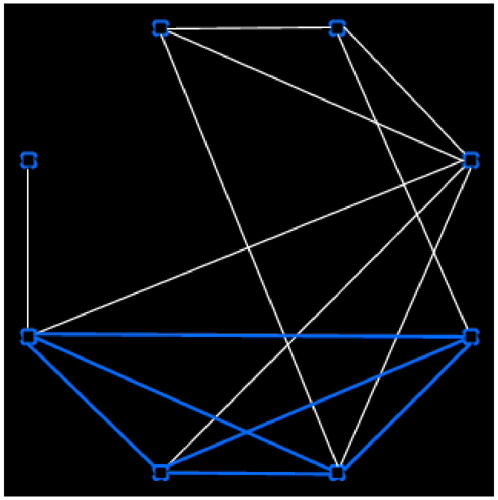
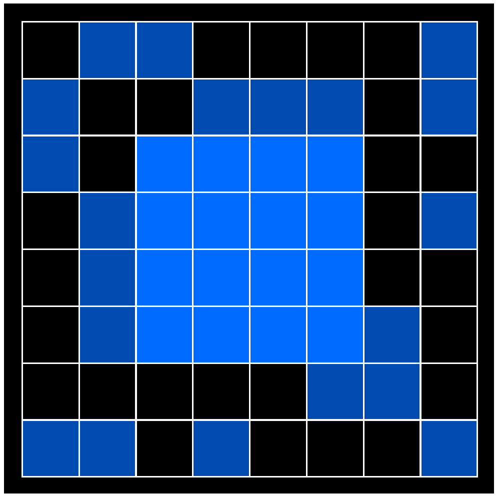
Statistical and computational trade-offs in estimation of sparse principal components

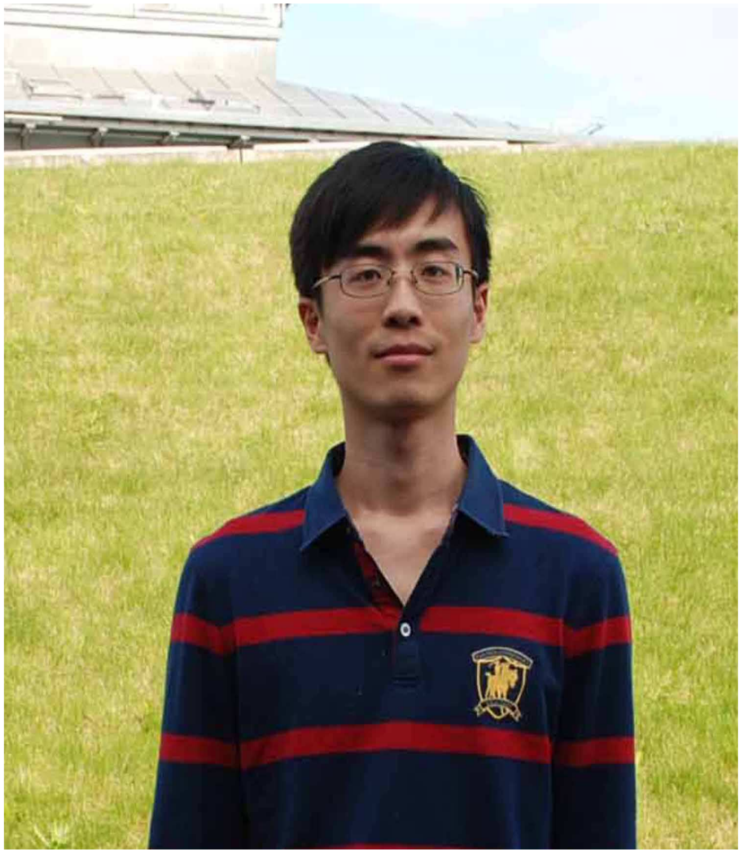




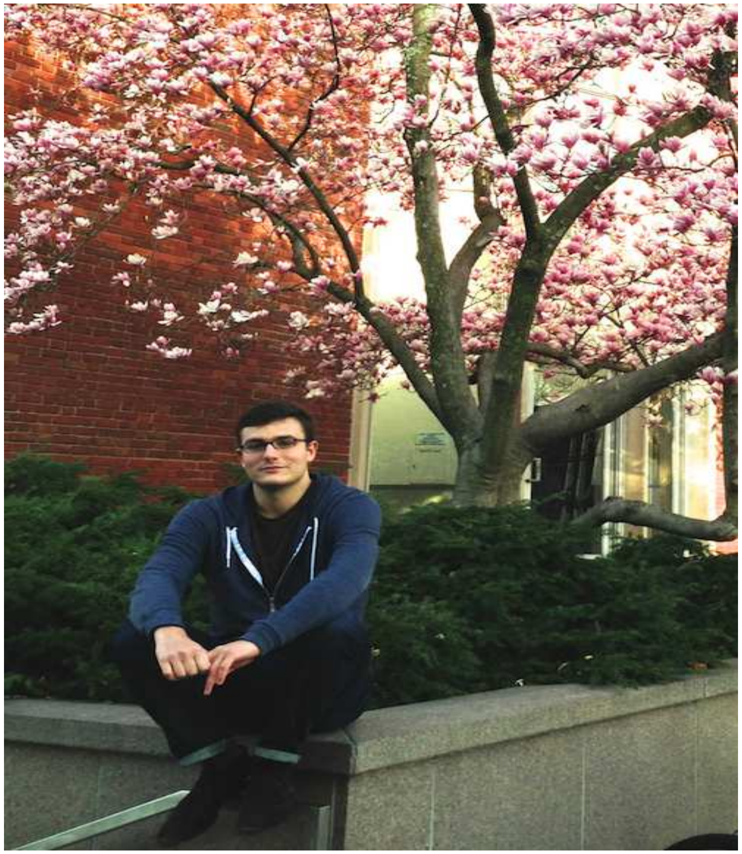
Richard Samworth

University of Cambridge

Collaborators



Tengyao Wang



Quentin Berthet

Statistical and computational trade-offs

Is there a fundamental trade-off between statistical and computational efficiency?

Statistical and computational trade-offs

Is there a fundamental trade-off between statistical and computational efficiency?

Could it be that no (randomised) polynomial time algorithm can attain the minimax rate?

Statistical and computational trade-offs

A growing body of work strongly suggests there is such a fundamental trade-off.

- **Sparse principal component detection** (Berthet and Rigollet, 2013)
- **Convex relaxation algorithms** (Chandrasekaran and Jordan, 2013)
- **Submatrix signal detection** (Ma and Wu, 2013; Chen and Xu, 2014)
- **Sparse linear regression** (Zhang et al., 2014)
- **Community detection** (Hajek et al., 2014).

The area introduces new connections between Statistics and theoretical computer science.

Principal Components Analysis (PCA)

Let $p \geq 2$ and $\mathcal{P}$ denote all distributions P on $\mathbb{R}^p$ with $\int_{\mathbb{R}^p} x \, dP(x) = 0$ and $\Sigma(P) := \int_{\mathbb{R}^p} xx^\top \, dP(x)$ finite.

Let $\lambda_1(P) > \lambda_2(P) \geq \ldots \geq \lambda_p(P) \geq 0$ denote the eigenvalues of Σ and let $v_1(P)$ denote the *first principal component* — an eigenvector corresponding to $\lambda_1(P)$.

If $X_1, \ldots, X_n \overset{iid}{\sim} P$, then we can estimate $v_1(P)$ using the top eigenvector, $\hat{v}$, of $\hat{\Sigma} := n^{-1} \sum_{i=1}^n X_i X_i^\top$.

PCA fails in high dimensions

For unit vectors $u, v \in \mathbb{R}^p$**, let** $\Theta(u,v) := \cos^{-1}(|u^\top v|)$**, and**

$$L(u,v) := \sin\Theta(u,v) = \{1-(u^\top v)^2\}^{1/2} = \frac{1}{\sqrt{2}}\|uu^\top - vv^\top\|_2.$$

Let $P = N_p(0, I_p + \theta vv^\top)$**, where** $\theta > 0$**. If** $p = p_n \to \infty$ **with** $p/n \to c \in (0,1)$**, then**

$$L(\hat{v}, v)^2 \xrightarrow{\text{a.s.}} \begin{cases} 1 - \frac{1-c/\theta^2}{1+c/\theta} & \textbf{if } \theta > \sqrt{c} \\ 1 & \textbf{if } \theta \le \sqrt{c} \end{cases}$$

(Paul, 2007; Johnstone and Lu, 2009).

Sparse PCA

Sparse PCA (Zou et al., 2006) **is designed to remedy the situation and aid interpretability in high dimensions.**

Assume $v_1(P) \in B_0(k)$ **where** $k \ll p$**, and**

$$B_0(k) := \left\{ u = (u_1, \ldots, u_p)^\top : \sum_{j=1}^{p} \mathbb{1}_{\{u_j \neq 0\}} \leq k,\ \|u\|_2 = 1 \right\}.$$

Much work on theoretical properties, e.g. minimax rates over subgaussian classes (Vu and Lei, 2013; Cai et al., 2013)**.**

Restricted Covariance Concentration

The *directional variance* of P along a unit vector $u \in \mathbb{R}^p$ is $V(u) := u^\top \Sigma u$, with empirical counterpart $\hat{V}(u) := u^\top \hat{\Sigma} u$.

For $\ell \in \{1, \ldots, p\}$ and $C \in (0, \infty)$, say $P \in \mathrm{RCC}_p(n, \ell, C)$ if

$$\mathbb{P}\biggl\{ \sup_{u \in B_0(\ell)} |\hat{V}(u) - V(u)| \geq C \max\biggl(\sqrt{\frac{\ell \log(p/\delta)}{n}}, \frac{\ell \log(p/\delta)}{n} \biggr) \biggr\} \leq \delta$$

for all $\delta > 0$. Further, let

$$\mathrm{RCC}_p(\ell, C) := \cap_{n \in \mathbb{N}} \mathrm{RCC}_p(n, \ell, C);$$
$$\mathrm{RCC}_p(C) := \cap_{\ell=1}^{p} \mathrm{RCC}_p(\ell, C).$$

Subgaussian distributions satisfy RCC

If Q **is a mean zero distribution on** $\mathbb{R}^p$ **and** $Y \sim Q$**, write** $Q \in \mathrm{subgaussian}(\sigma^2)$ **if** $\mathbb{E}(e^{u^\top Y}) \leq e^{\sigma^2 \|u\|^2/2}$ **for all** $u \in \mathbb{R}^p$**.**

For every $\sigma \in (0, \infty)$**, we have**

$$\mathrm{subgaussian}_p(\sigma^2) \subseteq \mathrm{RCC}_p\biggl(16\sigma^2\Bigl(1 + \frac{9}{\log p}\Bigr)\biggr).$$

Minimax rates over RCC classes

Let

$$\mathcal{P}_p(n,k,\theta) := \{P \in \mathrm{RCC}_p(2,1) \cap \mathrm{RCC}_p(2k,1) : \\ v_1(P) \in B_0(k), \lambda_1(P) - \lambda_2(P) \geq \theta\}.$$

For $7 \leq k \leq p^{1/2}$ **and** $0 < \theta \leq \frac{1}{16(1+\frac{9}{\log p})}$**, we have**

$$\inf_{\hat{v}} \sup_{P \in \mathcal{P}_p(n,k,\theta)} \mathbb{E}_P L\bigl(\hat{v}, v_1(P)\bigr) \geq \min\biggl\{\frac{1}{1660}\sqrt{\frac{k \log p}{n\theta^2}}, \frac{5}{18\sqrt{3}}\biggr\}.$$

(Vu and Lei, 2013). ■

Minimax rates over RCC classes II

Let $\hat{v}^k_{\max}(\hat{\Sigma}) := \mathrm{sargmax}_{u \in B_0(k)}\, u^\top \hat{\Sigma} u$ **be the** k**-sparse maximum eigenvector of** $\hat{\Sigma}$**.**

If $2k \log p \leq n$**, then** $\hat{v}^k_{\max}(\hat{\Sigma})$ **satisfies**

$$\sup_{P \in \mathcal{P}_p(n,k,\theta)} \mathbb{E}_P L\bigl(\hat{v}^k_{\max}(\hat{\Sigma}), v_1(P)\bigr) \leq 2\sqrt{2}\Bigl(1 + \frac{1}{\log p}\Bigr)\sqrt{\frac{k \log p}{n\theta^2}}.$$

Computationally efficient estimators

Computing $\hat{v}^k_{\max}(\hat{\Sigma})$ involves a search of all $k \times k$ submatrices of $\hat{\Sigma}$.

Let $\mathcal{M}_1 := \{M \in \mathbb{R}^{p \times p} : M \succeq 0, \mathrm{tr}(M) = 1\}$ and $\mathcal{M}_{1,1}(k^2) := \{M \in \mathcal{M}_1 : \mathrm{rank}(M) = 1, \|M\|_0 \leq k^2\}$. Then

$$\max_{u \in B_0(k)} u^\top \hat{\Sigma} u = \max_{u \in B_0(k)} \mathrm{tr}(\hat{\Sigma} u u^\top) = \max_{M \in \mathcal{M}_{1,1}(k^2)} \mathrm{tr}(\hat{\Sigma} M).$$

Drop rank constraint and replace sparsity constraint with ℓ_1 penalty (d'Aspremont et al., 2007) to obtain $\hat{v}^{\mathrm{SDP}}$.

Computing $\hat{v}^{\mathrm{SDP}}$

Algorithm 1: Pseudo-code for computing $\hat{v}^{\mathrm{SDP}}$

Input: $\mathbf{X} = (X_1, \ldots, X_n)^\top \in \mathbb{R}^{n \times p}$, $\lambda > 0, \epsilon > 0$

begin

Step 1: Set $\hat{\Sigma} \leftarrow n^{-1}\mathbf{X}^\top\mathbf{X}$.

Step 2: For $f(M) := \mathrm{tr}(\hat{\Sigma}M) - \lambda\|M\|_1$**, find** $\hat{M} \in \mathcal{M}_1$ **with** $f(\hat{M}) \geq \max_{M \in \mathcal{M}_1} f(M) - \epsilon$.

Step 3: Let $\hat{v}^{\mathrm{SDP}} := \hat{v}^{\mathrm{SDP}}_{\lambda,\epsilon} \leftarrow \mathrm{sargmax}_{u:\|u\|_2=1}\, u^\top \hat{M} u$.

end

Output: $\hat{v}^{\mathrm{SDP}}$

Further detail on Step 2

Rewrite

$$\max_{M \in \mathcal{M}_1} \mathrm{tr}(\hat{\Sigma} M) - \lambda \|M\|_1 = \max_{M \in \mathcal{M}_1} \min_{U \in \mathcal{U}} \mathrm{tr}\big((\hat{\Sigma} + U)M\big),$$

where $\mathcal{U} := \{U \in \mathbb{R}^{p \times p} : U^\top = U, \|U\|_\infty \le \lambda\}$. Since RHS is linear in both M and U, we can use proximal gradient methods (Nemirovski, 2004) to obtain after N iterations that

$$\max_{M \in \mathcal{M}_1} \min_{U \in \mathcal{U}} \mathrm{tr}\big((\hat{\Sigma} + U)M\big) - \min_{U \in \mathcal{U}} \mathrm{tr}\big((\hat{\Sigma} + U)\hat{M}\big) \le \frac{\lambda^2 p^2 + 1}{\sqrt{2}N}.$$

Thus Step 2 takes $O(\frac{\lambda^2 p^2 + 1}{\epsilon})$ iterations in the worst case. With $\lambda := 4\sqrt{\frac{\log p}{n}}$ and $\epsilon := \frac{\log p}{4n}$, the overall algorithm has worst-case complexity $O\big(\max(p^5, \frac{np^3}{\log p})\big)$.

Risk bound for $\hat{v}^{\mathrm{SDP}}$

For an arbitrary $P \in \mathcal{P}_p(n,k,\theta)$ **and** $X_1,\ldots,X_n \stackrel{iid}{\sim} P$**, let** $\hat{v}^{\mathrm{SDP}}(\mathbf{X})$ **denote the output of Algorithm 1 with input** $\mathbf{X} := (X_1,\ldots,X_n)^\top$, $\lambda := 4\sqrt{\frac{\log p}{n}}$ **and** $\epsilon := \frac{\log p}{4n}$**.**

If $4\log p \le n \le k^2 p^2 \log p$**, and** $\theta \in (0,1]$**, then**

$$\sup_{P \in \mathcal{P}_p(n,k,\theta)} \mathbb{E}_P L\bigl(\hat{v}^{\mathrm{SDP}}(\mathbf{X}), v_1(P)\bigr) \le (16\sqrt{2}+2)\sqrt{\frac{k^2 \log p}{n\theta^2}}.$$

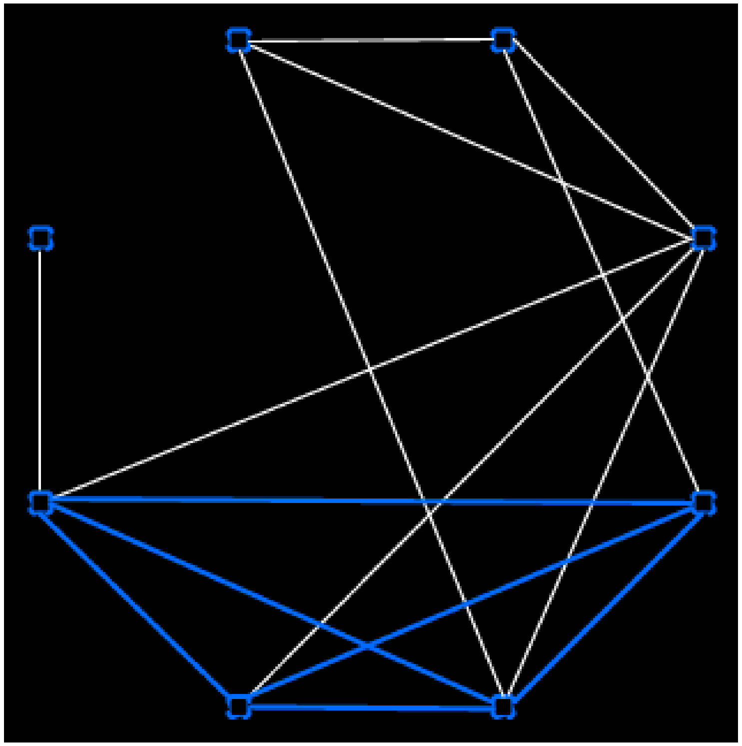
The planted clique problem

Let $\mathbb{G}_m$ be the set of all undirected graphs with m vertices.

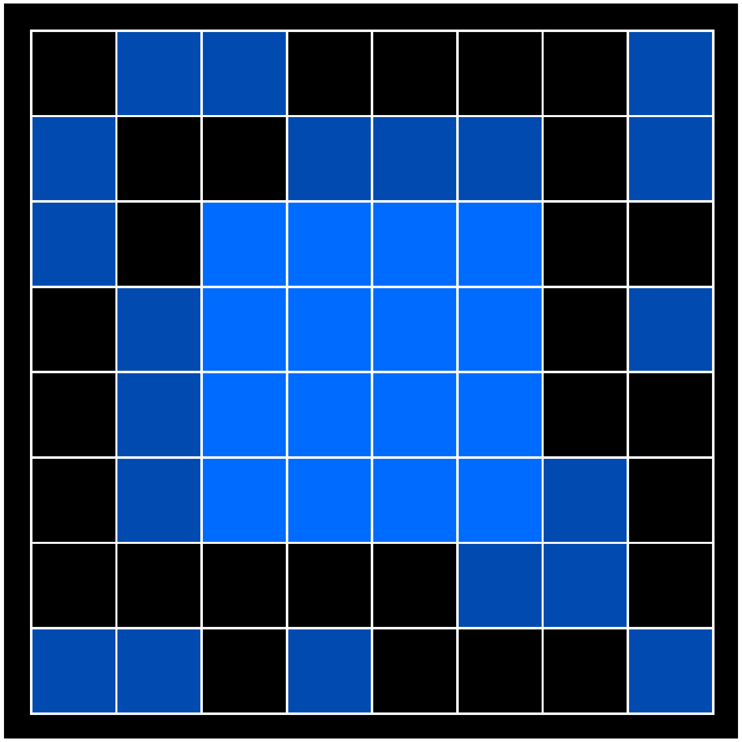
For $\kappa \in \{1, \ldots, m\}$, the planted clique distribution picks κ vertices uniformly and connects all edges between these vertices (the 'planted clique'). All other pairs of vertices are joined independently with probability $1/2$.

Can we locate the planted clique quickly?

Planted clique

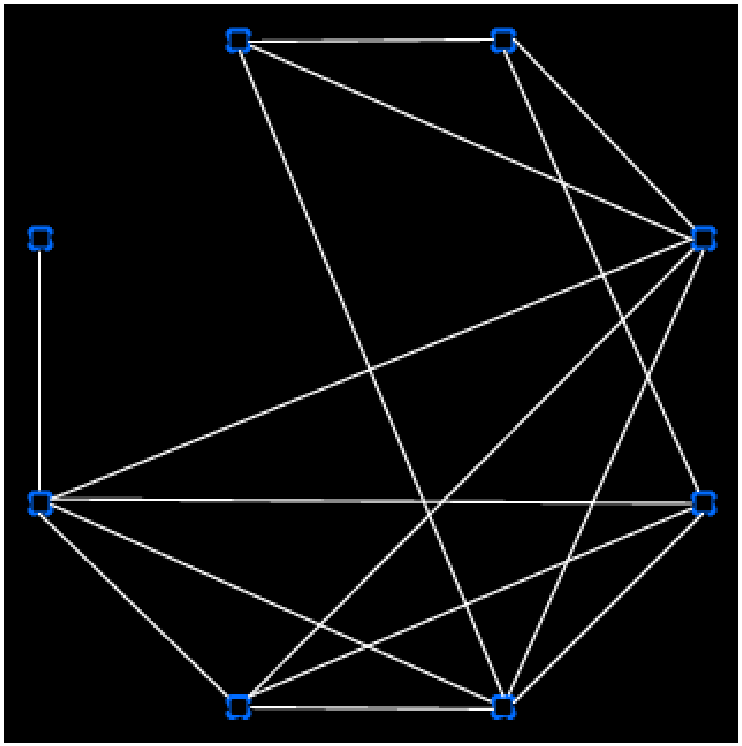


Graph

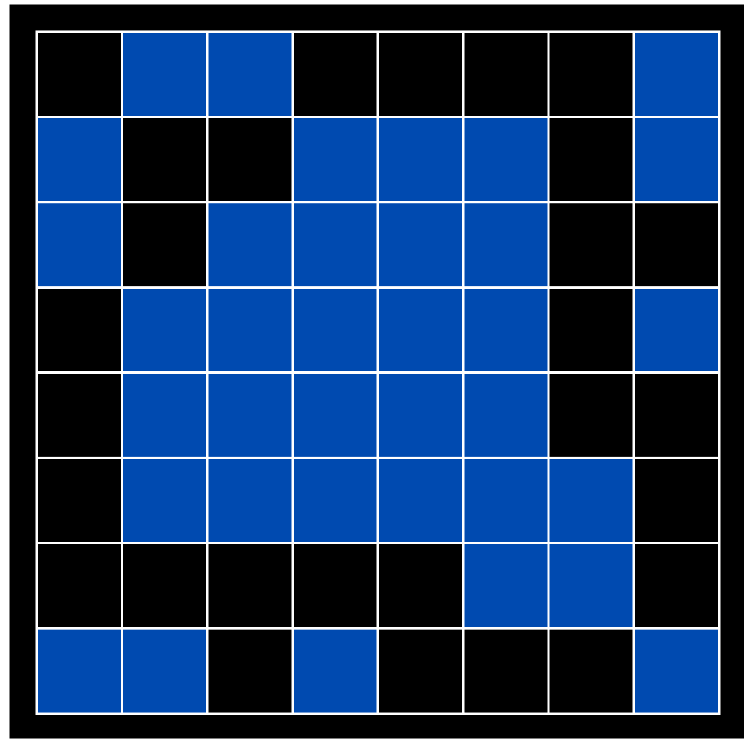


Adjacency matrix

Planted clique



Graph



Adjacency matrix

Finding the planted clique is hard

For a standard Erdős–Rényi graph, the maximal clique K_m satisfies $\frac{|K_m|}{2\log_2 m} \xrightarrow{\text{a.s.}} 1$. If $\liminf_{m\to\infty} \frac{\kappa}{2\log_2 m} > 1$, the planted clique is a.a.s. the unique maximum clique.

If $\kappa > C\sqrt{m \log m}$, then a.a.s., the planted clique vertices have largest degree (Kučera, 1995).

If $\kappa > c\sqrt{m}$ for some $c > 0$, then spectral-based methods can find planted clique a.a.s. (Alon et al., 1998).

No known randomised polynomial time algorithms when $\kappa = o(\sqrt{m})$, and substantial evidence against their existence (Jerrum, 1992; Feige and Krauthgamer, 2003; Feldman et al., 2013).

Planted clique hypothesis

(A1) **For any sequence $\kappa = \kappa_m$ such that $\kappa \leq m^\beta$ for some $0 < \beta < 1/2$, there is no randomised polynomial time algorithm that can correctly identify the planted clique with probability tending to 1 as $m \to \infty$.**

Similar (often stronger) hypotheses have been used in theoretical computer science in

- **testing k-wise independence** (Alon et al., 2007)
- **approximating Nash equilibria** (Hazan and Krauthgamer, 2011)
- **sparse submatrix detection** (Ma and Wu, 2013)
- **in cryptographic applications** (e.g. Juels and Peinado, 2000).

Computational lower bound

Assume (A1) and let $\alpha \in (0,1)$**. Let** $k := \lfloor n^{2/(5-\alpha)} \rfloor$**,** $p := n$ **and** $\theta := n^{(1-\alpha)/(5-\alpha)}/1000$**. For** $P \in \mathcal{P}_p(n,k,\theta)$**, let** $\boldsymbol{X}$ **be an** $n \times p$ **matrix with independent rows having distribution** P**. Then every sequence** $(\hat{v}^{(n)})$ **of randomised polynomial time estimators of** $v_1(P)$ **satisfies**

$$\sqrt{\frac{n\theta^2}{k^{1+\alpha}\log p}} \sup_{P \in \mathcal{P}_p(n,k,\theta)} \mathbb{E}_P L\bigl(\hat{v}^{(n)}(\boldsymbol{X}), v_1(P)\bigr) \to \infty.$$

Algorithm 2: Pseudo-code for a planted clique algorithm based on a hypothetical randomised polynomial time sparse principal component estimation algorithm.

Input: $m \in \mathbb{N}$, $\kappa \in \{1, \ldots, m\}$, $G \in \mathbb{G}_m$, $L \in \mathbb{N}$

begin

Step 1: Let $n \leftarrow \lfloor 9m/(10L) \rfloor$, $p \leftarrow n$, $k \leftarrow \lfloor \kappa/L \rfloor$. Draw $u_1, \ldots, u_n, w_1, \ldots, w_p$ uniformly at random without replacement from $V(G)$. Form $\mathbf{A} = (A_{ij}) \leftarrow (\mathbb{1}_{\{u_i \sim w_j\}}) \in \mathbb{R}^{n \times p}$ and $\mathbf{X} \leftarrow \mathrm{diag}(\xi_1, \ldots, \xi_n)(2\mathbf{A} - \mathbf{1}_{n \times p})$, where $\xi_1, \ldots, \xi_n$ are independent Rademacher random variables

Step 2: Use the estimator $\hat{v}^{(n)}$ to compute $\hat{v} = \hat{v}^{(n)}(\mathbf{X}/\sqrt{750})$.

Step 3: Let $\hat{S} = \hat{S}(\hat{v})$ be the lexicographically smallest k-subset of $\{1, \ldots, p\}$ such that $(\hat{v}_j : j \in \hat{S})$ contains the k largest coordinates of $\hat{v}$ in absolute value.

Step 4: Let $\mathrm{nb}(u, W) := \mathbb{1}_{\{u \in W\}} + \sum_{w \in W} \mathbb{1}_{\{u \sim w\}}$ for $u \in V$ and $W \subseteq V$. Set $\hat{K} := \bigl\{u \in V : \mathrm{nb}(u, \{w_j : j \in \hat{S}\}) \geq 3k/4\bigr\}$.

end

Output: $\hat{K}$

Proof heuristics

Let $L := \lceil \log n \rceil$**, let** $m := \lceil 10Lp/9 \rceil$ **and** $\kappa := Lk$**. Let** $(\boldsymbol{\epsilon}, \boldsymbol{\gamma}) = (\epsilon_1, \ldots, \epsilon_n, \gamma_1, \ldots, \gamma_p)$ **be independent Bern(**κ/m**).**

Let $\mathbf{Y} = (Y_1, \ldots, Y_n)^\top$ **have independent components conditional on** γ**, each distributed as**

$$Y := \xi\big\{(1-\epsilon)R + \epsilon(\boldsymbol{\gamma} + \tilde{R})\big\},$$

where ξ**,** ϵ **and** R **are independent,** ξ **is a Rademacher random variable,** $\epsilon \sim \mathrm{Bern}(\kappa/m)$**,** $R = (R_1, \ldots, R_p)^\top$ **has independent Rademacher components, and** $\tilde{R} = (\tilde{R}_1, \ldots, \tilde{R}_p)^\top$ **with** $\tilde{R}_j := (1-\gamma_j)R_j$**.**

Then $d_{\mathrm{TV}}\big(\mathcal{L}(\boldsymbol{X}), \mathcal{L}(\boldsymbol{Y})\big) \leq 18/(5L)$ **(Diaconis and Freedman, 1980), and** $Q_{\boldsymbol{\gamma}} := \mathcal{L}(Y|\boldsymbol{\gamma}) \in \cap_{\ell=1}^{\lfloor 20p/(9k) \rfloor} \mathrm{RCC}_p(\ell, 750)$**.**

Proof heuristics II

Suppose the r.p.t. estimator $\hat{v}^{(n)}$ **of** $v_1(P)$ **satisfied**

$$\sup_{P \in \mathcal{P}_p(n,k,\theta)} \mathbb{E}_P L\bigl(\hat{v}^{(n)}(\mathbf{X}), v_1(P)\bigr) \leq K_0 \sqrt{\frac{k^{1+\alpha}\log p}{n\theta^2}}.$$

Let $N_\gamma := \sum_{j=1}^p \gamma_j$ **and** $\Gamma_0 := \{g : |N_g - p\kappa/m| \leq k/20\}$**. If** $g \in \Gamma_0$**, then** $\mathcal{L}(\frac{Y_1}{\sqrt{750}}|\gamma = g) \in P_p(n,k,\theta)$ **for** $\theta \leq \frac{\kappa}{750m}(N_g - 1)$ **and large** $n \in \mathcal{N}$**. So**

$$\mathbb{E}\Bigl\{L\Bigl(\hat{v}^{(n)}\bigl(\tfrac{\boldsymbol{Y}}{\sqrt{750}}\bigr), v_1(Q_\gamma)\Bigr) \Bigm| \gamma = g\Bigr\} \leq 1000K_0 n^{-\frac{5(1-\alpha)}{2(5-\alpha)}}\sqrt{\log n}.$$

Deduce that $\bigl|\{j \in \hat{S}\bigl(\hat{v}^{(n)}(\boldsymbol{X}/\sqrt{750})\bigr) : w_j \in K\}\bigr| > 3k/4$ **w.h.p. and** $\mathbb{P}(\hat{K} \neq K) \to 0$**.**

Summary

- We introduce new classes of distributions for studying the estimation problem in Sparse PCA.
- Minimax rates are obtained, but the upper bound is only attained by a super-polynomial time procedure.
- Under a Planted Clique Assumption, rates of convergence for randomised polynomial time algorithms are necessarily worse.

References

- Alon, N., Andoni, A., Kaufman, T., Matulef, K., Rubinfeld, R., and Xie, N. (2007) Testing k-wise and almost k-wise independence.*Proceedings of the thirty-ninth ACM Symposium on Theory of Computing*, 496–505.
- Alon, N., Krivelevich, M. and Sudakov, B. (1998) Finding a large hidden clique in a random graph. *Proceedings of the ninth annual ACM-SIAM symposium on Discrete algorithms*, 594–598.
- Berthet, Q. and Rigollet P. (2013) Complexity theoretic lower bounds for sparse principal component detection. *J. Mach. Learn. Res. (COLT)*, 30, 1046–1066.
- Cai, T. T., Ma, Z. and Wu, Y. (2013) Sparse PCA: Optimal rates and adaptive estimation. *Ann. Statist.*, 41, 3074–3110.
- Chandrasekaran, V. and Jordan, M. I. (2013) Computational and statistical tradeoffs via convex relaxation. *Proc. Nat. Acad. Sci.*, 110, E1181–E1190.

- **Chen, Y. and Xu, J. (2014) Statistical-computational tradeoffs in planted problems and submatrix localization with a growing number of clusters and submatrices. Available at `http://arxiv.org/abs/1402.1267`.**

- **dAspremont, A., El Ghaoui, L., Jordan, M. I. and Gert R. G. Lanckriet (2007). A direct formulation for sparse PCA using semidefinite programming. *SIAM review*, 49, 434–448.**

- **Feige, U. and Krauthgamer, R. (2003) The probable value of the Lovàsz–Schrijver relaxations for a maximum independent set. *SIAM J. Comput.*, 32, 345–370.**

- **Feldman, V., Grigorescu, E., Reyzin, L., Vempala, S. S. and Xiao, Y. (2013) Statistical Algorithms and a Lower Bound for Detecting Planted Cliques. *Proceedings 45th Annual ACM Symposium on Theory of Computing*, 655–664.**

- **Hajek, B., Wu, Y. and Xu, J. (2014) Computational lower bounds for community detection on random graphs. Available at `http://arxiv.org/abs/1406.6625`.**

- **Hazan, E. and Krauthgamer, R. (2011) How hard is it to approximate the best nash equilibrium?**

SIAM J. Comput., 40, 79–91.

- Jerrum, M. (1992) Large cliques elude the Metropolis process. *Random Structures Algorithms*, 3, 347–359.
- Johnstone, I. M. and Lu, A. Y. (2009) On consistency and sparsity for principal components analysis in high dimensions. *J. Amer. Statist. Assoc.*, 104, 682–693.
- Juels, A. and Peinado, M. (2000) Hiding cliques for cryptographic security. *Des. Codes Cryptogr.*, 20, 269-280.
- Kučera, L. (1995) Expected complexity of graph partitioning problems. *Discrete Appl. Math.*, 57, 193–212.
- Ma, Z. and Wu, Y. (2013) Computational barriers in minimax submatrix detection. Available at `http://arxiv.org/abs/1309.5914`.
- Nemirovski, A. (2004) Prox-method with rate of convergence $O(1/t)$ for variational inequalities with

Lipschitz continuous monotone operators and smooth convex-concave saddle point problems. *SIAM J. Optim.* 15, 229–251.

- **Paul, D. (2007) Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. *Statist. Sinica*, 17, 1617–1642.**
- **Vu, V. Q. and Lei, J. (2013) Minimax sparse principal subspace estimation in high dimensions. *Ann. Statist.*, 41, 2905–2947.**
- **Wang, T., Berthet, Q. and Samworth, R. J. (2014) Statistical and computational trade-offs in estimation of sparse principal components. Available at `http://arxiv.org/abs/1408.5369`.**
- **Zhang, Y., Wainwright, M. J. and Jordan, M. I. (2014) Lower bounds on the performance of polynomial-time algorithms for sparse linear regression. Available at `http://arxiv.org/abs/1402.1918`.**
- **Zou, H., Hastie, T. and Tibshirani, R. (2006) Sparse Principal Component Analysis. *J. Comp. Graph. Statist.*, 15, 265–286.**