# Sparse PCA – theory and practice

Tengyao Wang

University College London

SUSTech Stats-DS Talk

4 Nov 2020

# Collaborators

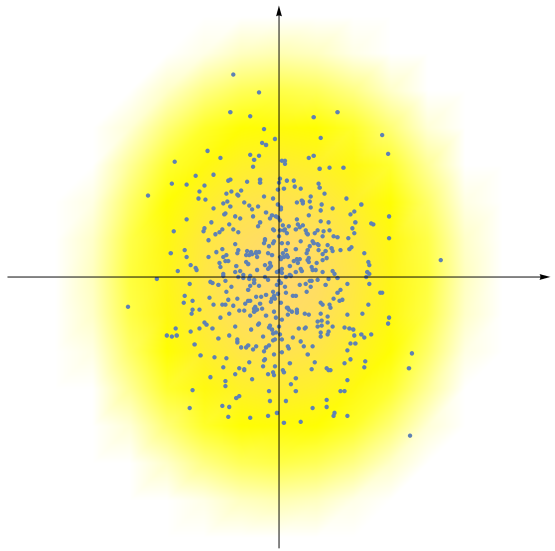
Quentin Berthet
Google Brain (Pairs)

Milana Gataric
EBI (Cambridge)

Richard Samworth
University of Cambridge

# Sparse PCA

$X_1, \ldots, X_n \in \mathbb{R}^p$ independent centred Gaussians with an unknown covariance matrix $\Sigma$.



$$X \sim N_p(0, \Sigma)$$

$\Sigma$ has spectral gap $\theta > 0$ and a $k$-sparse leading eigenvector

$$v \in B_0(k) = \{u : \|u\|_2 = 1, \|u\|_0 \leq k\}.$$

Estimation problem: estimate $v$ using $X_1, \ldots, X_n$.

Loss function: $L(\hat{v}, v) = \sin \Theta(\hat{v}, v)$

# Sparse PCA

Many different estimators have been proposed:

- SCoTLASS estimator (Jolliffe, Trendafilov and Uddin, 2003)
- Sparse linear regression based estimator (Zou, Hastie and Tibshirani, 2006)
- Semidefinite relaxation estimator (d'Aspremont et al. 2007)
- Diagonal thresholding estimator (Johnstone and Lu, 2009)
- Iterative thresholding estimator (Ma, 2013)
- ...

Applications in high-dimensional data sets:

- Signal processing (Majumdar, 2009)
- Computer vision (Wang, Lu and Yang, 2013; Naikal, Yang and Sastry, 2011)
- Biomedical research (Chun and Sündüz, 2009; Tan, Petersen and Witten, 2014)
- ...

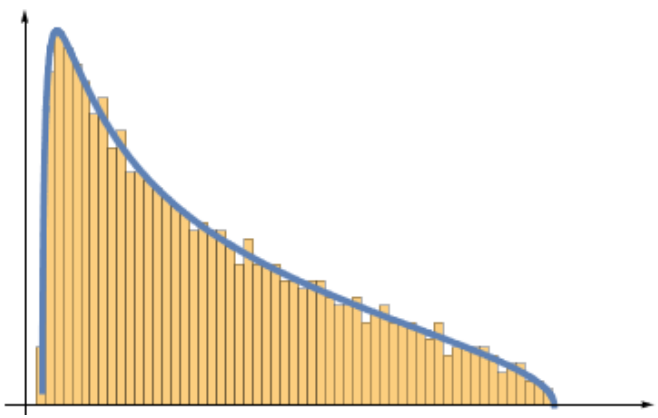
# Motivation for introducing sparsity

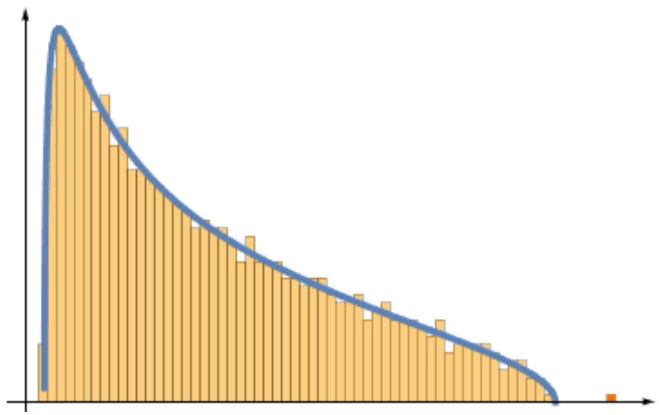
Why sparse PCA?

- Applications: enhanced interpretability of the principal components
- Theory: classical PCA is inconsistent in high dimensional settings.

$$\Sigma = I_p + \theta vv^\top, \qquad p/n \to c$$

Spectrum of $\hat{\Sigma} = n^{-1}\sum_{i=1}^n X_i X_i^\top$





$$\theta \le \sqrt{c} : L(\hat{v}_{\max}, v) \to 1 \qquad \theta > \sqrt{c} : L(\hat{v}_{\max}, v) \to \frac{c + c/\theta}{c + \theta}$$

# Sparse leading eigenvector estimator

Maximum likelihood estimator

$$\hat{v} = \hat{v}^k_{\max}(\hat{\Sigma}) = \arg\max_{u \in B_0(k)} u^\top \hat{\Sigma} u.$$

By a curvature lemma from Vu and Lei (2013),

$$L(\hat{v}, v)^2 = \|\hat{v}\hat{v}^\top - vv^\top\|_{\mathsf{F}}^2 \le \frac{2}{\theta} \mathbf{tr}\big((\hat{\Sigma} - \Sigma)(\hat{v}\hat{v}^\top - vv^\top)\big).$$

Upper bound the loss using empirical process theory

$$\mathbb{E} L(\hat{v}, v) \le \frac{4}{\theta} \mathbb{E} \sup_{u \in B_0(2k)} \big|u^\top (\hat{\Sigma} - \Sigma) u\big| \le C \sqrt{\frac{k \log p}{n\theta^2}}.$$

Key step: controlling the empirical process $u^\top(\hat{\Sigma} - \Sigma)u$ over $B_0(2k)$.

# Family of distributions

Restricted Covariance Concentration: $\mathbf{P} \in \mathrm{RCC}_p(n, \ell, A)$ if for all $\delta > 0$,

$$\mathbf{P}\left\{ \sup_{u \in B_0(\ell)} \left| u^\top (\hat{\Sigma} - \Sigma) u \right| \geq A \max\left( \sqrt{\frac{\ell \log(p/\delta)}{n}}, \frac{\ell \log(p/\delta)}{n} \right) \right\} \leq \delta.$$

Satisfied by subgaussian distributions.

$\mathbf{P} \in \mathcal{P}_p(n, k, \theta)$: distributions in $\mathrm{RCC}_p(n, 2k, 1)$ and $\mathrm{RCC}_p(n, 2, 1)$ with $k$-sparse leading eigenvector, spectral gap $\geq \theta$.

General upper bound: for $n \geq 2k \log p$,

$$\sup_{\mathbf{P} \in \mathcal{P}_p(n,k,\theta)} \mathbb{E}_{\mathbf{P}} L(\hat{v}^k_{\max}, v) \leq C \sqrt{\frac{k \log p}{n \theta^2}}.$$

# Minimax lower bound

The estimator $\hat{v}^k_{\max}$ is minimax optimal: for $k \leq \sqrt{p}$, $\theta$ bounded,

$$\inf_{\hat{v}} \sup_{\mathbf{P} \in \mathcal{P}_p(n,k,\theta)} \mathbb{E}_{\mathbf{P}} L(\hat{v}, v) \geq c \min\left(\sqrt{\frac{k \log p}{n\theta^2}}, 1\right).$$

Minimax optimal rate of estimation $\asymp \sqrt{\frac{k \log p}{n\theta^2}}$.

One problem remains: it is NP-hard to calculate $\hat{v}^k_{\max}$.

Especially problematic since sparse PCA is typically used on large datasets.

# Semidefinite programming

Semidefinite relaxation estimator: first studied by d'Aspremont et al. (2007), a polynomial time estimator.

Analogous to the $\ell_1$ relaxation used in sparse linear regression.

Original problem:

$$\begin{aligned} \hat{v}_{\max}^k \quad = \quad & \arg\max_u \quad u^\top \hat{\Sigma} u \\ & \text{subject to} \quad u^\top u = 1, \|u\|_0 \leq k. \end{aligned}$$

Non-convex problem.

# Semidefinite programming

Semidefinite relaxation estimator: first studied by d'Aspremont et al. (2007), a polynomial time estimator.

Analogous to the $\ell_1$ relaxation used in sparse linear regression.

Original problem:

$$\begin{aligned} \hat{v}_{\max}^k \quad = \quad & \arg\max_{u} \quad \mathbf{tr}\left(uu^\top \hat{\Sigma}\right) \\ & \text{subject to} \quad \mathbf{tr}(uu^\top) = 1, \|uu^\top\|_0 \leq k^2. \end{aligned}$$

Non-convex problem.

# Semidefinite programming

Semidefinite relaxation estimator: first studied by d'Aspremont et al. (2007), a polynomial time estimator.

Analogous to the $\ell_1$ relaxation used in sparse linear regression.

Matrix form:

$$\begin{aligned} \hat{M} \quad = \quad & \arg\max_{M} \quad \mathbf{tr}(\hat{\Sigma} M) \\ & \text{subject to} \quad \mathrm{rk}(M) = 1, \mathbf{tr}(M) = 1, \|M\|_0 \leq k^2, M \succeq 0. \end{aligned}$$

Two sources of non-convexity: rank constraint and $\ell_0$ constraint.

# Semidefinite programming

Semidefinite relaxation estimator: first studied by d'Aspremont et al. (2007), a polynomial time estimator.

Analogous to the $\ell_1$ relaxation used in sparse linear regression.

Matrix form (relaxed):

$$\begin{aligned} \hat{M} \quad = \quad & \arg\max_{M} \quad \mathbf{tr}(\hat{\Sigma} M) \\ & \text{subject to} \quad \mathbf{tr}(M) = 1, \|M\|_1 \leq k, M \succeq 0. \end{aligned}$$

Convex problem.

# Semidefinite programming estimator

Penalised version of the SDP estimator

$$\begin{aligned} \hat{M} &= \arg\max_{M} \quad \mathbf{tr}(\hat{\Sigma} M) - \lambda \|M\|_1 \\ &\qquad \text{subject to} \quad \mathbf{tr}(M) = 1, M \succeq 0. \end{aligned}$$

$$\hat{v}^{\mathrm{SDP}} = \text{leading eigenvector of } \hat{M}.$$

Solve the SDP (up to statistical precision) by first-order proximal methods, e.g. Nemirovski (2004), Nesterov (2005).

Overall complexity $O(p^5 \vee np^3)$.

# Statistical properties of the SDP estimator

Choosing $\lambda = 4\sqrt{\frac{\log p}{n}}$ and $\epsilon = \frac{\log p}{4n}$, if $4\log p \le n \le k^2 p^2 \log p$ and $\theta \le 1$, then

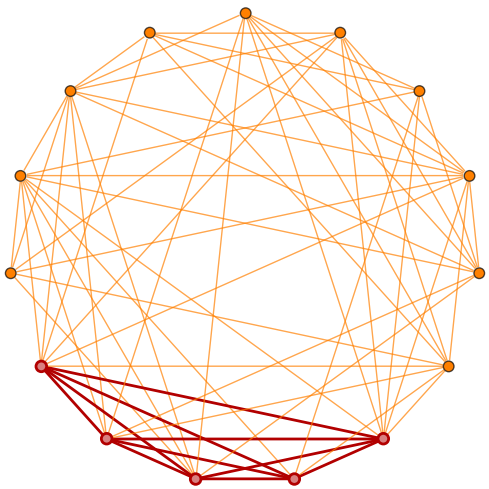
$$\sup_{\mathbf{P} \in \mathcal{P}_p(n,k,\theta)} \mathbb{E}_{\mathbf{P}} L(\hat{v}^{\mathrm{SDP}}, v) \le C\sqrt{\frac{k^2 \log p}{n\theta^2}}.$$

Computationally efficient, but statistically suboptimal.

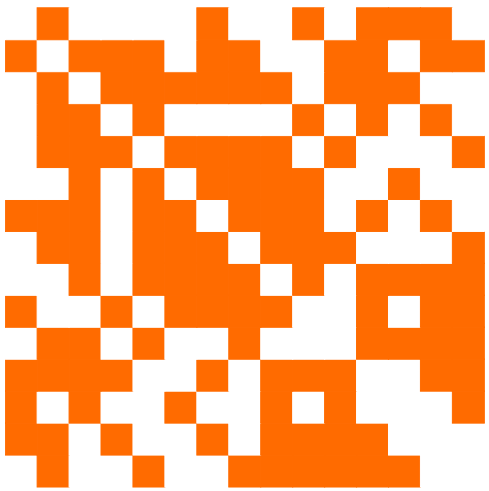
Can any (randomised) polynomial algorithm achieve the minimax rate? or a rate of the order $O\left(\sqrt{\frac{k^{1+\alpha} \log p}{n\theta^2}}\right)$ for any $0 < \alpha < 1$.

# A complexity theoretic problem

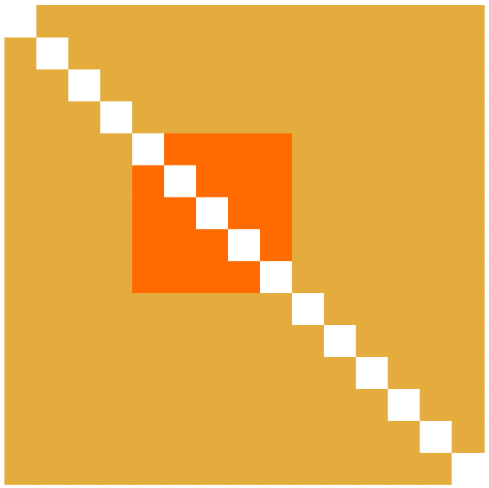
Planted Clique Problem: given $m$ vertices, select $\kappa$ of them to form a clique, then independently draw remaining edges with probability $1/2$. How to find the planted clique?



$G \sim \mathcal{G}_{m,\kappa}$



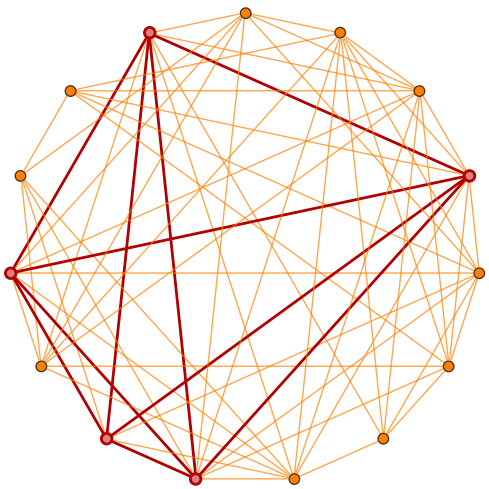
$\mathrm{Adj}(G)$



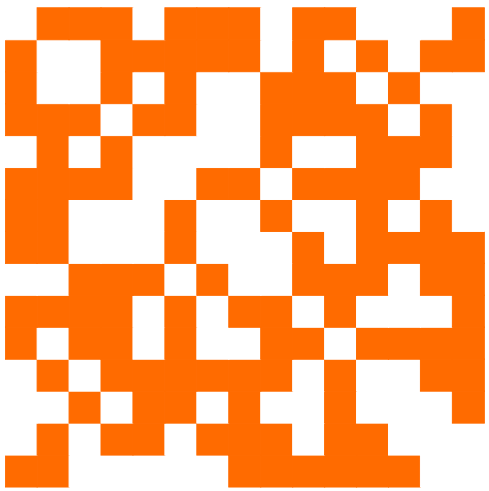
$\mathbb{E}\{\mathrm{Adj}(G)\}$

# A complexity theoretic problem

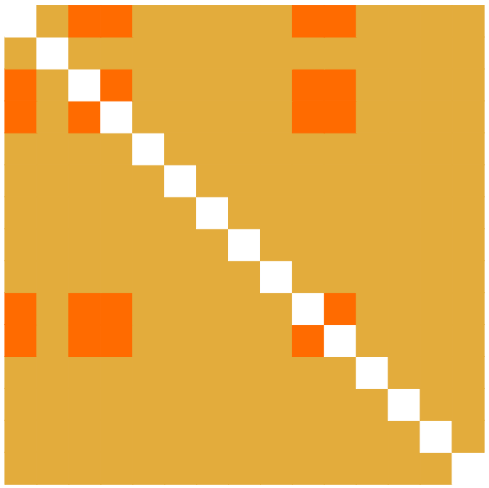
Planted Clique Problem: given $m$ vertices, select $\kappa$ of them to form a clique, then independently draw remaining edges with probability $1/2$. How to find the planted clique?



$G \sim \mathcal{G}_{m,\kappa}$



$\mathrm{Adj}(G)$



$\mathbb{E}\{\mathrm{Adj}(G)\}$

# A complexity theoretic problem

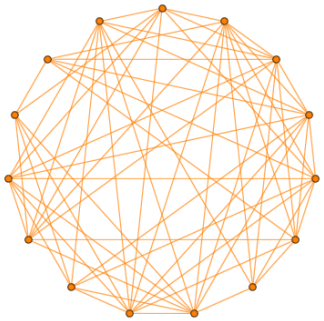
Planted Clique Problem: given $m$ vertices, select $\kappa$ of them to form a clique, then independently draw remaining edges with probability $1/2$. How to find the planted clique?

- $\kappa \geq (2+\delta)\log_2 m$: max clique
- $\kappa \geq c\sqrt{m}$: spectral methods
- $\kappa = O(m^{1/2-\delta})$: no known randomised polynomial time algorithms. Jerrum (1992), Feige and Krauthgamer (2003) and Feldman et al. (2013) show that some large subclasses of polynomial time algorithms will fail.

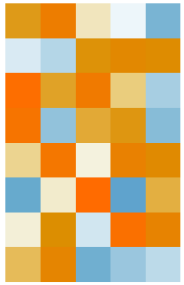
Planted Clique Hypothesis: For any sequence of $\kappa = \kappa_m$ such that $\kappa \leq m^{1/2-\delta}$, there is no randomised polynomial time algorithm that can identify the planted clique with asymptotic probability 1.

# A reduction argument

We use the hardness of the planted clique problem to derive a computational lower bound for the sparse PCA estimation problem.

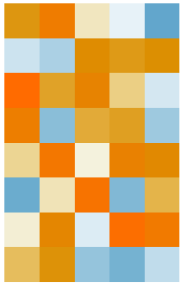


$G \sim \mathcal{G}_{m,\kappa}$



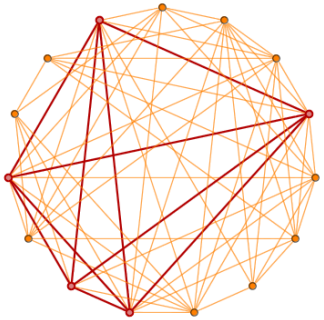
$X$

$\overset{d_{\mathrm{TV}}}{\approx}$



$Y \sim \mathbf{P} \in \mathcal{P}_p(n,k,\theta)$

$\downarrow$



$\longleftarrow$ $\hat{v}(X)$ $\overset{d_{\mathrm{TV}}}{\approx}$ $\hat{v}(Y)$

$\mathbb{E}L(\hat{v},v) \leq \sqrt{\frac{k^{1+\alpha}\log p}{n\,\theta^2}}$ will imply asymptotic probability 1 identification of the planted clique for $\kappa \asymp m^{1/2-\delta}$ for some $\delta > 0$ depending on $\alpha$.

# Details of the reduction

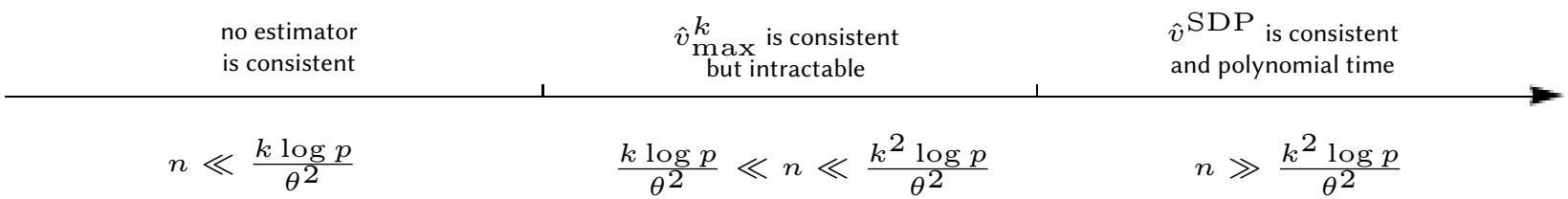
- $n = p \approx m/\log m$, $k \approx \kappa/\log m$.
- Take a random $n \times p$ submatrix $A$ of $\mathrm{Adj}(G)$ and change all $0$ to $-1$. Then independently flip signs of each row with probability $1/2$ to get matrix $X$.
- $X$ does not have independent rows, but a similar construction by 'sampling with replacement' gives $Y$ that has independent rows.
- A lemma by Diaconis and Freedman (1980) show $X$ and $Y$ are close in total variation distance, hence $\hat{v}(X)$ and $\hat{v}(Y)$ are close.
- Columns of $Y$ correspond to vertices of $G$. The $k$ columns that give rise to the largest coordiantes of $\hat{v}(Y)$ in absolute value correspond to a set of vertices in $G$ with high clique density.
- Reconstruct the entire clique from this vertex set of high clique density.

# Computational lower bound for sparse PCA

Theorem. Assume the Planted Clique Hypothesis, fix some $\alpha \in (0,1)$. If $k = O(p^{1/2-\delta})$, $n = o(p \log p)$, $\theta \le k^2/(1000p)$ and $\frac{k^{(1+\alpha)} \log p}{n\theta^2} \to 0$, then any sequence of randomised polynomial time estimators $(\hat{v}^{(n)})$ satisfies

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

Take home message: the $O\left(\sqrt{\frac{k^2 \log p}{n\theta^2}}\right)$ rate achieved by $\hat{v}^{\mathrm{SDP}}$ is the best uniform rate that we can hope for.



# High effective sample size regime

For a subclass $\tilde{\mathcal{P}}_p(n,k,\theta) \subset \mathcal{P}_p(n,k,\theta)$, a variant of $\hat{v}^{\mathrm{SDP}}$ can achieve the minimax rate in the high effective sample size regime.

$\hat{v}^{\mathrm{MSDP}}$: obtain $\hat{M} = \arg\max_{M \succeq 0, \mathbf{tr}(M)=1} \mathbf{tr}(\hat{\Sigma} M) - \lambda \|M\|_1$, let $S = \{j : \hat{M}_j j > \tau\}$,

$$\hat{v}^{\mathrm{MSDP}}_{S^c} = 0, \qquad \hat{v}^{\mathrm{MSDP}}_{S} = \text{leading eigenvector of } \hat{\Sigma_{SS}}.$$

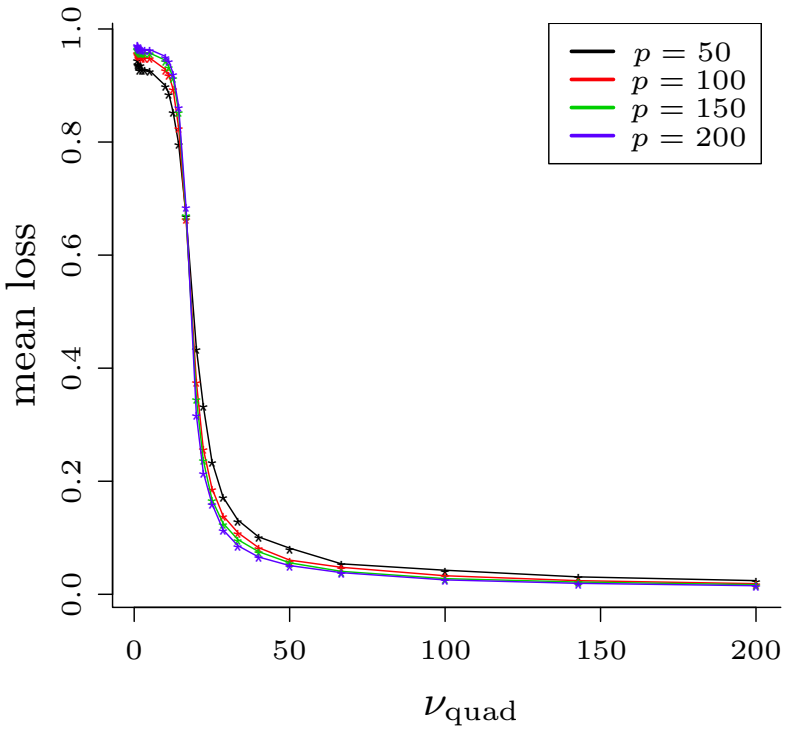
Performance of $\hat{v}^{\mathrm{MSDP}}$ in the high effective sample size regime: assume $\log p \le n$, $\theta^2 \le B\sqrt{k}$, $p \ge \theta\sqrt{n/k}$, set $\lambda = 4\sqrt{\frac{\log p}{n}}$, $\tau = \left(\frac{\log p}{Bn}\right)^2$,

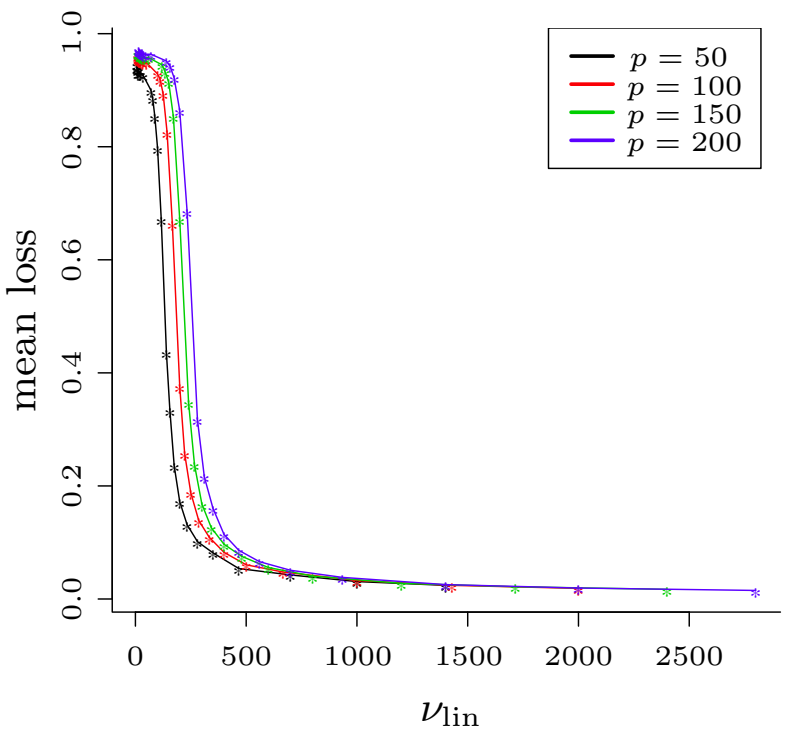
$$\sup_{\mathbf{P} \in \tilde{\mathcal{P}}_p(n,k,\theta)} \mathbb{E}_{\mathbf{P}} L(\hat{v}^{\mathrm{MSDP}}, v) \le C\sqrt{\frac{k \log p}{n\theta^2}}.$$

# Numerical experiments

$$X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} N_p(0, I_p + \theta v v^\top), v = (\tfrac{1}{\sqrt{k}}, \ldots, \tfrac{1}{\sqrt{k}}, 0, \ldots, 0)^\top.$$

Plot the average loss of $\hat{v}^{\mathrm{SDP}}$ against $\nu_{\mathrm{quad}} = \frac{n\theta^2}{k^2 \log p}$ or $\nu_{\mathrm{lin}} = \frac{n\theta^2}{k \log p}$

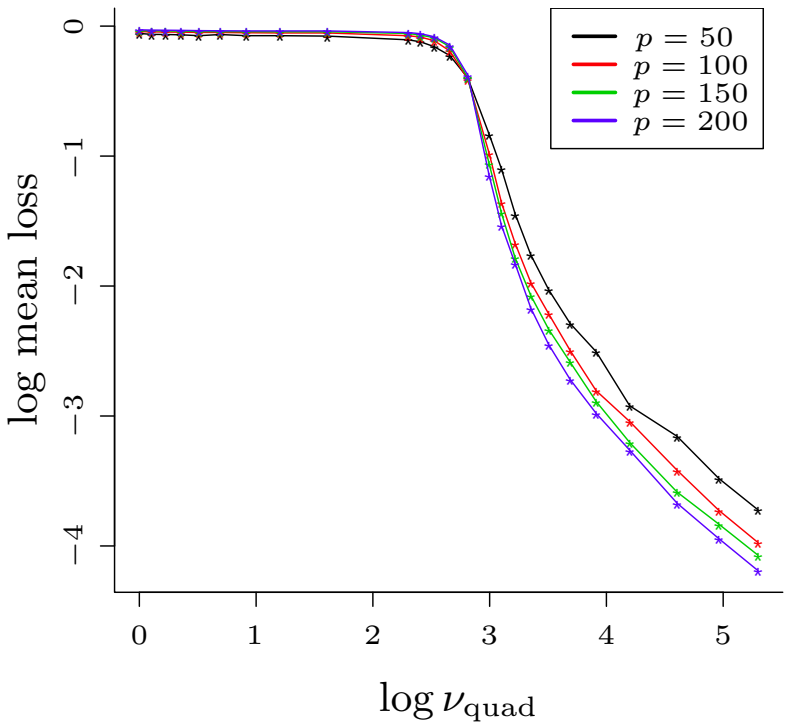


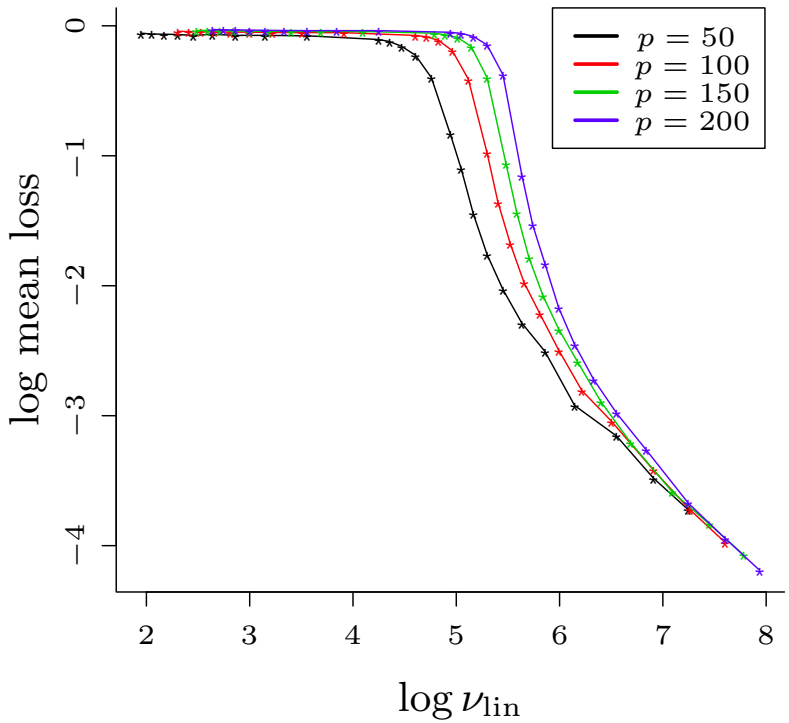



# Numerical experiments

$$X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} N_p(0, I_p + \theta v v^\top), v = (\tfrac{1}{\sqrt{k}}, \ldots, \tfrac{1}{\sqrt{k}}, 0, \ldots, 0)^\top.$$

Plot the average loss of $\hat{v}^{\mathrm{SDP}}$ against $\nu_{\mathrm{quad}} = \frac{n\theta^2}{k^2 \log p}$ or $\nu_{\mathrm{lin}} = \frac{n\theta^2}{k \log p}$



# Sparse PCA in practice

Our story so far: the SDP estimator is essentially the best polynomial-time estimator for sparse PCA.

In practice, almost no one uses the SDP estimator:

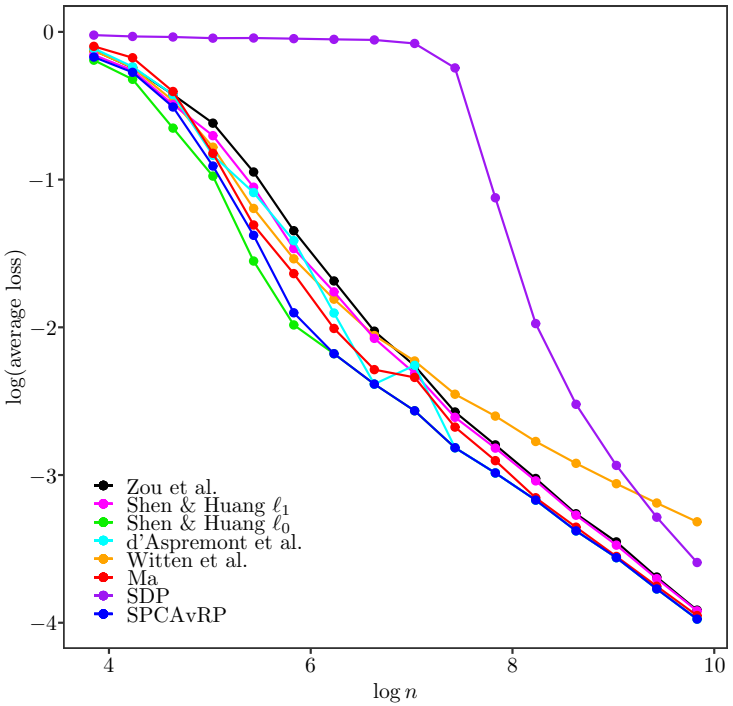
- Computationally expensive $O(p^5 \vee np^3)$.
- Poor finite sample performance.

Other options are available, but most are iterative in nature and depend on initialisers.

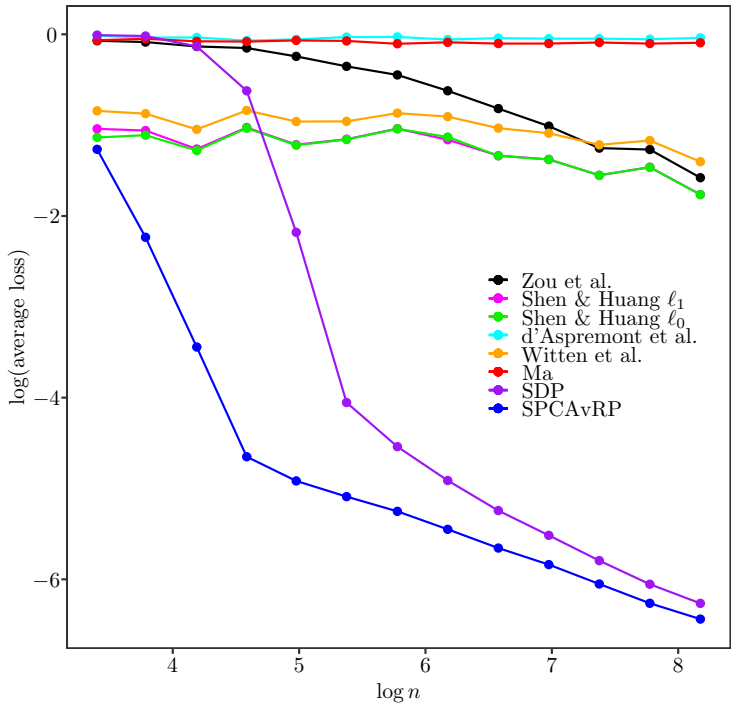
# Comparison of different sparse PCA methods

Define, for $J_q := \mathbf{1}_q \mathbf{1}_q^\top / q \in \mathbb{R}^{q \times q}$,

$$\Sigma_{(1)} = \begin{pmatrix} 2J_{10} & & \\ & J_{10} & \\ & & \mathbf{0} \end{pmatrix} + I_{100}, \quad \Sigma_{(2)} = \begin{pmatrix} 10J_{10} & & \\ & 9.9J_{30} & \\ & & I_{160} \end{pmatrix} + 0.01 I_{200}.$$




$\Sigma_{(1)}$




$\Sigma_{(2)}$

# Sparse PCA via random projection

- Given a sample covariance matrix $\hat{\Sigma}$
- Randomly select $S \subseteq [p]$ coordinates and consider the leading eigenvector/eigenvalue of $\hat{\Sigma}_{S,S}$ (which we call an axis-aligned random projection)
- Repeat the above for $A \times B$ random projections
- Choose $A$ best projections according to the leading eigenvalue of the corresponding submatrices of $\hat{\Sigma}$.
- Aggregate the eigenvectors of these $A$ best projections to identify signal coordinates.
- Use the estimate signal coordinates to estimate the sparse PC.

# Pseudocode of SPCAvRP

Given $S \subseteq [p]$, let $P_S \in \mathbb{R}^{p\times p}$ be diagonal with $j$th diagonal entry $\mathbb{1}_{\{j\in S\}}$.

**Input**: $x_1, \ldots, x_n \in \mathbb{R}^p$, $A, B \in \mathbb{N}$, $d, \ell \in [p]$.
Generate $\{P_{a,b} : a \in [A], b \in [B]\}$ independently and uniformly from $\mathcal{P}_d$.
Compute $\{P_{a,b}\hat{\Sigma}P_{a,b} : a \in [A], b \in [B]\}$, where $\hat{\Sigma} := n^{-1}\sum_{i=1}^n x_i x_i^\top$.
For $a = 1, \ldots, A$
  For $b = 1, \ldots, B$
    Compute $\hat{\lambda}_{a,b} := \lambda_1(P_{a,b}\hat{\Sigma}P_{a,b})$ and $\hat{v}_{a,b} \in v_1(P_{a,b}\hat{\Sigma}P_{a,b})$.
  Compute $b^*(a) := \arg\max_{b\in[B]} \hat{\lambda}_{a,b}$.
Compute $\hat{w} = (\hat{w}^{(1)}, \ldots, \hat{w}^{(p)})^\top$, where

$$\hat{w}^{(j)} := \frac{1}{A}\sum_{a=1}^{A} \left|\hat{v}^{(j)}_{a,b^*(a)}\right|,$$

and let $\hat{S}_1 \subseteq [p]$ be the index set of the $\ell$ largest components of $\hat{w}$.
**Output**: $\hat{v}_1 := \arg\max_{v\in\mathcal{S}^{p-1}} v^\top P_{\hat{S}_1}\hat{\Sigma}P_{\hat{S}_1}v$.

# Theoretical guarantees

Computational complexity: $O(\min\{np^2 + ABd^3 + \ell^3, ABnd^2 + p + \ell^3\})$.

Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} N_p(0, I_p + \theta_1 v_1 v_1^\top)$, where $v_1 = k^{-1/2}(\mathbf{1}_k^\top, \mathbf{0}_{p-k}^\top)^\top$. Let $\hat{v}_1$ be from SPCAvRP using $X_1, \ldots, X_n$, $A$, $B$, $d$ and $\ell$. Assume $p \geq \max(4, 2k)$, $n \geq 4\max(d, \ell)\log p$, and that there exists $t \in \{1, \ldots, k\}$ such that

$$\{1 - F_{\text{HG}}(t-1; d, k, p)\}B \geq 3\log p \tag{1}$$

and

$$2400\sqrt{\frac{k^2 d \log p}{t^2 n\theta_1^2}} \leq \min\{1, (p-k)d^{-1/2}k^{-1}\}. \tag{2}$$

Then with probability at least $1 - p^{-3} - pe^{-A/(32k^2)}$ we have for $\theta_1 \leq 1$ that

$$L(\hat{v}_1, v_1) \leq 240\sqrt{\frac{\ell \log p}{n\theta_1^2}}\max\left(1, \frac{k}{\ell}\right) + \sqrt{\max\left(1 - \frac{\ell}{k}, 0\right)}.$$

# Discussion

When $k \le \ell \lesssim k$, the loss is bounded w.h.p. by a constant multiple of $\sqrt{k \log p/(n\theta_1^2)}$, which is the minimax optimal rate. When $\ell < k$, we incur an additional loss of order $\sqrt{1 - \ell/k}$.

As $t$ increases, (1) is strengthened and (2) is weakened. When $t = 1$,

$$F_{\mathrm{HG}}(0; d, k, p) = \frac{\binom{p-k}{d}}{\binom{p}{d}} \le 1 - k/p,$$

so $B \ge 3k^{-1}p \log p$ suffices for (1). But from (2), for the minimax rate when $\theta_1 \le 1$, we need $n \gtrsim k^2 d\theta_1^{-2} \log p$, the high effective sample size regime.

When $t \asymp k$, we only need $n \gtrsim d\theta_1^{-2} \log p$, so include medium and high effective sample size regimes, but then need $B$ to be exponentially large.

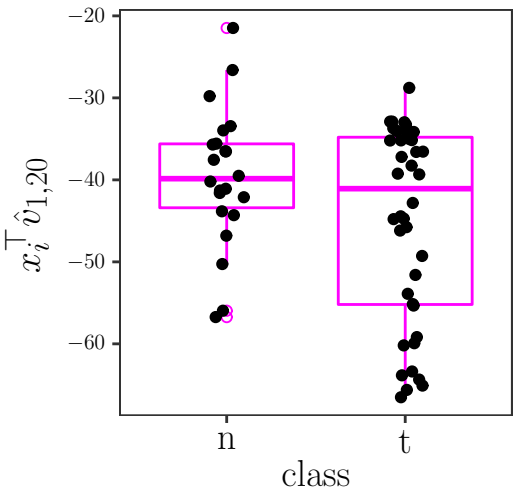
# Example with microarray data

- Colon data set: $p = 2000$, $n = 62$ (42 tumor and 20 healthy)
- Below, we project onto the first PC (in SPCAvRP, we choose $\ell = 20$)

PCA

$W_{2000} = 5.68$

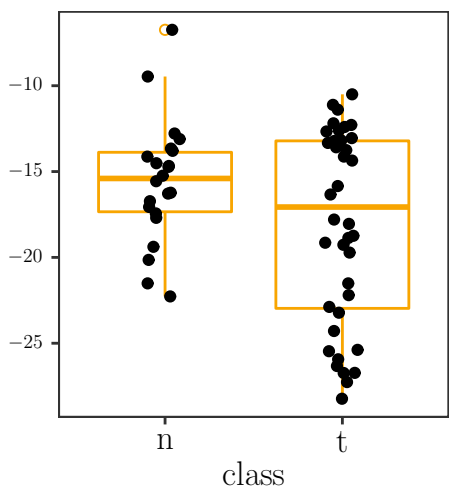
$p_t$ = 0.0416




Zou et al.

$W_{20} = 2.92$

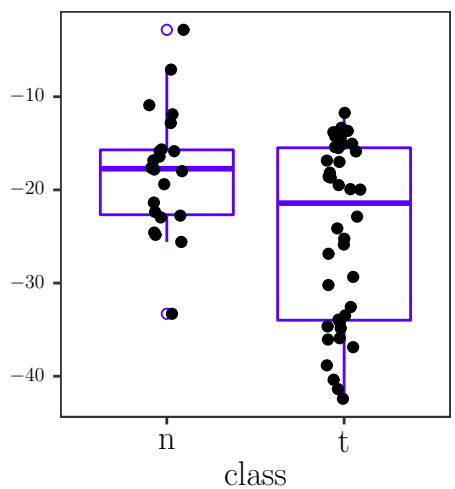
$p_t$ = 0.0435




SPCAvRP

$W_{20} = 6.54$

$p_t$ = 0.0028



# Summary

- We formulated computational lower bounds for sparse PCA by linking Sparse PCA with the Planted Clique problem.
- Rate obtained by SDP methods cannot be improved, but SDP works poorly in practice.
- Random projections offer a very general methodology for handling high-dimensional data.
- They are particularly effective in Sparse PCA because we can identify good projections and aggregate.

**Main references:**


- Wang, T., Berthet, Q. and Samworth, R. J. (2016) Statistical and computational trade-offs in estimation of sparse principal components. *Ann. Statist.*, **44**, 1896–1930.
- Gataric, M., Wang, T. and Samworth, R. J. (2018) Sparse principal component analysis via random projections. *J. Roy. Statist. Soc., Ser. B*, **82**, 329–359.
- **R** package SPCAvRP on CRAN.

# References


- 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 on arxiv.
- Chun, H. and Sündüz, K. (2009) Expression quantitative trait loci mapping with multivariate sparse partial least squares regression. *Genetics*, **182**, 79–90.
- d'Aspremont, A. El Ghaoui, L., Jordan, M. I. and Lanckriet, G. R. G. (2007) A direct formulation for sparse PCA using semidefinite programming. *SIAM Rev.*, **49**, 434–448.
- Diaconis, P. and Freedman D. (1980) Finite exchangeable sequences. *Ann. Probab.*, **8**, 745–764.
- 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., Perkins, W. and Vempala, S. (2015) On the Complexity of Random Satisfiability Problems with Planted Solutions. *Proceedings of the forty-seventh annual ACM Symposium on Theory of Computing*, to appear.

# References


- Gao, C., Ma, Z. and Zhou, H. H. (2014) Sparse CCA: Adaptive estimation and computational barriers. Available on arxiv.
- Hajek, B., Wu, Y. and Xu, J. (2014) Computational lower bounds for community detection on random graphs. Available on arxiv.
- 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.
- Jolliffe, I. T., Trendafilov, N. T. and Uddin, M. (2003) A modified principal component technique based on the LASSO, *J. Comput. Graph. Statist.*, **12**, 531–547.
- Ma, Z. (2013) Sparse principal component analysis and iterative thresholding. *Ann. Statist.*, **41**, 772–801.
- Ma, Z. and Wu, Y. (2015) Computational barriers in minimax submatrix detection. *Ann. Statist.*, to appear.
- Majumdar, A. (2009) Image compression by sparse PCA coding in curvelet domain. *Signal, image and video processing*, **3**, 27–34.

# References


- Naikal, N., Yang A. Y. and Sastry S. S. (2011) Informative feature selection for object recognition via sparse PCA. *Computer Vision (ICCV), 2011 IEEE International Conference*, 818–825.
- 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.
- Nesterov, Yu. (2005) Smooth minimization of nonsmooth functions. *Math. Program., Ser. A*, **103**, 127–152.
- Tan, K. M., Petersen, A. and Witten, D. (2014) Classification of RNA-seq Data. In S. Datta and D. Nettleton (Eds.) *Statistical Analysis of Next Generation Sequencing Data*, 219–246. Springer, New York.
- Vu, V. Q. and Lei, J. (2013) Minimax sparse principal subspace estimation in high dimensions. *Ann. Statist.*, **41**, 2905–2947.
- Wang, D., Lu, H.-C. and Yang, M.-H. (2013) Online object tracking with sparse prototypes. *IEEE transactions on image processing*, **22**, 314–325.
- Wang, T., Berthet, Q. and Samworth, R. J. (2015) Statistical and computational trade-offs in estimation of sparse principal components. *Ann. Statist.*, to appear.
- Zou, H., Hastie, T. and Tibshirani, R. (2006) Sparse principal components analysis. *J. Comput. Graph. Statist.*, **15**, 265–86.