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

Chia-Man Hung
Master Data Science
chia-man.hung@polytechnique.edu

Zhengying Liu
Master Data Science
zhengying.liu@polytechnique.edu

Abstract

In the paper "Statistical and computational trade-offs in estimation of sparse principal components" by Wang, Tengyao, Quentin Berthet, and Richard J. Samworth, the authors managed to prove that under some widely-believed assumptions from computational complexity theory, there is a fundamental trade-off between statistical and computational performance in the problem of finding good Sparse PCA estimators. In order to prove this, they introduced a class of general and robust conditions called Restricted Covariance Concentration (RCC) condition on probability distributions. For models satisfying these conditions, the authors gave a Sparse PCA estimator that is computable in randomised polynomial time. This estimator has good theoretical performance which just has a small factor of difference compared to the minimax rate of convergence. Finally, they showed that this small factor is fundamental and this trade-off is inevitable if the Planted Clique problem is hard (e.g. NP-hard).

1 Introduction

In this article, we give some review notes for the paper "Statistical and computational trade-offs in estimation of sparse principal components" by Wang, Tengyao, Quentin Berthet, and Richard J. Samworth [7]. The paper is published in 2016. It introduced some recent research on Sparse Principle Component Analysis (Sparse PCA), which is a popular approach to remedy the inconsistency of ordinary PCA estimator in high-dimensional settings. We first list the main points made in the paper to briefly introduce its general ideas. Then we give some important definitions and theorems introduced and proven by the authors, along with some of our comments. We also give some experimental results with some remarks.

2 Overview

Some main points made in the paper are listed as follows:

1. Classical PCA breaks down in some high-dimensional settings [5];
2. Sparse PCA can overcome this and gives additional interpretability;
3. Sparse PCA has gained high popularity and many different estimators are proposed;
4. Some of these estimators have good theoretical properties, e.g. attain the minimax rate of convergence [3] [6];
5. But these estimators are not computable in polynomial time;
6. One main question treated in the paper: Does there exist an estimator that is computable in (randomised) polynomial time and that attains the minimax rate of convergence?

7. Some studied this problem by considering the "Planted Clique" problem, but their approach was not suitable for sparse PCA; [1] [2]
8. One first contribution of the paper is to introduce the Restricted Covariance Concentration (RCC) condition, which is satisfied by sub-Gaussian distributions;
9. There exists an estimator $\hat{v}^{\text{SDP}}$ that is computable in polynomial time and has a worst case performance close to the minimax rate of convergence (by a factor of $\sqrt{k}$, where k is the level of sparsity);
10. The main result of the paper: Assuming Planted Clique hypothesis, there exists a fundamental trade-off between statistical and computational efficiency in the estimation of sparse principal components in an effective sample size regime;
11. Statistical and computational trade-offs have also been studied in many different domains and will be a key challenge for theoreticians in the coming years.

Now for each point, we give more details and recall some important definitions and propositions. In most cases, we will not go into the details of the proofs.

2.1 Classical Principle Component Analysis

(Point 1)

Principle Component Analysis (PCA) is one of the oldest and most widely-used dimension reduction devices in statistics. It is mathematically defined as an orthogonal linear transformation that projects the data to a new coordinate system such that the greatest variance by some projection of the data comes to lie on the first coordinate (called the first principal component), the second greatest variance on the second coordinate, and so on. It consists of the following procedures.

Let $X \in \mathbb{R}^{n\times p}$ be a data matrix with column-wise zero empirical mean, i.e. $\sum_{i=1}^{n} X_{ij} = 0$ for $j = 1, ..., p$. Then the **first principle component** of X is defined as

$$\hat{v}_1 = \underset{\|v\|=1}{\arg\max} \|Xv\|^2 = \underset{\|v\|=1}{\arg\max}\, v^\top \hat{\Sigma} v, \tag{1}$$

which is the direction with the greatest empirical variance. Here

$$\hat{\Sigma} = \frac{1}{n} X^\top X \tag{2}$$

is the empirical covariance matrix. Then by a procedure similar to Schmidt normalization, we can consider the projection of the data points onto the orthogonal complement of $\hat{v}_1$ and repeat the above process and find a $\hat{v}_2$, then $\hat{v}_3$ and so on.

If we consider the first p components, this is equivalent to do an SVD for the matrix X

$$T = XV$$

where $V = [\hat{v}_1\, \hat{v}_2 \dots \hat{v}_p] \in \mathbb{R}^{p\times p}$ is an orthogonal matrix and T has orthogonal columns, ordered decreasingly according to their norms.

In this article, we will mainly consider the first principle component.

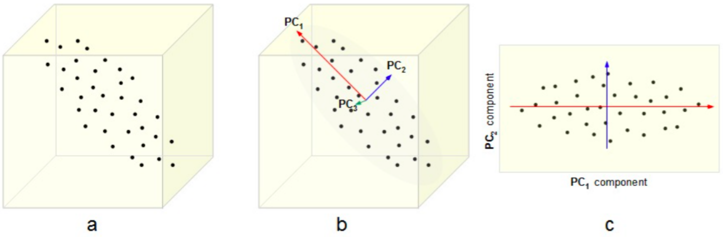


Figure 1: PCA is an orthogonal transformation.

PCA is convenient but has the following weaknesses:

1. It is not robust to outliers. A single error in measurements can strongly impact PCA;
2. It is sensitive to the scaling of the data, and it is difficult to decide which scaling is the best;
3. It may break down in some high-dimensional settings, e.g. when $p \approx n$ or $p \gg n$.

Concerning the last weakness, Paul (2007) [5] considered following situation.

Let the rows in X (i.e. $X_1, ..., X_n$) be independent $N_p(0, \Sigma)$ random vectors, with

$$\Sigma = I_p + \theta v_1 v_1^\top$$

for some $\theta > 0$ and a unit vector $v_1 \in \mathbb{R}^p$. In this case, the first principle component is v_1 and $\lambda_1(\Sigma) = 1 + \theta, \lambda_2(\Sigma) = 1$. Thus the difference between the largest and the second largest eigenvalue is $\theta = \lambda_1(\Sigma) - \lambda_2(\Sigma)$. The classical PCA estimate would be $\hat{v}_1$, the leading unit eigenvector of the empirical covariance matrix (2). However, in the high-dimensional setting where $p = p_n$ is such that $p/n \to c \in (0, 1)$, Paul showed that

$$|\hat{v}_1^\top v_1| \overset{\text{a.s.}}{\to} \begin{cases} \sqrt{\frac{1-c/\theta^2}{1+c/\theta}}, & \text{if } \theta > \sqrt{c}, \\ 0, & \text{if } \theta \le \sqrt{c}, \end{cases}$$

which means that

$\hat{v}_1$ is inconsistent as an estimator of v_1 in this asymptotic regime.

2.2 Sparse PCA

(Points 2,3)

To remedy this inconsistency, sparse PCA has been proposed. In the simplest case, v_1 is assumed to be in the k-sparse unit Euclidean sphere in $\mathbb{R}^p$, given by

$$B_0(k) = \{u \in \mathbb{R}^p : \|u\|_0 \le k, \|u\|_2 = 1\}.$$

Then the first k-sparse principle component is given by

$$\hat{v}^k_{\text{max}} \in \underset{u \in B_0(k)}{\arg\max}\, u^\top \hat{\Sigma} u. \tag{3}$$

We notice that the only difference between this definition and that in (1) is the constraint $\|u\|_0 \le k$. This constraint is quite essential. As the number of non-zero terms in $\hat{v}^k_{\text{max}}$ is bounded by k, this estimator is likely to be more robust in high-dimensional settings.

Given this definition, several questions come naturally in mind:

1. Does this estimator have good theoretical properties?
2. How do we even measure the performance of this estimator?
3. Is it easy to compute $\hat{v}^k_{\text{max}}$?

2.3 Theoretical properties of sparse PCA estimators and minimax rate of convergence

(Points 4,5,6)

To evaluate the theoretical properties of sparse PCA estimates, one common performance measure in the literature is whether the estimator attains the minimax rate of convergence.

For a class $\mathcal{P}$ of distributions and a loss function $L(v_1, \hat{v})$, we can consider the minimax rate defined by

$$\inf_{\hat{v}} \sup_{P \in \mathcal{P}} \mathbb{E}_P[L(v_1, \hat{v})]$$

where $\hat{v}$ runs over all possible estimators. In [6], Vu and Lei showed that for a certain class $\mathcal{P}_p(n, k)$ of sub-Gaussian distributions and in a particular asymptotic regime, one has

$$\inf_{\hat{v}} \sup_{P \in \mathcal{P}_p(n,k)} \mathbb{E}_P[1 - v_1^\top \hat{v}] \asymp \frac{k \log p}{n}. \tag{4}$$

Here "$\asymp$" means asymptotic equivalence.

We recall that a random vector $X \in \mathbb{R}^p$ is said to be sub-Gaussian if

$$\mathbb{E}[e^{u^\top X}] \leq e^{\sigma^2 \|u\|^2/2}$$

for any $u \in \mathbb{R}^p$ and some $\sigma \in \mathbb{R}$.

It is also shown in [6] that $\hat{v}^k_{\text{max}}$ attains the minimax rate of convergence in (4), which seems to give a satisfying answer to the question of sparse PCA estimation.

However, existing sparse PCA estimators that attain the minimax rate of convergence have one common unsettling feature: **they are not computable in polynomial time**. For example, to solve (3), one may have to test all $\binom{p}{k}$ possible choices for the non-zero terms in $\hat{v}^k_{\text{max}}$, which easily becomes infeasible when p and k increase even a little bit.

So an important question is addressed by the authors of the paper:

Is it possible to find an estimator of v_1 that is computable in (randomised) polynomial time, and that attains the minimax optimal rate of convergence when the sparsity of v_1 is allowed to vary with the sample size?

Following this question, we now introduce the main approach of the paper.

2.4 Restricted Covariance Concentration

(Point 8)

The notion of Restricted Covariance Concentration (RCC) condition is an important contribution of this paper. This condition is satisfied by all Gaussian and sub-Gaussian distributions, which are the main classes considered in the literature. RCC turns out to be very convenient when analyzing convergence rates in sparse PCA.

Let's first introduce some notations. We consider the class $\mathcal{P}$ of the distributions P on $\mathbb{R}^p$ such that $\mathbb{E}_P(X) = 0$ and that the covariance matrix $\Sigma(P)$ of P is finite. Let $\lambda_1(P), \lambda_2(P), ..., \lambda_p(P)$ be the eigenvalues of $\Sigma(P)$ in decreasing order. When $\lambda_1(P) > \lambda_2(P)$, the first principle component $v_1(P)$ is well-defined up to sign. At last, the data matrix $X \in \mathbb{R}^{n \times p}$ and any estimator of v_1 are defined in a conventional way.

As for the performance measure of the estimators, the authors adopted the following loss function

$$L(u, v) := \sqrt{1 - (u^\top v)^2}.$$

To define the RCC condition, we need to introduce the directional variance of P along a unit vector $u \in \mathbb{R}^p$ given by $V(u) := \mathbb{E}_P(u^\top X_1)^2 = u^\top \Sigma u$ and its empirical counterpart $\hat{V}(u) := n^{-1} \sum_{i=1}^n (u^\top X_i)^2 = u^\top \hat{\Sigma} u$.

Restricted Covariance Concentration condition. For $l \in \{1, ..., p\}$ and $C > 0$ we say $P \in \text{RCC}_p(n, l, C)$ if

$$\mathbb{P}\left\{ \sup_{u \in B_0(l)} |\hat{V}(u) - V(u)| \geq C \max\left(\sqrt{\frac{l \log(p/\delta)}{n}}, \frac{l \log(p/\delta)}{n} \right) \right\} \leq \delta \tag{5}$$

for any $\delta > 0$.

Then Proposition 1 of the article shows that all Gaussian and sub-Gaussian distributions satisfy this condition with some specific parameters.

After introducing the RCC condition, the authors consider the following classes of distributions:

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

The above classes can be considered as a generalization of the distribution classes (e.g. sub-Gaussian distributions) considered in other articles.

Then just as what Vu and Lei showed in [6] for sub-Gaussian distributions, the authors showed in Theorems 2 and 3 that, under some mild conditions on the parameters, $\hat{v}^k_{\text{max}}$ attains the minimax rate of convergence for distributions in $\mathcal{P}_p(n, k, \theta)$ which is asymptotically equivalent to

$$\sqrt{\frac{k \log p}{n\theta^2}}. \tag{6}$$

Notice that here the $\hat{v}^k_{\text{max}}$ is defined as

$$\hat{v}^k_{\text{max}} := \text{sargmax}_{u \in B_0(k)} u^\top \hat{\Sigma} u \tag{7}$$

where sargmax denotes the smallest element of the argmax in the lexicographic ordering. In this way $\hat{v}^k_{\text{max}}$ is well-defined and is guaranteed to be a measurable function on $\hat{\Sigma}$.

2.5 Semidefinite relaxation estimator $\hat{v}^{\text{SDP}}$

(Point 9)

As discussed before, $\hat{v}^k_{\text{max}}$ may be very difficult to compute. Thus the authors proposed some relaxation to obtain an estimator that is computable in polynomial time. They also gave Algorithms 1 and 2 for computing this estimator $\hat{v}^{\text{SDP}}$.

As the main difficulty in (7) is the constraint $u \in B_0(k)$, one can replace this constraint by applying some relaxation. Let $\mathcal{M}$ be the set of all $p \times p$ non-negative definite real symmetric matrices, $\mathcal{M}_1 := \{M \in \mathcal{M} : \text{tr}(M) = 1\}$ and $\mathcal{M}_{1,1}(k^2) := \{M \in \mathcal{M}_1 : \text{rank}(M) = 1, \|M\|_0 = k^2\}$. One observes that

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

As the two conditions in the definition of $\mathcal{M}_{1,1}(k^2)$ are not convex, we can adopt the standard semidefinite relaxation approach (i.e. dropping the rank constraint and replace the ℓ_0 norm by the ℓ_1 norm) and consider the problem

$$\max_{M \in \mathcal{M}_1} \text{tr}(\hat{\Sigma} M) - \lambda \|M\|_1. \tag{8}$$

Now we have a convex optimisation problem and we can then apply Algorithm 1 in the paper to output an estimator $\hat{v}^{\text{SDP}}$. The most important step in Algorithm 1 is Step 2: For $f(M) := \text{tr}(\hat{\Sigma} M) - \lambda \|M\|_1$, compute a $\hat{M}^\epsilon$ such that $f(\hat{M}^\epsilon) \geq \max_{M \in \mathcal{M}_1} f(M) - \epsilon$. The paper gives Algorithm 2 to implement this step. The key point of Algorithm 2 is that the optimisation problem in Step 2 can be rewritten in a saddlepoint formulation:

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

where $\mathcal{U} := \{U \in \mathbb{R}^{p \times p} : U^\top = U, \|U\|_\infty \leq \lambda\}$.

If we take $\lambda = 4\sqrt{\frac{\log p}{n}}$ and $\epsilon = \frac{\log p}{4n}$, the overall complexity of Algorithm 1 and Algorithm 2 is given by

$$O(\max(p^5, \frac{np^3}{\log p})),$$

which is indeed polynomial.

As for the theoretical property of $\hat{v}^{\text{SDP}}$, Lemma 4 and Theorem 5 show that, under some mild assumptions on the parameters and taking $\lambda = 4\sqrt{\frac{\log p}{n}}$ and $\epsilon = \frac{\log p}{4n}$ as above, the worst case risk of $\hat{v}^{\text{SDP}}$ for distributions in the class $\mathcal{P}_p(n, k, \theta)$ is bounded by

$$\min \left\{ (16\sqrt{2} + 2)\sqrt{\frac{k^2 \log p}{n\theta^2}}, 1 \right\}$$

which is

$$O\left(\sqrt{\frac{k^2 \log p}{n\theta^2}}\right).$$

We notice that this quantity only differs from (6) by a factor of $\sqrt{k}$.

However, the authors showed that in some asymptotic regime, this gap of $\sqrt{k}$ is essential and cannot be improved in the case where the estimators are computable in polynomial time, assuming that the **Planted Clique** problem is hard.

2.6 Planted Clique problem

(Points 7,10)

The authors in the paper used a polynomial time reduction from the Planted Clique problem to the sparse principle component estimation problem to prove that, if Planted Clique is hard, then sparse PCA estimation problem is hard too. They claimed that any randomised polynomial time algorithm with a faster rate of convergence could be adapted to solve instances of the planted clique problem. So in this section, we will talk about what the Planted Clique problem is.

A (undirected) graph G is defined to be the ordered pair (V, E) where V is a countable set and $E \subset \{e \subset V : |e| = 2\}$. A **clique** C is a subset of V such that $\{x, y\} \in E$ for all distinct $x, y \in C$. The problem of finding a clique of maximum size in a given graph G is known to be NP-complete [4].

An easier problem called Planted Clique only consider randomly generated input graphs with a clique "planted" in. A more formal formulation of this problem is given in the following.

Let $\mathbb{G}_m$ denote the set of all undirected graphs with m vertices. Let $\mathcal{G}_{m,\kappa}$ be a distribution on $\mathbb{G}_m$ constructed by first picking κ distinct vertices uniformly at random and connecting all edges (the "planted clique"), then joining each remaining pair of distinct vertices by an edge independently with probability 1/2.

Then the **Planted Clique problem** takes as input graphs randomly sampled from the distribution $\mathcal{G}_{m,\kappa}$ and the goal is to find an algorithm that can locate a maximum clique K_m with high probability.

When $\kappa = \kappa_m \geq c\sqrt{m}$ for some constant $c > 0$, there exist polynomial time algorithms solving this problem. But below this threshold, e.g. when $\kappa = O(m^{1/2-\delta})$ for some $\delta > 0$, many works in the literature showed the hardness of this problem and might suggest that there is no randomised polynomial time algorithm solving the Planted Clique problem in this regime. Thus the authors made the following assumption:

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

We notice that among the assumptions (A1)(τ) for different τ, the case with $\tau = 0$ (i.e. (A1)(0) is the strongest one. When $\tau > 0$, (A1)(τ) is weaker and thus the results relying on the correctness of (A1)(τ) are stronger.

From (A1)(τ), the authors established the main result of this paper: Theorem 6. Given its importance, we state this theorem in details.

Theorem 6. *$\tau \in [0, 1/6)$, assume $(A1)(\tau)$, and let $\alpha \in (0, \frac{1-6\tau}{1-2\tau})$. For any $n \in \mathbb{N}$, let $(p, k, \theta) = (p_n, k_n, \theta_n)$ be parameters indexed by n such that $k = O(p^{1/2-\tau-\delta})$ for some $\delta \in (0, 1/2 - \tau)$, $n = o(p \log p)$ and $\theta \leq k^2/(1000p)$. Suppose further that*

$$\frac{k^{1+\alpha} \log p}{n\theta^2} \to 0$$

as $n \to \infty$. Let X be an $n \times p$ matrix with independent rows, each having distribution P. Then every sequence $(\hat{v}^{(n)})$ of randomised polynomial time estimators of $v_1(P)$ satisfies

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

as $n \to \infty$.

A more colloquial way of stating this theorem would be, if the assumption (A1)(τ) is correct and the number of samples n satisfies

$$\frac{k^{1+\alpha} \log p}{\theta^2} \ll n \ll p \log p,$$

then for the distribution class $\mathcal{P}_p(n, k, \theta)$, the worst case risk (or rate of convergence) of any sequence of randomised polynomial time estimators will be

$$\gg \sqrt{\frac{k^{1+\alpha} \log p}{n\theta^2}},$$

which means that there is indeed no polynomial time estimator for v_1 attaining the minimax rate of convergence. We recall that for the class $\mathcal{P}_p(n, k, \theta)$, the minimax rate of convergence is given by

$$\sqrt{\frac{k \log p}{n\theta^2}}$$

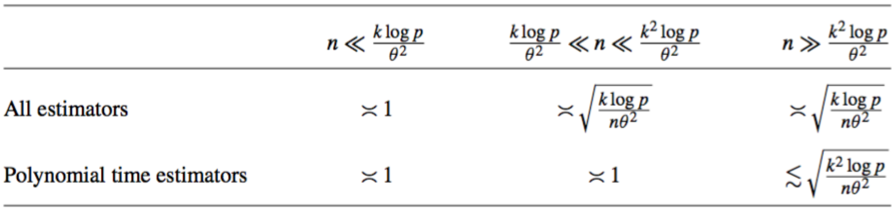
as in (6), according to Theorems 2 and 3 of the paper.

The proof of Theorem 6 mainly consists of constructing a random matrix $\mathbf{Y} \in \mathbb{R}^{n \times p}$ with rows belonging to the class $\mathcal{P}_p(n, k, \theta)$ (conditionally), from a graph with a planted clique. This is a polynomial time reduction and thus each polynomial time estimator will solve the Planted Clique problem in polynomial time, which contradicts the hypothesis (A1)(τ).

To summarise the results of Theorems 2, 3, 5 and 6, the authors gave the following clear and useful table of the rate of convergence of best estimator in different asymptotic regimes.

TABLE 1

Rate of convergence of best estimator in different asymptotic regimes



	$n \ll \frac{k \log p}{\theta^2}$	$\frac{k \log p}{\theta^2} \ll n \ll \frac{k^2 \log p}{\theta^2}$	$n \gg \frac{k^2 \log p}{\theta^2}$
All estimators	$\asymp 1$	$\asymp \sqrt{\frac{k \log p}{n\theta^2}}$	$\asymp \sqrt{\frac{k \log p}{n\theta^2}}$
Polynomial time estimators	$\asymp 1$	$\asymp 1$	$\lesssim \sqrt{\frac{k^2 \log p}{n\theta^2}}$

We will call the three asymptotic regimes in Table 1 low, moderate and high effective sample size regime respectively. Theorem 6 is mainly concerned with the moderate effective sample size regime, i.e. where

$$\frac{k \log p}{\theta^2} \ll n \ll \frac{k^2 \log p}{\theta^2}.$$

This fact raises the question of whether computationally efficient procedures could attain a faster rate of convergence in the high effective sample size regime, i.e. where $n \gg \frac{k^2 \log p}{\theta^2}$.

Then Theorem 7 together with Algorithm 3 of the paper give a satisfying answer to this question, considering a subclass of $\mathcal{P}_p(n, k, \theta)$. A variant of $\hat{v}^{\text{SDP}}$ is proposed and attains the minimax optimal rate of convergence $\sqrt{\frac{k \log p}{n\theta^2}}$ in high effective sample size regime.

2.7 Main contributions of the paper

In our opinion, some main contributions of this paper could be listed as follows.

1. Addressed the question of computational efficiency for sparse PCA estimation;
2. Proposed a more robust condition, RRC condition, and define a class of more flexible distribution classes $\mathcal{P}_p(n, k, \theta)$;
3. Gave the minimax optimal rate of convergence for the class $\mathcal{P}_p(n, k, \theta)$.
4. Studied the rate of convergence for different kinds of estimators in different asymptotic regimes and gave a satisfying answer, namely Table 1;
5. Gave two estimators $\hat{v}^{\text{SDP}}$ and $\hat{v}^{\text{MSDP}}$ (and their corresponding algorithms) that are computable in polynomial time and that have a computational performance close (or equal) to the minimax rate of convergence for $\mathcal{P}_p(n, k, \theta)$ (or its subclass);
6. Adapted existing techniques to prove the fundamental trade-off between statistical and computational performance in sparse PCA estimation, assuming the hardness of the Planted Clique problem.

3 Implementation

In this section, we present our implementation for the experiments in 4 such that they can be easily reproducible.

3.1 Environment

Our code is written in `python` and we use the `numpy` package for random data generation and matrix manipulation and the `math` package for some basic computation.

3.2 Code Structure & Implementation Details

Our two core classes are namely *data_generator.py* and *sdp_estimator.py*. The former one enables us to draw $X_1, ..., X_n \overset{\text{i.i.d.}}{\sim} N_p(0, \Sigma)$, where $\Sigma := I_p + \theta v_1 v_1^T$, given p, k, n, θ. The latter one computes the semidefinite relaxation estimator $\hat{v}^{\text{SDP}}$ by following Algorithms 1 and 2 in the paper.

In Algorithm 2, for the projection $\Pi_{\mathcal{M}_1}(A)$ where A is a symmetric matrix $A = (A_{ij}) \in \mathbb{R}^{p \times p}$, we need to first decompose $A := PDP^T$ for some orthogonal P and diagonal $D = \text{diag}(d)$, where $d = (d_1, ..., d_p)^T \in \mathbb{R}^p$. This is done by doing an SVD. In the case where A is a real symmetric matrix, the first matrix given by the SVD is exactly the P we look for. The only issue is that in the implementation of `numpy.linalg.svd`, in the case where A contains negative eigenvalues, they will be turned into positive values in the diagonal matrix and the sign of some values in the last matrix will be changed. In other words, the diagonal matrix given by the SVD differs from the one we want by their signs. This is solved by simply checking the sign between the first matrix and the last one given by the SVD. In Step 3 in Algorithm 1, we adopt a simple implementation, which is using `numpy.linalg.eig` for the sake of simplicity and clarity. Although other methods such as the Lanczos method would require less operations, we note that this would not decrease the overall time complexity.

In the *main.py* class, we combine the two core classes to compute the loss between a generated v_1 and a semidefinite estimator. We follow the definition of the loss function in the paper. $L(v_1, \hat{v}^{\text{SDP}}) = \{1 - (v_1^T \hat{v}^{\text{SDP}})^2\}^{\frac{1}{2}}$. By iterating this process a number of times, we compute the mean loss.

4 Experiments and Results

In this section, we first repeat the numerical experiments done in the paper and confirm their results. Then, we try a slightly different setting that still lies in the context of the theorems in the paper.

4.1 Repeated Experiments

To make this review self-contained, we recall the content of the experiments done in the paper as the pseudo-code described as follows.

```
• Choose $p \in \{50, 100, 150, 200\}$. $k = \lfloor p^{1/2} \rfloor$.
• For $\nu_{\text{lin}} = 1$ to 1000, $n = \nu_{\text{lin}} k \log p$ (resp. $n = \nu_{\text{quad}} k^2 \log p$) do
    – For $i_{\text{rep}} = 1$ to $N_{\text{rep}} = 100$ do
        ∗ Generate $v_1$ by setting $v_{1,j} := k^{-1/2}$ for $j = 1, ..., k$ and $v_{1,j} := 0$ for $j = k+1, ..., p$.
        ∗ Draw $X_1, ..., X_n \overset{\text{i.i.d.}}{\sim} N_p(0, \Sigma)$, where $\Sigma := I_p + \theta v_1 v_1^T$ and $\theta = 1$.
        ∗ Compute the semidefinite estimator $\hat{v}^{SDP}$ of the data matrix $X := (X_1, ..., X_n)^T$.
        ∗ Compute the loss $L(v_1, \hat{v}^{\text{SDP}})$.
    – Compute the mean loss.
• Report the mean losses on a chart.
```

In Figure 2, we show the average loss of the estimator $\hat{v}^{\text{SDP}}$ over $N_{\text{rep}} = 100$ repetitions. The top left panel is against $\nu_{\text{quad}} := \frac{n\theta^2}{k \log p}$. The top right one is against $\nu_{\text{lin}} := \frac{n\theta^2}{k^2 \log p}$. To examine their tail behaviour, we replot them under logarithmic scales in the bottom left and bottom right panels.

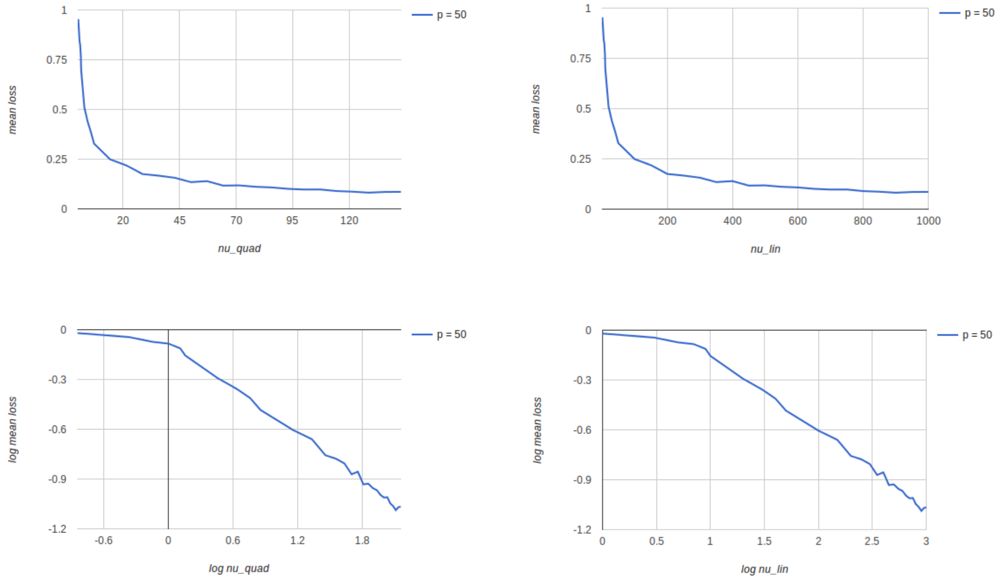


Figure 2: Average loss of the estimator $\hat{v}^{\text{SDP}}$ over $N_{\text{rep}} = 100$ repetitions against effective sample sizes ν_{quad} (top left) and ν_{lin} (top right). The tail behaviour under both scalings is examined under logarithmic scales in the bottom left and bottom right panels. (As we repeat the experiments, we try to organize in the same way. This caption is taken from the paper.)

Our results are very similar to those in the paper. Again, the top left panel shows a sharp phase transition for the mean loss, as predicted by Theorems 5 and 6 of the paper. The top right panels show that in the high effective sample size regime, $\hat{v}^{\text{SDP}}$ converges at rate $\sqrt{\frac{k \log p}{n\theta^2}}$, which is the same as that of the modified semidefinite relaxation estimator in Theorem 7 of the paper.

Due to the fact that the experiments are relatively time-consuming, we choose only to carry out the simulations for the setting where $p = 50$. Just to give an idea, for $p = 50$, it takes approximately three hours to obtain 30 data points (mean losses, each with 100 repetitions). We expect the experiments of $p = 100$ to take around a day.

4.2 Slightly Different Experiments

In this subsection, we vary the value of the parameter θ in the expressions of the covariance matrix

$$\Sigma := I_p + \theta v_1 v_1^T.$$

In the case where $\theta = 5$, the gap between the first eigenvalue and the second one is greater than that of $\theta = 1$. We expect the semidefinite estimator to be closer to the first eigenvector, i.e. the loss of the semidefinite estimator to be smaller. Results shown in Figure 3 correspond to our expectation. We observe that the curves of $\theta = 5$ have a similar shape to those of $\theta = 1$ and that the transition phase has almost disappeared, as predicted by Theorem 5 of the paper.

5 Conclusion

We gave an overview of the paper "Statistical and computational trade-offs in estimation of sparse principal components". We listed the main points made in the paper and we discussed each point in greater detail. After the discussion, we listed, from our point of view, the main contributions that the authors of the paper made. Then we implemented the algorithms proposed in the paper and carried out some numerical experiments. The results correspond very well to the theoretical analysis.

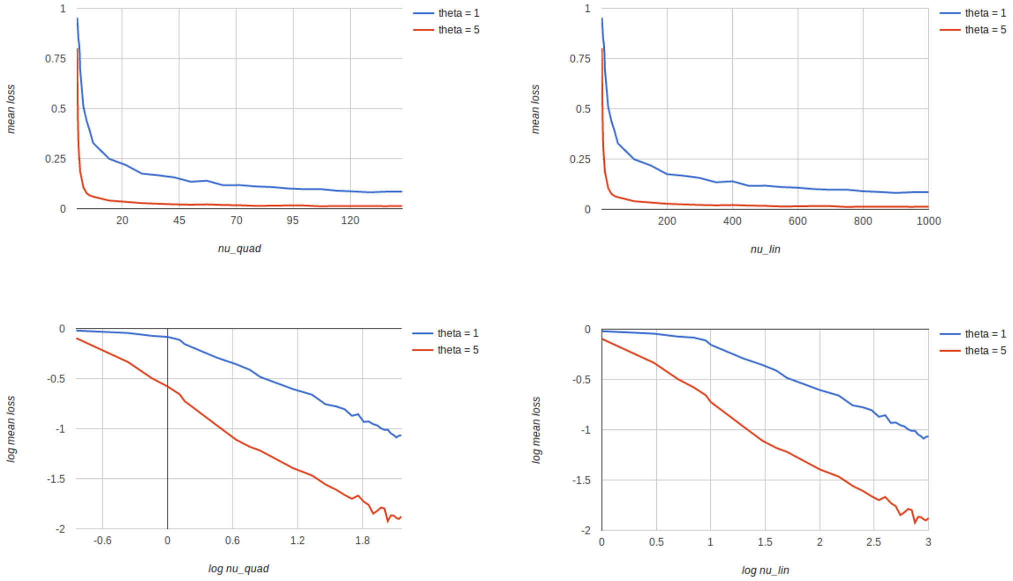


Figure 3: Average loss of the estimator $\hat{v}^{\text{SDP}}$ over $N_{\text{rep}} = 100$ repetitions against effective sample sizes ν_{quad} (top left) and ν_{lin} (top right) in two different settings: $\theta = 1$ (blue) and $\theta = 5$ (red). The tail behaviour under both scalings is examined under logarithmic scales in the bottom left and bottom right panels.

All of the statistical and computational analyses in the paper are performed over the the classes of distributions $\mathcal{P}_p(n, k, \theta)$ underpinned by the Restricted Covariance Concentration condition. Further research could be conducted to study other classes of distributions.

Epilogue

This is the first time that we two read an article in statistics on a state-of-the-art subject in detail. It was really not obvious at the beginning. We did not understand the notations and were not familiar with this domain, etc. But after reading it 4 or 5 times, the structure and the logic of the paper became clearer and clearer to us and we became more and more confident. So we would like to say that we are happy to have such experience of mini research in statistics. This will help us to be more confident when possible challenges in this domain occur to us in the future.

Acknowledgment

We would like to express our special thanks of gratitude to our professor Christophe Giraud, who taught us the foundations of high-dimensional data analysis and introduced us to this paper.

References

[1] Q. Berthet and P. Rigollet. Complexity theoretic lower bounds for sparse principal component detection. In *COLT*, pages 1046–1066, 2013.

[2] Q. Berthet, P. Rigollet, et al. Optimal detection of sparse principal components in high dimension. *The Annals of Statistics*, 41(4):1780–1815, 2013.

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

[4] R. M. Karp. Reducibility among combinatorial problems. In *Complexity of computer computations*, pages 85–103. Springer, 1972.

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

[6] V. Q. Vu, J. Lei, et al. Minimax sparse principal subspace estimation in high dimensions. *The Annals of Statistics*, 41(6):2905–2947, 2013.

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