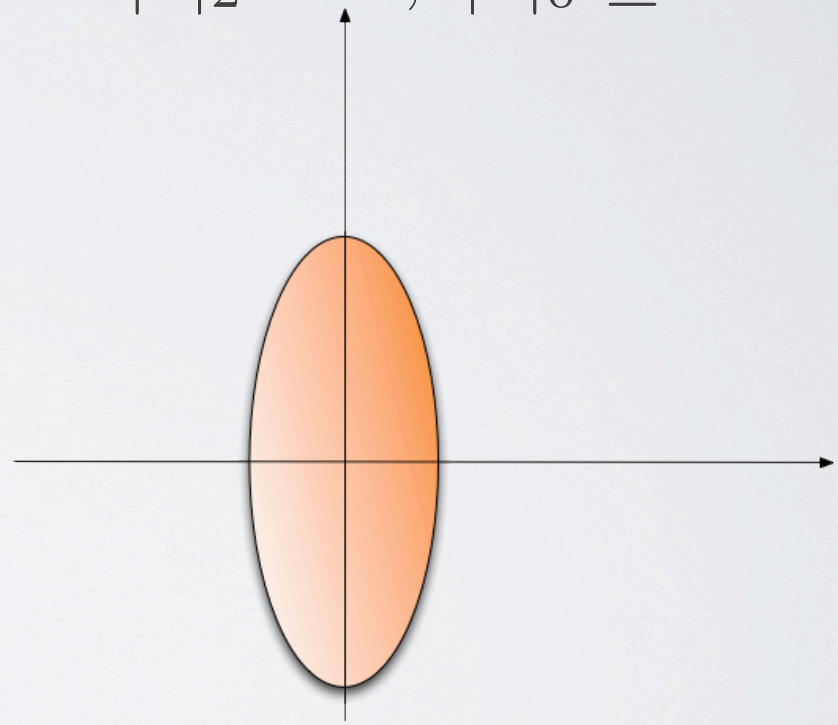
Testing for sparse principal component

$$H_0 \,:\, \Sigma = I_p$$

$$H_1 \,:\, \Sigma = I_p + \theta v v^\top,$$
$$|v|_2 = 1\,,\ |v|_0 \leq k$$



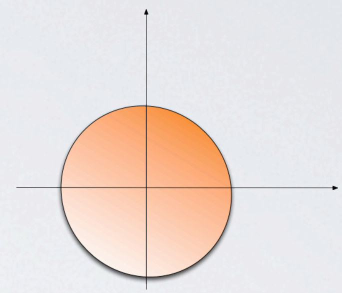
Isotropic



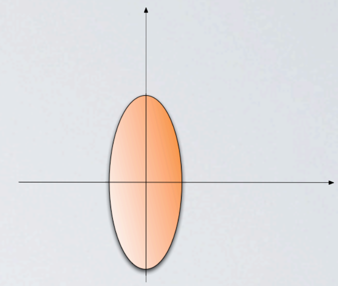
Sparse principal direction

Testing for sparse principal component

$H_0 \,:\, \Sigma = I_p$



$H_1 \,:\, \Sigma = I_p + \theta v v^\top,$
$|v|_2 = 1\,,\ |v|_0 \le k$

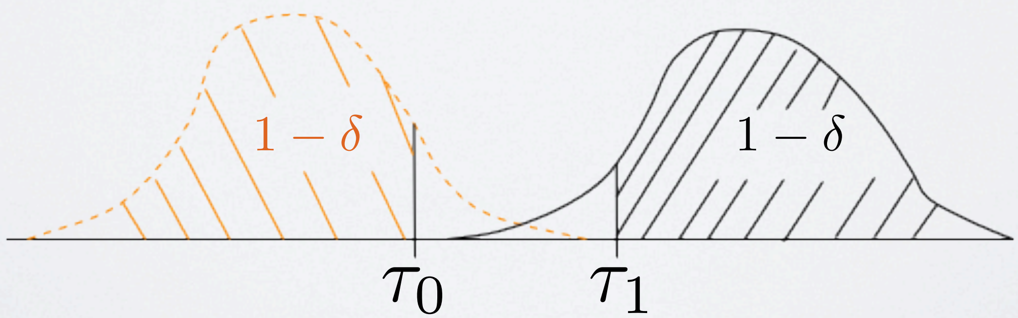


minimum detection level θ ?

Goal: find a statistic $\varphi : \mathbf{S}_p^+ \mapsto \mathbb{R}$ such that

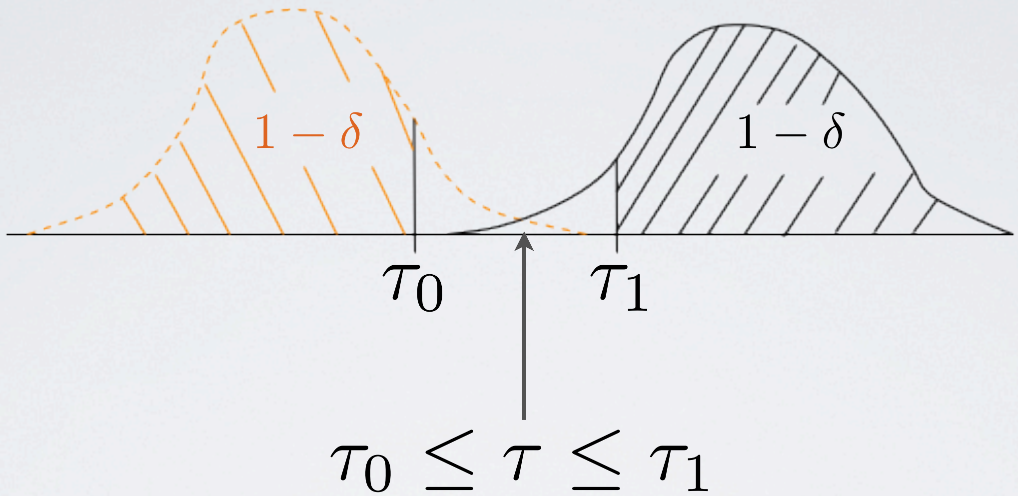
$$\mathbf{P}_{H_0}(\varphi(\hat\Sigma) < \tau_0) \ge 1 - \delta \longrightarrow \text{small under } H_0$$

$$\mathbf{P}_{H_1}(\varphi(\hat\Sigma) > \tau_1) \ge 1 - \delta \longrightarrow \text{large under } H_1$$



$\mathbf{P}_{H_0}(\varphi(\hat{\Sigma}) < \tau_0) \geq 1 - \delta$ ⟶ small under H_0

$\mathbf{P}_{H_1}(\varphi(\hat{\Sigma}) > \tau_1) \geq 1 - \delta$ ⟶ large under H_1



Take the test: $\psi(\hat{\Sigma}) = \mathbf{1}\{\varphi(\hat{\Sigma}) > \tau\}$. It satisfies:

$$\mathbf{P}_{H_0}(\psi = 1) \vee \max_{\substack{|v|_2 = 1 \\ |v|_0 \leq k}} \mathbf{P}_{H_1}(\psi = 0) \leq \delta$$

Sparse eigenvalue

k-sparse eigenvalue:

$$\varphi(\hat{\Sigma}) = \lambda^k_{\max}(\hat{\Sigma}) = \max_{\substack{|x|_2 = 1 \\ |x|_0 \leq k}} x^\top \hat{\Sigma}\, x = \max_{|S|=k} \lambda_{\max}(\hat{\Sigma}_S)$$

Note that: $\lambda^k_{\max}(I_p) = 1 \quad \text{and} \quad \lambda^k_{\max}(I_p + \theta v v^\top) = 1 + \theta$

Smaller fluctuations than the largest eigenvalue $\lambda_{\max}(\hat{\Sigma})$

Upper bounds w.p. $1-\delta$

Under the **null hypothesis**:

$$\lambda^k_{\max}(\hat{\Sigma}) \le 1 + 8\sqrt{\frac{k\log(9ep/k) + \log(1/\delta)}{n}} =: \tau_0$$

Under the **alternative hypothesis**:

$$\lambda^k_{\max}(\hat{\Sigma}) \ge 1 + \theta - 2(1+\theta)\sqrt{\frac{\log(1/\delta)}{n}} =: \tau_1$$

Can detect as soon as $\tau_0 < \tau_1$, which yields

$$\theta \ge C\sqrt{\frac{k\log(p/k)}{n}}$$

Minimax lower bound

Fix $\nu > 0$ (small).

Then there exists a constant $C_\nu > 0$ such that if

$$\theta < \bar{\theta} := \sqrt{\frac{k \log\left(C_\nu p/k^2 + 1\right)}{n}} \wedge \frac{1}{\sqrt{2}}$$

Then

$$\inf_{\psi} \left\{ \mathbf{P}_0^n(\psi = 1) \vee \max_{\substack{|v|_2 = 1 \\ |v|_0 \leq k}} \mathbf{P}_v^n(\psi = 0) \right\} \geq \frac{1}{2} - \nu$$

See also Arias-Castro, Bubeck and Lugosi (12)

Computational issues

To compute $\lambda^k_{\max}(\hat{\Sigma})$, need to compute $\binom{p}{k}$ eigenvalues

Can be used to find cliques in graphs: NP-complete pb.

Need an approximation...

Semidefinite relaxation 101

Cauchy-Schwarz

$$\lambda^k_{\max}(A) = \text{max.} \quad x^\top A x \quad \text{subject to } x^\top x = 1,\ |x|_0 \le k$$

$$\mathrm{SDP}_k(A) = \text{max.} \quad \mathrm{Tr}(AZ) \quad \text{subject to } \mathrm{Tr}(Z) = 1,\ |Z|_1 \le k$$

$$|xx^\top|_1 \le k$$

$$Z = xx^\top \qquad \mathrm{rank}(Z) = 1 \quad Z \succeq 0$$

Semidefinite program program (SDP) introduced by d'Aspremont, El Gahoui, Jordan and Lanckriet (2004).

Testing procedure: $\mathbf{1}\{\mathrm{SDP}_k(\hat{\Sigma}) > \tau\}$

Defined even if solution of SDP has rank > 1

Performance of SDP

For the **alternative**: relaxation of $\lambda^k_{\max}(\hat{\Sigma})$ so

$$\mathsf{SDP}_k(\hat{\Sigma}) \geq \lambda^k_{\max}(\hat{\Sigma})$$

For the **null**: use dual (Bach *et al.* 2010)

$$\mathsf{SDP}_k(A) = \min_{U \in \mathbf{S}_p^+} \{\lambda_{\max}(A+U) + k|U|_\infty\}$$

For any $U \in \mathbf{S}_p^+$ this gives an upper bound on $\mathsf{SDP}_k(\hat{\Sigma})$

Enough to look only at **minimum dual perturbation**

$$\mathsf{MDP}_k(\hat{\Sigma}) = \min_{z \geq 0} \left\{\lambda_{\max}(\mathsf{st}_z(\hat{\Sigma})) + kz\right\}$$

Upper bounds w.p. $1-\delta$

$*\text{DP} \in \{\text{SDP}, \text{MDP}\}$

Under the **null hypothesis**:

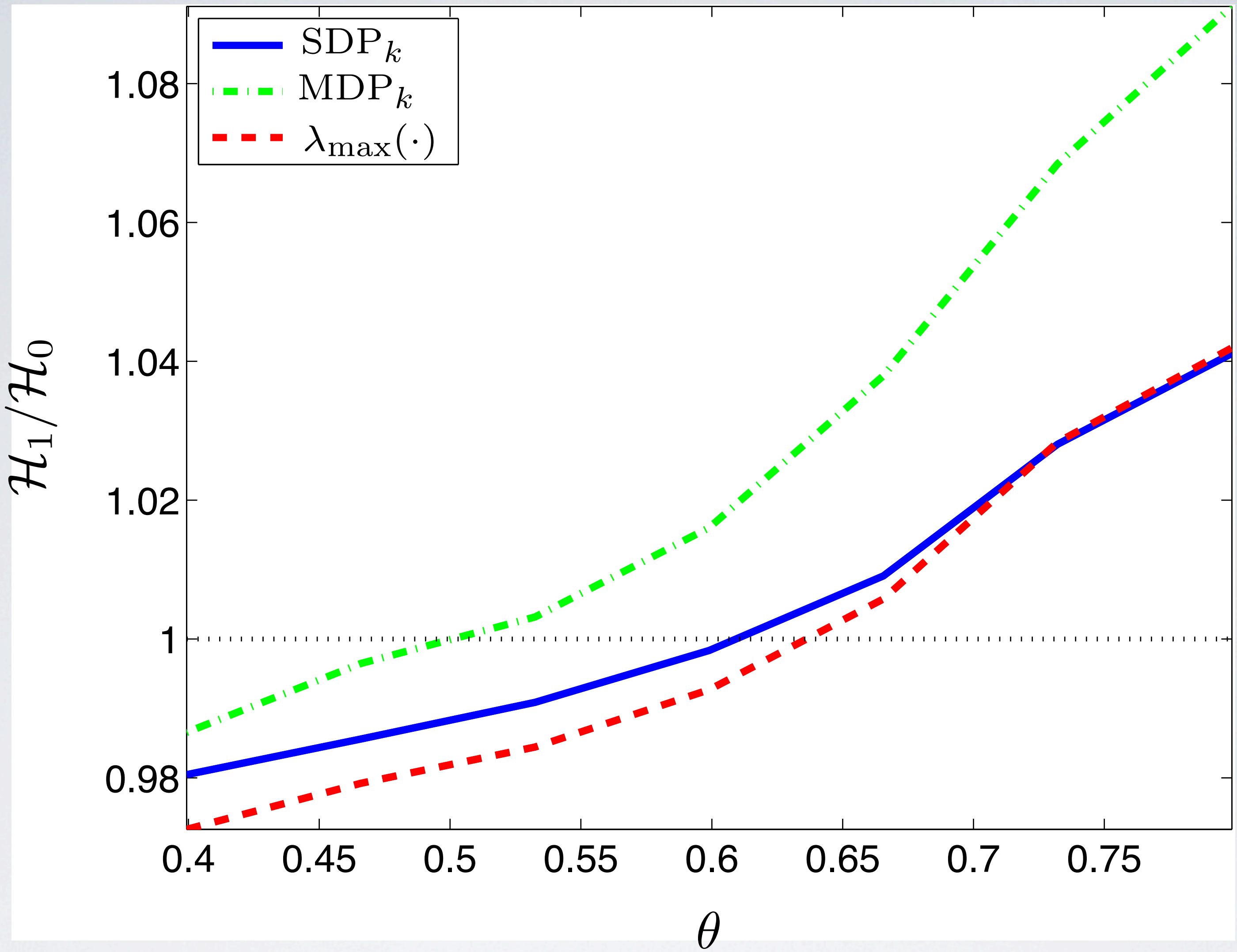
$$*\text{DP}_k(\hat{\Sigma}) \leq 1 + 10\sqrt{\frac{k^2 \log(ep/\delta)}{n}} =: \tau_0$$

Under the **alternative hypothesis**:

$$*\text{DP}_k(\hat{\Sigma}) \geq 1 + \theta - 2(1+\theta)\sqrt{\frac{\log(1/\delta)}{n}} =: \tau_1$$

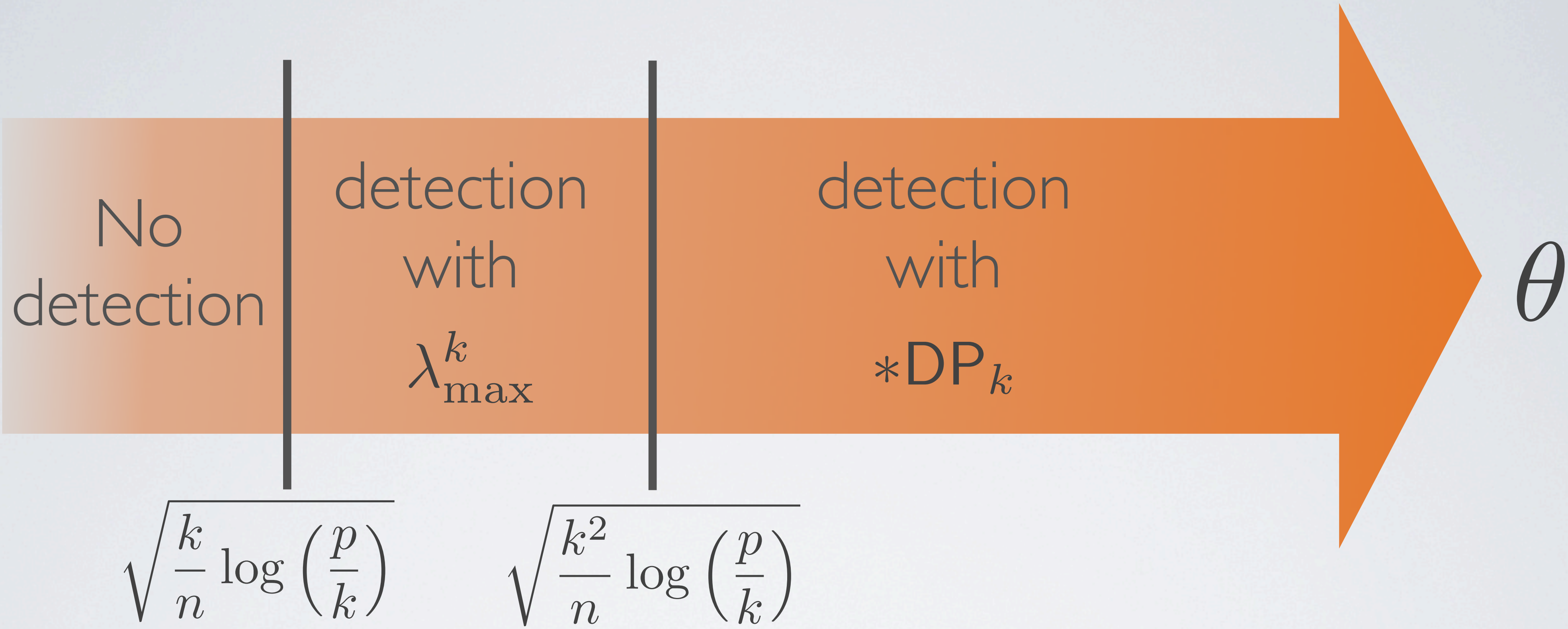
Can detect as soon as $\tau_0 < \tau_1$, which yields

$$\theta \geq C\sqrt{\frac{k^2 \log(p/k)}{n}}$$



Ratio of 5% quantile under $\mathcal{H}_1$ over 95% quantile under $\mathcal{H}_0$, versus signal strength θ. When this ratio is larger than one, both type I and type II errors are below 5%.

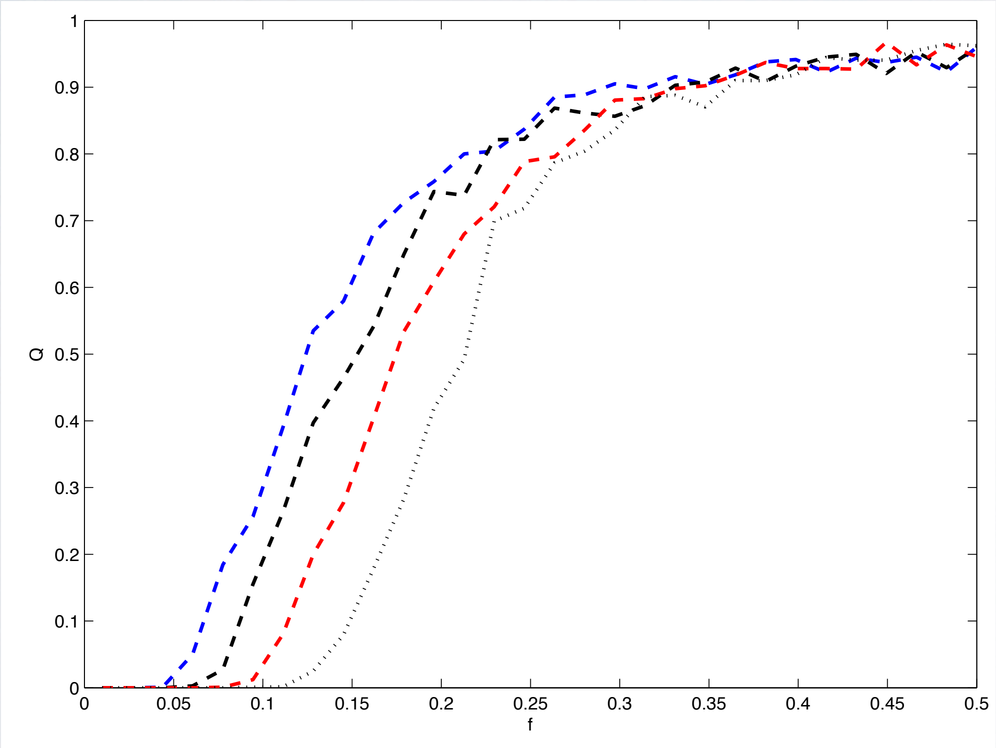
Summary

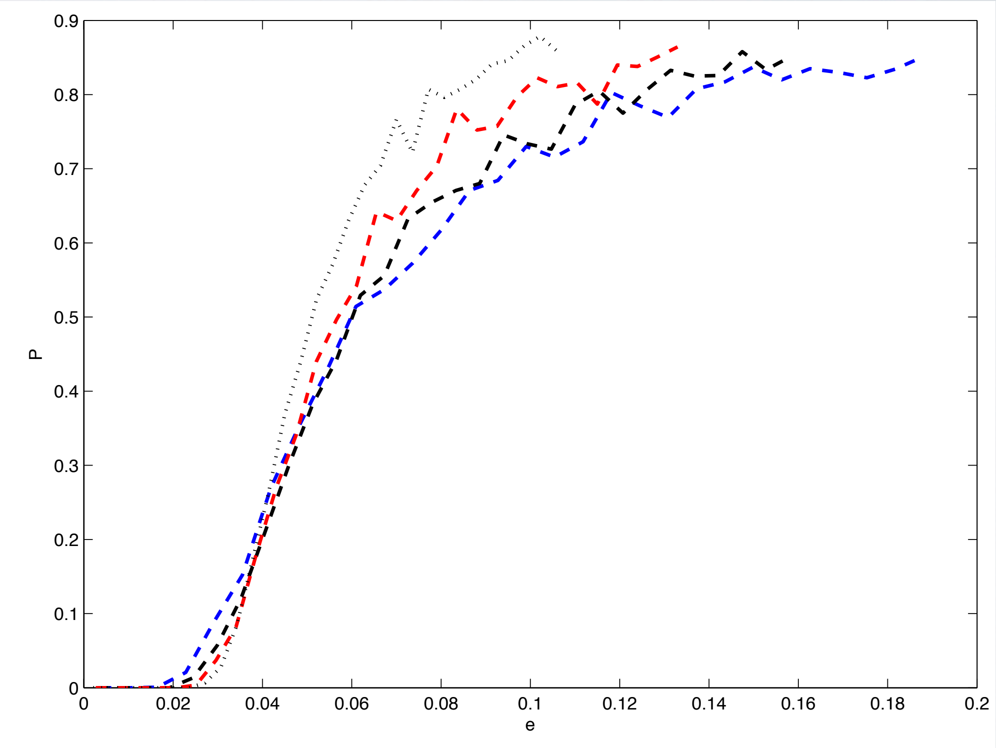


Can we tighten the gap?

Numerical evidence

Fix type I error at 1%, plot type II error of MDP_k
p={50, 100, 200, 500}, $k=\sqrt{p}$





$$\frac{k}{n}\log\left(\frac{p}{k}\right)$$

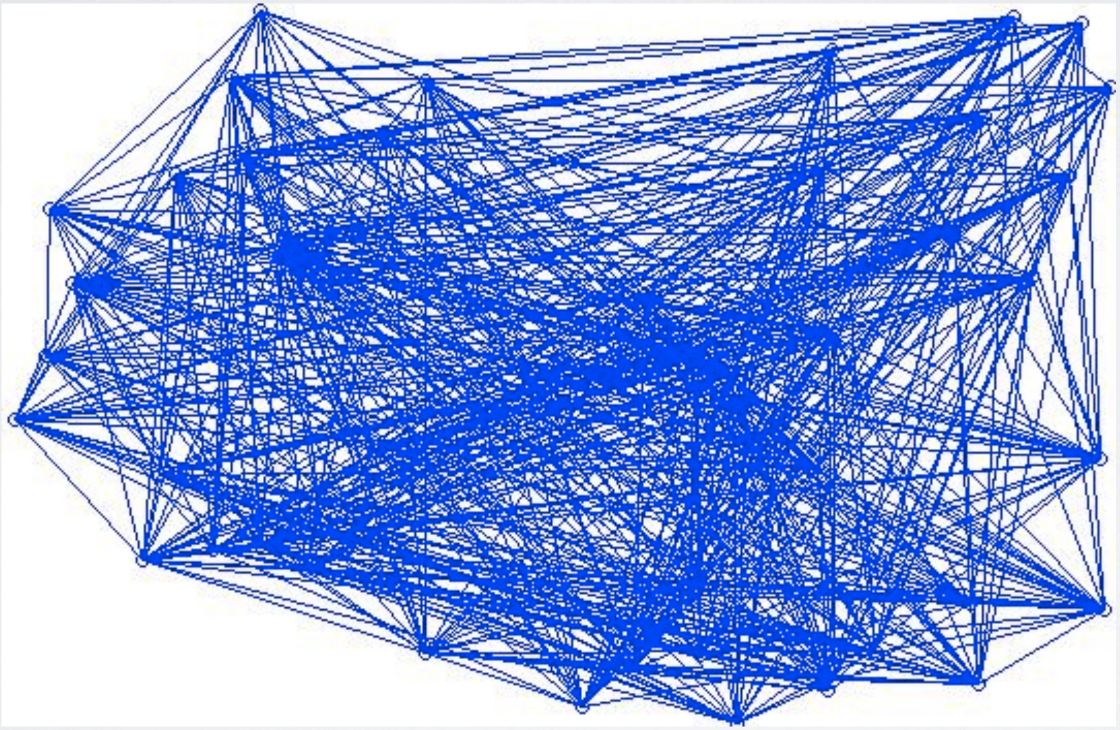
minimax optimal scaling

$$\frac{k^2}{n}\log\left(\frac{p}{k}\right)$$

proved scaling

Random graphs

A random (Erdos-Renyi) graph on N vertices is obtained by drawing edges at random with probability 1/2

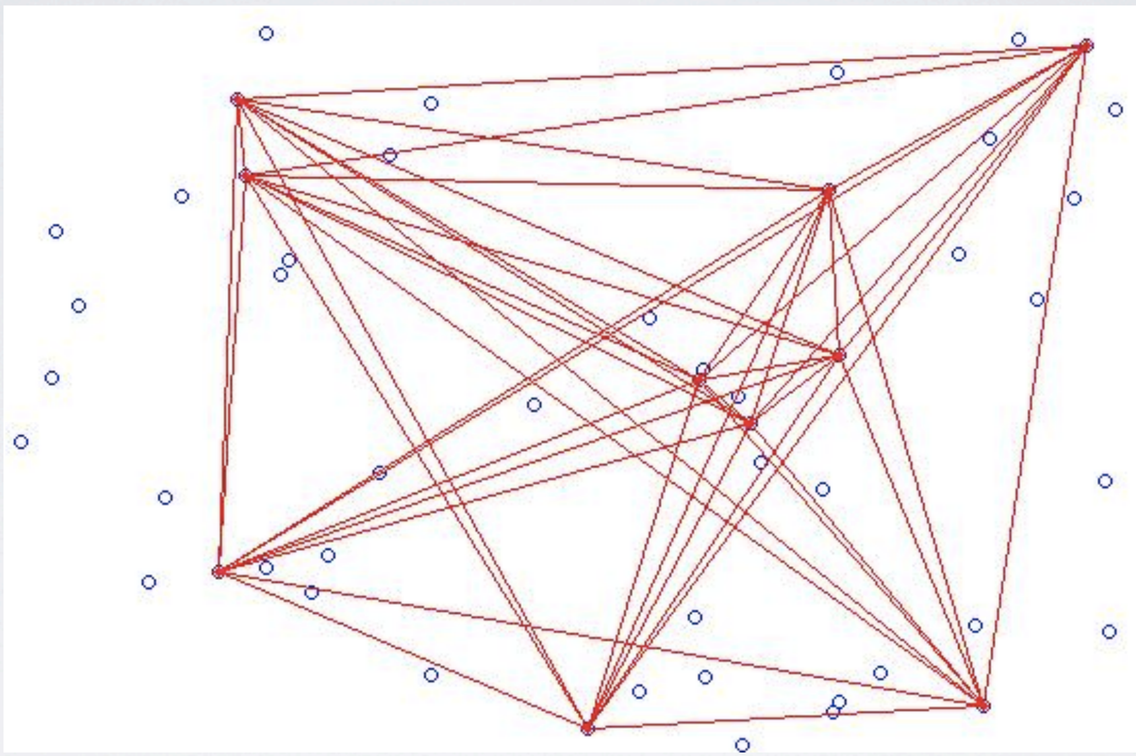


N=50

largest clique is of size $2 \log N$ =7.8 asymp. almost surely

Hidden clique

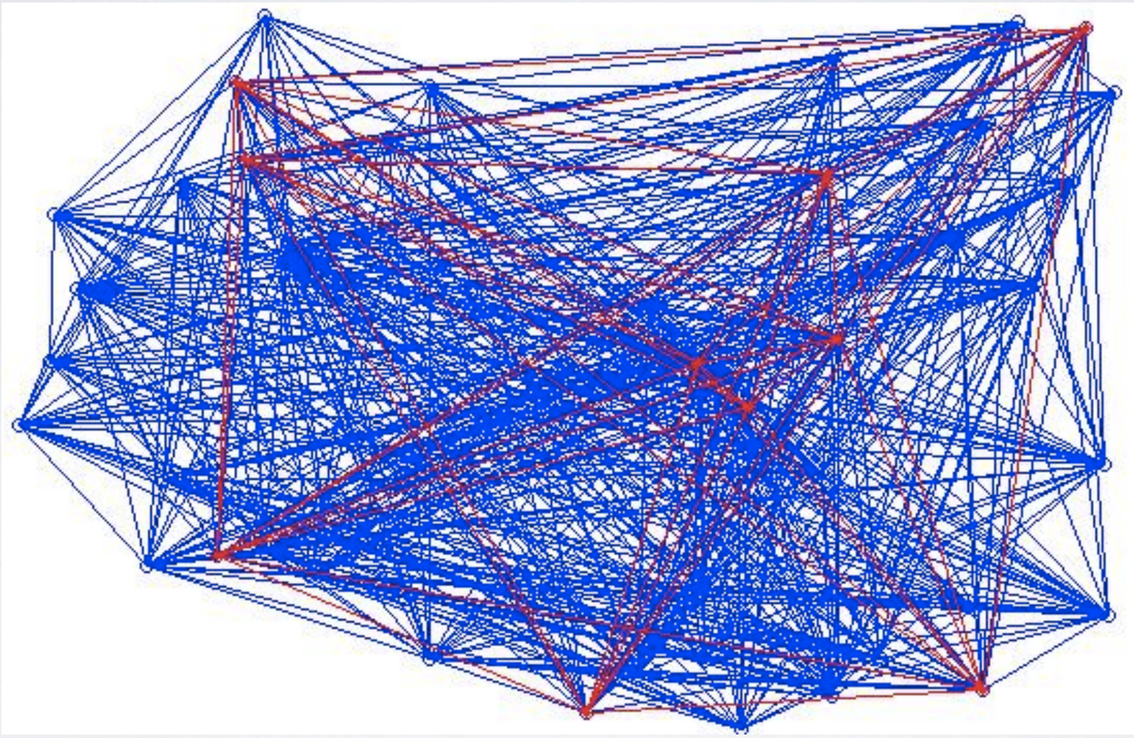
We can hide a clique (here of size 10) in this graph



Choose points arbitrarily and draw a clique

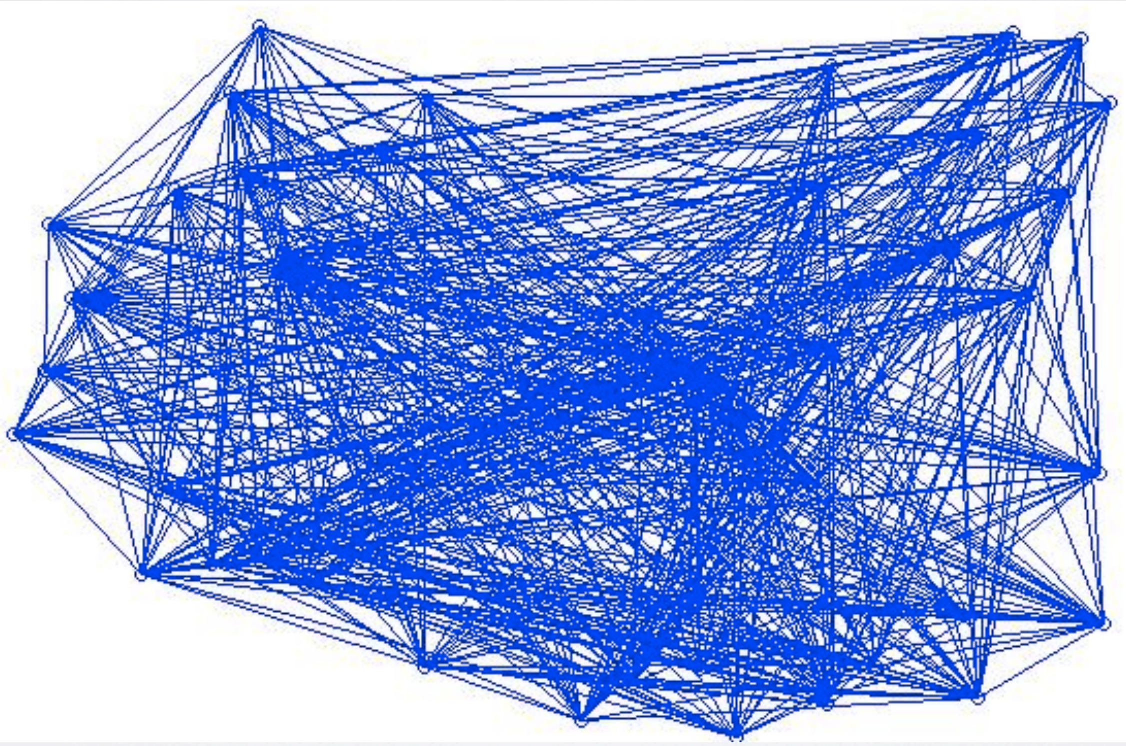
Hidden clique

embed in the original random graph



Hidden clique

Question: is there a hidden clique in this graph?



Hidden clique problem

It is believed that it is hard to find/test the presence of a clique in a random graph (Alon, Arora, Feige, Hazan, Krauthgamer,... Cryptosystems are based on this fact!)

Conjecture: It is hard to find cliques of size between

$2 \log N$ and $\sqrt{N}$

- Alon, Krivelevich, Sudakov 98
- Feige and Krauthgamer 00
- Dekel *et al.* 10
- Feige and Ron 10
- Ames and Vavasis 11

Canonical example of **average case complexity**

Hidden clique problem

It seems related to our problem but not trivially (the randomness structure is very fragile)

Note that all our results extend to **sub-Gaussian** r.v.

Theorem. If we could prove that there exists $C > 0$ such that under the null hypothesis it holds

$$\mathsf{SDP}_k(\hat{\Sigma}) \leq 1 + C\sqrt{\frac{k^{\alpha}\log(ep/\delta)}{n}}$$

for some $\alpha \in (1,2)$, then it can be used to test the presence of a clique of size $\mathsf{polylog}(N)N^{\frac{1}{4-\alpha}}$

Remarks

Unlike usual hardness results, this one is for one (actually two) method only (not for all methods).

In progress: we can remove this limitation using bi-cliques (need to carefully deal with independence)

Conclusion

- Optimal rates for sparse detection
- Computationally efficient methods with suboptimal rate
- First(?) link between sparse detection and average case complexity
- Opens the door to new statistical lower bounds:
  complexity theoretic lower bounds
- Evidence that heuristics **cannot** be optimal