MATH-F-305 – Projet de recherche

Statistical and Computational Trade-offs in Estimation of Sparse Principal Components

Robin Petit

May 2019

Table of Contents

- Given $X_1, \ldots, X_n$ i.i.d. samples from p-dimensional distribution P, Principal Component Analysis (PCA) projects the samples onto the space spanned by *principal components*, i.e. the set of orthogonal vectors maximising variance.

- Given $X_1, \ldots, X_n$ i.i.d. samples from p-dimensional distribution P, Principal Component Analysis (PCA) projects the samples onto the space spanned by *principal components*, i.e. the set of orthogonal vectors maximising variance.
- The first principal component is the leading eigenvector of the sample covariance matrix $\hat{\Sigma}$.

- Given $X_1, \dots, X_n$ i.i.d. samples from p-dimensional distribution P, Principal Component Analysis (PCA) projects the samples onto the space spanned by *principal components*, i.e. the set of orthogonal vectors maximising variance.
- The first principal component is the leading eigenvector of the sample covariance matrix $\hat{\Sigma}$.
- Very efficient if $p \ll n$, but breaks down if $p_n \sim n$ or $\lim_{n \to +\infty} p_n/n = +\infty$ and estimators become inconsistent (Johnstone and Lu, 2009; Paul, 2007).

- Given $X_1, \dots, X_n$ i.i.d. samples from p-dimensional distribution P, Principal Component Analysis (PCA) projects the samples onto the space spanned by *principal components*, i.e. the set of orthogonal vectors maximising variance.
- The first principal component is the leading eigenvector of the sample covariance matrix $\hat{\Sigma}$.
- Very efficient if $p \ll n$, but breaks down if $p_n \sim n$ or $\lim_{n \to +\infty} p_n/n = +\infty$ and estimators become inconsistent (Johnstone and Lu, 2009; Paul, 2007).
- Sparse PCA intends to improve interpretability of projection and to remedy this inconsistency. In the simplest case, it is assumed that the leading eigenvector v_1 of $\hat{\Sigma}$ belongs to $B_0(k) \coloneqq \{u \in \mathbb{R}^p : \|u\|_0 \leq k, \|u\|_2 = 1\}$.

(Wang et al., 2016) detailed a *trade-off* between statistical and computational efficiency:

- In general, well performing estimators are hard to compute, e.g. $v_{\max}^k(\hat\Sigma) := \mathrm{argmax}_{u \in B_0(k)}\, u^\top \hat\Sigma u$ attains minimax rate but is NP-hard (Berthet and Rigollet, 2013a,b; Birnbaum et al., 2013; Cai et al., 2013).
- Under some distributional assumptions, interesting rates can be achieved while being *easily* computable.

Some definitions (Sipser, 2012):

- P is the class of problems solvable in polynomial time.

Some definitions (Sipser, 2012):

- P is the class of problems solvable in polynomial time.
- NP is the class of problems verifiable in polynomial time.

Some definitions (Sipser, 2012):

- P is the class of problems solvable in polynomial time.
- NP is the class of problems verifiable in polynomial time.
- Conjectured that P $\subsetneq$ NP.

Some definitions (Sipser, 2012):

- P is the class of problems solvable in polynomial time.
- NP is the class of problems verifiable in polynomial time.
- Conjectured that P $\subsetneq$ NP.
- A problem Q is said NP-hard if it is at least as hard as every problem in NP (every problem $H \in$ NP is reducible to Q in polytime).

Some definitions (Sipser, 2012):

- P is the class of problems solvable in polynomial time.
- NP is the class of problems verifiable in polynomial time.
- Conjectured that P $\subsetneq$ NP.
- A problem Q is said NP-hard if it is at least as hard as every problem in NP (every problem $H \in$ NP is reducible to Q in polytime).
- A problem Q is said NP-complete if it is NP-hard and $Q \in$ NP.

- A clique in an undirected graph is a complete subgraph (every two nodes are connected). A k-clique is a clique of size k.
- The *clique problem* (denoted CLIQUE) consists in determining whether a graph contains a clique of specified size k.
- CLIQUE $\in$ NP.
- Finding the largest clique of a graph is NP-hard.

Planted Clique Problem

The *planted clique problem* is a variant of CLIQUE. Consider the following random process:

- Sample a random graph $G \sim \mathcal{G}(n, 1/2)$ (Erdős-Rényi),
- with probability $1/2$, sample uniformly $W \in \binom{V(G)}{k}$ and join each pair of vertices of W (W induces a clique) in G.

The planted clique problem consists in determining whether such a graph contains a clique of size $\geq k$.

Objectives of the paper

1. Restrict analysis to finding first principal component (i.e. maximising directional variance).
2. Find appropriate classes of probability distributions with interesting minimax rate ($\mathcal{P}_p(n, k, \theta)$).
3. Find estimators behaving well w.r.t. this rate ($\hat{v}^{\mathrm{SDP}}$).
4. Find a lower bound for estimators computable in polytime.

Table of Contents

Definition 1

The k-sparse unit ball is defined by:

$$B_0(k) := \left\{ u \in \mathbb{R}^p \text{ s.t. } \|u\|_0 \leq k, \|u\|_2 = 1 \right\}.$$

Definition 1

The k-sparse unit ball is defined by:

$$B_0(k) := \left\{u \in \mathbb{R}^p \text{ s.t. } \|u\|_0 \leq k, \|u\|_2 = 1\right\}.$$

Definition 2 (Restricted Covariance Concentration)

A distribution P is said to satisfy a *Restricted Covariance Concentration* condition with parameters p, n, ℓ, C if for all $\delta > 0$:

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

which is denoted $P \in \mathrm{RCC}_p(n, \ell, C)$

Definition 3

For $\theta > 0$ (signal-to-noise measure), define:

$$\mathcal{P}_p(n,k,\theta) \coloneqq \Big\{P \in \mathrm{RCC}_p(n,2,1) \cap \mathrm{RCC}_p(n,2k,1) \text{ s.t. } v_1(P) \in B_0(k), \lambda_1(P) - \lambda_2(P) \geq \theta\Big\}.$$

Definition 3

For $\theta > 0$ (signal-to-noise measure), define:

$$\mathcal{P}_p(n,k,\theta) \coloneqq \Big\{P \in \mathrm{RCC}_p(n,2,1) \cap \mathrm{RCC}_p(n,2k,1) \text{ s.t.} \\ v_1(P) \in B_0(k), \lambda_1(P) - \lambda_2(P) \geq \theta\Big\}.$$

Definition 4

Consider the loss function:

$$L(u,v) \coloneqq \left(1 - (u^\top v)^2\right)^{\frac{1}{2}} = \frac{1}{\sqrt{2}}\left\|uu^\top - vv^\top\right\|_2.$$

Definition 5

Consider the following notations:

- $\mathcal{M}$ is the set of non-negative definite real symmetric matrices;
- $\mathcal{M}_1 \coloneqq \{M \in \mathcal{M} \text{ s.t. } \operatorname{Tr} M = 1\}$;
- $\mathcal{M}_{1,1}(k^2) \coloneqq \{M \in \mathcal{M}_1 \text{ s.t. } \operatorname{rank} M = 1, \|M\|_0 = k^2\}$.

Theorem 1

For $2k \log p \leq n$, $\hat{v}^k_{\max}(\hat{\Sigma}) \coloneqq \operatorname{argmax}_{u \in B_0(k)} u^\top \hat{\Sigma} u$ satisfies:

$$\begin{aligned}
\sup_{P \in \mathcal{P}_p(n,k,\theta)} \mathbb{E}_P L(\hat{v}^k_{\max}(\hat{\Sigma}), v_1(P)) &\leq 2\sqrt{2}\left(1 + \frac{1}{\log p}\right)\sqrt{\frac{k \log p}{n\theta^2}} \\
&\leq 7\sqrt{\frac{k \log p}{n\theta^2}}
\end{aligned}$$

Theorem 1

For $2k \log p \leq n$, $\hat{v}^k_{\max}(\hat{\Sigma}) \coloneqq \operatorname{argmax}_{u \in B_0(k)} u^\top \hat{\Sigma} u$ satisfies:

$$\begin{aligned}\sup_{P \in \mathcal{P}_p(n,k,\theta)} \mathbb{E}_P L(\hat{v}^k_{\max}(\hat{\Sigma}), v_1(P)) &\leq 2\sqrt{2}\left(1 + \frac{1}{\log p}\right)\sqrt{\frac{k \log p}{n\theta^2}} \\ &\leq 7\sqrt{\frac{k \log p}{n\theta^2}}\end{aligned}$$

Theorem 2

If $7 \leq k \leq \sqrt{p}$ and $0 < \theta \leq \frac{1}{16(1+\frac{9}{\log p})}$:

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

Lemma 3 (SM – Propostion 1)

Let $P \in \mathrm{RCC}_p(n, \ell, C)$ with $\ell \log p \leq n$. Then:

$$\mathbb{E}_P \sup_{u \in B_0(\ell)} \left|\hat{V}(u) - V(u)\right| \leq \left(1 + \frac{1}{\log p}\right) C \sqrt{\frac{\ell \log p}{n}}.$$

Lemma 3 (SM – Propostion 1)

Let $P \in \mathrm{RCC}_p(n, \ell, C)$ with $\ell \log p \leq n$. Then:

$$\mathbb{E}_P \sup_{u \in B_0(\ell)} \left|\hat{V}(u) - V(u)\right| \leq \left(1 + \frac{1}{\log p}\right) C \sqrt{\frac{\ell \log p}{n}}.$$

Lemma 4 (Curvature Lemma (Vu et al., 2013))

For $A \in \mathbb{R}^{p \times p}$ a symmetric matrix and E the projection onto the subspace spanned by the eigenvectors of A corresponding to the d largest eigenvalues, if $\delta_A := \lambda_d - \lambda_{d+1} \gneqq 0$, then for all F satisfying $0 \preceq F \preceq I$ and $\operatorname{Tr} F = d$:

$$\frac{\delta_A}{2} \|E - F\|_2^2 \leq \operatorname{Tr}\left(A(E - F)\right)$$

Proof (Theorem 1.) Fix $P \in \mathcal{P}_p(n,k,\theta)$. By Curvature lemma:

$$\begin{aligned}\frac{\theta}{2}\left\|vv^\top - \hat{v}\hat{v}^\top\right\|_2^{\color{red}2} &\leq \mathrm{Tr}(\Sigma(vv^\top - \hat{v}\hat{v}^\top)) \\ &\leq \mathrm{Tr}((\Sigma - \hat{\Sigma})(vv^\top - \hat{v}\hat{v}^\top)).\end{aligned}$$

$M = \frac{vv^\top - \hat{v}\hat{v}^\top}{\left\|vv^\top - \hat{v}\hat{v}^\top\right\|_2}$ has rank 2, trace 0 and non-zero entries in at most $2k$ rows and $2k$ columns. Hence $M = (xx^\top - yy^\top)/\sqrt{2}$ for some $x, y \in B_0(2k)$. Thus:

$$\begin{aligned}\mathbb{E}L(\hat{v}, v) = \frac{1}{\sqrt{2}}\mathbb{E}\left\|\hat{v}\hat{v}^\top - vv^\top\right\|_2 &\leq \frac{1}{\theta}\mathbb{E}[\mathrm{Tr}((\Sigma - \hat{\Sigma})(xx^\top - yy^\top))] \\ &\leq \frac{2}{\theta}\mathbb{E}\sup_{u \in B_0(2k)}\left|\hat{V}(u) - V(u)\right| \leq 2\sqrt{2}\left(1 + \frac{1}{\log p}\right)\sqrt{\frac{k\log p}{n\theta^2}}\end{aligned}$$

$\square$

SDP formulation from (d'Aspremont et al., 2005):

$$\begin{aligned} \max_{M \in \mathcal{M}_1} \ & \operatorname{Tr}(\hat{\Sigma} M) \\ \text{s.t. } & \|M\|_0 \leq k^2 \\ & \operatorname{rank} M = 1 \end{aligned}$$

With fixed sparsity level, formulation is equivalent to:

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

- Problem: rank and sparsity constraints are not convex.
- Solution: Drop rank constraint and relax sparsity constraint by ℓ_1 penalty.

$\hat{v}^{SDP}$ is then defined by the following *convex* optimisation problem (with parameter $\lambda > 0$):

$$\max_{M \in \mathcal{M}_1} \operatorname{Tr}(\hat{\Sigma} M) - \lambda \|M\|_1$$

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

begin

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

Step 2: For $f(M) \coloneqq \operatorname{Tr}(\hat{\Sigma} M) - \lambda \|M\|_1$, let $\hat{M}^\epsilon$ be an ϵ-maximiser of f in $\mathcal{M}_1$.

Step 3: Let $\hat{v}^{\mathrm{SDP}} \coloneqq \hat{v}^{\mathrm{SDP}}_{\lambda,\epsilon} \in \operatorname{argmax}_{u \text{ s.t. } \|u\|_1 = 1} u^\top \hat{M}^\epsilon u$.

end

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

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

Fully adaptative: $\hat{v}^{\mathrm{SDP}}$ is not necessarily sparse but can be forced sparse by keeping the k top components and set the $p-k$ remaining to 0 (then renormalising).

- Step 1 takes $O(np^2)$ flops
- Step 2 can be solved in $O(\frac{\lambda^2 p^2+1}{\epsilon})$ flops (provided algorithm based on the following equality: $\max_{M\in\mathcal{M}_1} \mathrm{Tr}(\hat{\Sigma}M) = \max_{M\in\mathcal{M}_1} \min_{U\in\mathcal{U}} \mathrm{Tr}((\hat{\Sigma}+U)M)$ where $\mathcal{U} = \{U \in \mathbb{R}^{p\times p}$ s.t. $U = U^\top, \|U\|_\infty \leq \lambda\}$).
- Step 3 requires $O(p^3)$ flops in worst case but under additional assumptions is feasible in $O(p^2)$.

Hence $\hat{v}^{\mathrm{SDP}}$ is computable in polytime. Is it *statistically efficient* though?

Theorem 5

Let $\Sigma \in \mathcal{M}$ s.t. $\theta = \lambda_1(\Sigma) - \lambda_2(\Sigma) > 0$. Let $\mathbf{X} \in \mathbb{R}^{n \times p}$. For arbitrary $\lambda > 0$ and $\epsilon > 0$, if $\left\|\Sigma - \hat{\Sigma}\right\|_\infty \leq \lambda$, then $\hat{v}^{\mathrm{SDP}}$ computed by Algorithm 1 with parameters $\mathbf{X}, \lambda, \epsilon$ satisifes:

$$L(\hat{v}^{\mathrm{SDP}}, v_1(\Sigma)) \leq \frac{4\sqrt{2}\lambda k}{\theta} + 2\sqrt{\frac{\epsilon}{\theta}}.$$

Theorem 6

For any $P \in \mathcal{P}_p(n, k, \theta)$ and $X_1, \dots, X_n \overset{\text{iid}}{\sim} P$, let $\hat{v}^{\mathrm{SDP}}(\mathbf{X})$ be the output of Algorithm 1 with parameters $\mathbf{X}$, $\lambda = 4\sqrt{\frac{\log p}{n}}$ and $\epsilon = \frac{\log p}{4n}$. If $4 \log p \leq n \leq k^2 p^2 \theta^{-2} \log p$ and $\theta \in (0, k]$, then:

$$\sup_{P \in \mathcal{P}_p(n,k,\theta)} \mathbb{E}_P L(\hat{v}^{\mathrm{SDP}}(\mathbf{X}), v_1(P)) \leq \min\left\{1, (16\sqrt{2} + 2)\sqrt{\frac{k^2 \log p}{n\theta^2}}\right\}$$

Lemma 7

Suppose $P \in \mathcal{P}_p(n, k, \theta)$, $X_1, \dots, X_n \overset{\text{iid}}{\sim} P$. Then:

$$\left\| \hat{\Sigma} - \Sigma \right\|_\infty \leq 2 \sup_{u \in B_0(2)} \left| \hat{V}(u) - V(u) \right|.$$

Proof (Lemma 7.) $e_{s,r}^+ := (e_s + e_r)/2$, $e_{s,r}^- := (e_s - e_r)/2$.

$$\begin{aligned}
\left\| \hat{\Sigma} - \Sigma \right\|_\infty &\leq \max_{r,s \in [1:p]} \left| \frac{1}{n} \sum_{i=1}^n \left((e_{s,r}^+)^\top X_i \right)^2 - \mathbb{E}\left[\left((e_{s,r}^+)^\top X_1 \right)^2 \right] \right| \\
&+ \max_{r,s \in [1:p]} \left| \frac{1}{n} \sum_{i=1}^n \left((e_{s,r}^-)^\top X_i \right)^2 - \mathbb{E}\left[\left((e_{s,r}^-)^\top X_1 \right)^2 \right] \right| \\
&\leq 2 \sup_{u \in B_0(2)} \left| \hat{V}(u) - V(u) \right|.
\end{aligned}$$

□

Proof (Theorem 6.) Fix $P \in \mathcal{P}_p(n, k, \theta)$.

$$\mathbb{E}_P L(\hat{v}^{\mathrm{SDP}}, v_1(P)) \le \frac{4\sqrt{2}\lambda k}{\theta} + 2\sqrt{\frac{\epsilon}{\theta}} + \mathbb{P}\left[\sup_{u \in B_0(2)} \left|\hat{V}(u) - V(u)\right| > 2\sqrt{\frac{\log p}{n}}\right].$$

Since $P \in \mathrm{RCC}_p(n, 2, 1)$, for every $\delta > 0$:

$$\mathbb{P}\left[\sup_{u \in B_0(2)} \left|\hat{V}(u) - V(u)\right| > \max\left\{\sqrt{\frac{2\log(p/\delta)}{n}}, \frac{2\log(p/\delta)}{n}\right\}\right] \le \delta.$$

Set $\delta \coloneqq \sqrt{(k^2 \log p)/(n\theta^2)}$. Since $4 \log p \le n$:

$$2\log(p/\delta) \le 1/2 + \log n - \log\log 2 \le n.$$

Furthermore, since $n \le k^2 p^2 \theta^{-2} \log p$:

$$2\log(p/\delta) = 2\log p + \log\left((n\theta^2)/(k^2 \log p)\right) \le 4\log p.$$

Finally:

$$\mathbb{P}\left[\sup_{u \in B_0(2)} \left|\hat{V}(u) - V(u)\right| > 2\sqrt{\frac{\log p}{n}}\right] \le \sqrt{\frac{k^2 \log p}{n\theta^2}}.$$

□

Polytime computable

It is conjectured that the Planted Clique problem is *hard* in the following sense (for $\tau = 0$):

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

However, a quasi-poly-time algorithm is known to solve PCP w.h.p. if $\kappa \geq 2 \log m$ and a poly-time algorithm is known to solve PCP w.h.p. if $\kappa \geq \sqrt{m}$.

Theorem 8

Fix $\tau \in [0, 1/6)$, assume $(A1)(\tau)$ and let $\alpha \in (0, \frac{1-6\tau}{1-2\tau})$. Let (p, k, θ) be indexed by n such that: (i) $k = O(p^{1/2-\tau-\delta})$ for some $\delta \in (0, 1/2 - \tau)$, (ii) $n = o(p \log p)$ and (iii) $\theta \leq k^2/(1000p)$. Also suppose that:

$$\frac{k^{1+\alpha} \log p}{n\theta^2} \xrightarrow[n \to +\infty]{} 0.$$

Let $\mathbf{X} \in \mathbb{R}^{n \times p}$ with rows iid P. Then every sequence of randomised polynomial time estimators $(\hat{v}^{(n)})_n$ of $v_1(P)$ satisfies:

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

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

begin

Step 1: Let $n \leftarrow \lfloor 9m/(10L) \rfloor$, $p = p_n$, $k \leftarrow \lfloor \kappa/L \rfloor$. Draw $u_1, \dots, u_n, w_1, \dots, w_p$ uniformly from $V(G)$. Form $\mathbf{A} = (\mathbb{1}_{\{u_i \sim w_j\}})_{ij}$ and $\mathbf{X} = \operatorname{diag}(\xi_1, \dots, \xi_n)(2\mathbf{A} - \mathbf{1})$.

Step 2: Use $\hat{v}^{(n)}$ to compute $\hat{v} \coloneqq \hat{v}^{(n)}$.

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

Step 4: For $u \in V(G)$ and $W \subset V(G)$, let $\operatorname{nb}(u, W) \coloneqq \mathbb{1}_{\{u \in W\}} + \sum_{w \in W} \mathbb{1}_{\{u \sim w\}}$. Set $\hat{K} \coloneqq \{u \in V(G) \text{ s.t. } \operatorname{nb}(u, \{w_j\}_{j \in \hat{S}}) \geq 3k/4\}$.

end

Output: $\hat{K}$

Algorithm 2: Pseudo-code for a Planted Algorithm algorithm.

Berthet, Q. and Rigollet, P. (2013a). Complexity theoretic lower bounds for sparse principal component detection. In *Conference on Learning Theory*, pages 1046–1066.

Berthet, Q. and Rigollet, P. (2013b). Computational lower bounds for sparse pca. *arXiv preprint arXiv:1304.0828*.

Birnbaum, A., Johnstone, I. M., Nadler, B., and Paul, D. (2013). Minimax bounds for sparse pca with noisy high-dimensional data. *Annals of statistics*, 41(3):1055.

Cai, T. T., Ma, Z., Wu, Y., et al. (2013). Sparse pca: Optimal rates and adaptive estimation. *The Annals of Statistics*, 41(6):3074–3110.

d'Aspremont, A., Ghaoui, L. E., Jordan, M. I., and Lanckriet, G. R. (2005). A direct formulation for sparse pca using semidefinite programming. In *Advances in neural information processing systems*, pages 41–48.

Johnstone, I. M. and Lu, A. Y. (2009). On consistency and sparsity for principal components analysis in high dimensions. *Journal of the American Statistical Association*, 104(486):682–693.

Paul, D. (2007). Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. *Statistica Sinica*, pages 1617–1642.

References II

Sipser, M. (2012). *Introduction to the Theory of Computation*. Cengage Learning.

Vu, V. Q., Cho, J., Lei, J., and Rohe, K. (2013). Fantope projection and selection: A near-optimal convex relaxation of sparse pca. In *Advances in neural information processing systems*, pages 2670–2678.

Wang, T., Berthet, Q., and Samworth, R. J. (2016). Statistical and computational trade-offs in estimation of sparse principal components. *The Annals of Statistics*, 44(5):1896–1930.