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 $n \approx cp$ it is inconsistent (Nadler, Paul, Onatski, ...)
eigenvectors are orthogonal
Empirical spectrum under the null

\[ H_0 : \Sigma = I_p \]

Spectrum of \( \hat{\Sigma} \)

Marcenko-Pastur distribution
Empirical spectrum under the alternative

\[ H_1 : \Sigma = I_p + \theta vv^\top \quad |v|_2 = 1 \]

The **BBP** (Baik, Ben Arous, Péché) transition \( \frac{p}{n} \rightarrow \alpha > 0 \)

\[ \theta \leq \sqrt{\alpha} \]

\[ \theta > \sqrt{\alpha} \]

Indistinguishable from the null
detection possible if \( \theta > \sqrt{\frac{p}{n}} \)

very strong signal!
Testing for sparse principal component

$H_0 : \Sigma = I_p$

$H_1 : \Sigma = I_p + \theta vv^\top,$

$|v|_2 = 1, \ |v|_0 \leq k$

Isotropic

Sparse principal direction
Testing for sparse principal component minimum detection level $\theta$?

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

\[
\begin{align*}
    P_{H_0}(\varphi(\hat{\Sigma}) < \tau_0) &\geq 1 - \delta \quad \text{small under } H_0 \\
    P_{H_1}(\varphi(\hat{\Sigma}) > \tau_1) &\geq 1 - \delta \quad \text{large under } H_1
\end{align*}
\]
\( \mathbf{P}_{H_0}(\varphi(\hat{\Sigma}) < \tau_0) \geq 1 - \delta \quad \rightarrow \quad \text{small under } H_0 \)

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

\[
\begin{align*}
\tau_0 & \leq \tau \leq \tau_1 \\
\mathbf{P}_{H_0}(\psi = 1) \lor \max_{|v|_2 = 1} \mathbf{P}_{H_1}(\psi = 0) \leq \delta
\end{align*}
\]

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

\[ \varphi(\hat{\Sigma}) = \lambda_{\text{max}}^k(\hat{\Sigma}) = \max \quad x^\top \hat{\Sigma} x = \max_{|S| = k} \lambda_{\text{max}}(\hat{\Sigma}_S) \]

\[ \text{Note that: } \lambda_{\text{max}}^k(I_p) = 1 \quad \text{and} \quad \lambda_{\text{max}}^k(I_p + \theta vv^\top) = 1 + \theta \]

Smaller fluctuations than the largest eigenvalue \[ \lambda_{\text{max}}(\hat{\Sigma}) \]
Upper bounds w.p. $1 - \delta$

Under the **null hypothesis**:

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

Under the **alternative hypothesis**:

$$\lambda_{\text{max}}^k(\hat{\Sigma}) \geq 1 + \theta - 2(1 + \theta)\sqrt{\frac{\log(1/\delta)}{n}} =: \tau_1$$

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

$$\theta \geq C\sqrt{\frac{k \log(p/k)}{n}}$$
Minimax lower bound

Fix $\nu > 0$ (small).
Then there exists a constant $C_\nu > 0$ such that if

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

Then

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

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

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

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

Need an approximation...
Semidefinite relaxation 101

\[
\text{SDP}_{\max}^k(A) = \max. \quad \text{Tr}(x^\top A x)
\]
subject to \[
\begin{align*}
\text{Tr}(x^\top Z x) &= 1 \\
Z |x|_0 &\leq k
\end{align*}
\]

\[
Z = xx^\top \quad \text{rank}(Z) = 1 \quad Z \succeq 0
\]

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

Testing procedure: \(1\{\text{SDP}_k(\hat{\Sigma}) > \tau\}\)

Defined even if solution of SDP has rank > 1
Performance of SDP

For the **alternative**: relaxation of $\lambda_{\text{max}}^k(\hat{\Sigma})$ so

$$\text{SDP}_k(\hat{\Sigma}) \geq \lambda_{\text{max}}^k(\hat{\Sigma})$$

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

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

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

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

$$\text{MDP}_k(\hat{\Sigma}) = \min_{z \geq 0} \left\{ \lambda_{\text{max}}(\text{st}_z(\hat{\Sigma})) + kz \right\}$$
Upper bounds w.p. $1 - \delta$

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 $\theta$. When this ratio is larger than one, both type I and type II errors are below 5%.
Summary

No detection

Detection with $\lambda_{\text{max}}^k$

Detection with $\ast \text{DP}_k$

Can we tighten the gap?
Numerical evidence

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

$$\frac{k}{n} \log \left( \frac{p}{k} \right)$$
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 $2 \log N = 7.8$ asymptotically 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 \((\text{Alon, Arora, Feige, Hazan, Krauthgamer,}... \text{Cryptosystems are based on this fact!})\)

Conjecture: It is hard to find cliques of size between \(2 \log N\) and \(\sqrt{N}\)

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

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

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