HARDNESS OF SPARSE PCA — Braverman Reading Group

# LECTURE #1

SEPTEMBER 29, 2015

LECTURER: YOUNG KUN KO — SCRIBE: KIRAN VODRAHALLI

## Contents



## 1 Introduction

Summary of paper one: Choosing sparse principal components is as difficult as planted clique.

The papers to reference for this talk are [Berthet and Rigollet '13], [Chen and Papailiopoulos, Rubenstein '15].

Braverman: Everything you should take seriously are teaching/requirements for graduating and research. Braverman: If you're going to fake results, don't fake anything too exciting.

Today I will mainly follow the Rigollet paper. Let me just mention that the Rigollet paper is an average-case hardness result, while the $'15$ paper is the worst-case hardness result. An average case lower bound is more impressive. I will also explain what these hardness assumptions are.

## 2 Sparse PCA

Let us say you are given some matrix $X$, which is the observed data. If you want to find a coordinate system that explains this data the most, you want to find a $v$ such that we solve the optimization problem

$$\text{argmax}_{v \in S^{d-1}} v^T X^T X v$$

where we can view the rows of $X$ being $x_i \in \mathbb{R}^d$. Each $x_i$ comes from some distribution. We want to find the coordinates that explain the data the most. So $X$ is a random matrix variable. This problem can be solved in polynomial time (find largest eigenvalue of $X^T X$, but in a lot of cases, we want to add an additional constraint: Namely that $\|v\|_0 \leq k$ - a sparse vector! Keeping this in mind, we will try to explain the difference between the Rigollet paper and the $'15$ paper.

In this Rigollet paper, the problem they consider is statistical hypothesis testing, which is basically distinguishing between the distribution where there exists some sparse vector $v$ and a distribution such that for any $v$ which is $k$-sparse, the quantity we are trying to maximize is bounded (upper bound, not a lower bound). Intuitively, we are distinguishing between two distributions such that in one distribution, we know that there exists such a $v$ such that $v^T X^T X v$ is large, and in some other distribution, with high probability for any $v$, $v^T X^T X v$ is small. The $X$s are generated from some distribution.

In the $'15$ paper, we have the assumption $A = X^T X \geq 0$ (PSD), $\|v\|_2 = 1$, $\|v\|_0 \leq k$, and $x^T A x \geq c$ (the 'yes' category, it is lower bounded) and the 'no' category, $x^T A x < s$ (there is no lower bound). Note

that there are no distributions involved here. Here $A$ is fixed. This is not average case, so we cannot construct $X$. The rows of $X$ must be drawn independently. We can take an empirical distribution which includes some bad instances. There is not a well-posed distinction between the formulations in the two papers.

## 2.1 Statistical Hypothesis Testing

More formally, statistical hypothesis testing can be framed in the following framework.

**Definition 2.1.** Empirical variance.
We have the empirical variance is given by $\hat{V}_n(u) = \frac{1}{n}\sum_{i=1}^n u^T X_i^T X_i u$, with respect to the vector $u$.

We have for any vector $u$ that is $k$-sparse, if the distribution is not skewed towards one vector, by some Chernoff Bernstein inequality this holds. If it is skewed, then you are not so lucky and you are skewed. This $\theta$ is denoted 'signal-strength'. We want to find the minimum $\theta$ such that we can distinguish between our two distributions. $\nu$ is a term that comes in the Chernoff bound.

At a high level, you sample $n$ independent copies of it and compute some statistics. The first case is the null hypothesis, so there is no special $k$-sparse direction that there explains anything, and the second one there is some $k$-sparse direction which does explain the variance. The null hypothesis should not depend on anything. The first probability is isotropic.

1. $\sup_{u\in S^{d-1}} P_0^{\times n}\left(|\hat{V}_n(u) - 1| > 4\sqrt{\frac{\log(1/\nu)}{n}} + \frac{4\log(1/\nu)}{n}\right) \leq \nu$: This is the CDF of the tails of the distribution.
2. $P_{v,\|v\|_0\leq k}^{\otimes n}\left(\hat{V}_n(v) - (1+\theta) < -2\sqrt{\frac{2\theta k\log(2/\nu)}{n}} - \frac{4\log(2/\nu)}{n}\right) \leq \nu$.

Then we write $H_0$ is a distribution such that property 1 holds, $H_1$ such that property 2 holds. Restating, the goal is to find the minimum $\theta$ such that you can distinguish between $H_0$ and $H_1$.

Another way to put it is to state that the union between Type $I$ and Type $II$ error is less than some constant.

Let us write down what is known about $\theta$. We have statistical lower bound on $\theta$ and SDP lower bound, meaning the SDP gap. If you try to pose that problem as some SDP, what is the $\theta$ such that SDP can distinguish between the two. So the overarching question in this line of research: It is not that hard to compute the statistical bounds, so in what cases are there computational issues when you have statistical bounds, you can't actually reach it without hitting some intractability. This problem is nonconvex, so this computational time could be quite bad.

The statistical bound is given by $\mathcal{O}\left(\sqrt{\frac{k\log(d)}{n}}\right)$, and the SDP lower bound on $\theta$ is $\mathcal{O}\left(Ck\sqrt{\frac{\log(d)}{n}}\right)$. So there is a gap of $\sqrt{k}$.

**Theorem 2.2.** *If there exists a test $T$ that distinguishes between $H_0$ and $H_1$ with signal strength $C \cdot k^{\frac{1}{2}+o(1)}\sqrt{\frac{\log(d)}{n}}$, such a $T$ can be used to solve planted clique.*

### 2.1.1 Planted Clique Conjecture

You are given a graph $G \sim G(n, \frac{1}{2})$ from Erdos-Renyi model. We could call this graph $H_0$. Then we can call $H_1$ $G \sim G(n, \frac{1}{2}, k)$. The goal is then to distinguish whether $G$ came from $H_0$ or $H_1$. If $k \leq 2\log(n)$, then $H_0$ and $H_1$ are statistically indistinguishable. If $k = \Omega(\sqrt{n})$, there exist polynomial time algorithms. This can be seen by looking at variance on number of edges, since number of edges are on order $n^2$, we see that variance is on the order of $n$. People did not seriously try to break $\Omega(\sqrt{n})$. Noga Alon tried for a bit.

Sum-of-square algorithms should fail? Note that you might not need SDP, maybe some fancy combinatorial algorithm.

Now let us return to the theorem. We define a class $R_\mu = R_0 \cap \{k \geq n^\mu\} \cap \{n < d\}$, and $R_0 = \{(d, n, k) | k \leq d^{0.49}, 15\sqrt{k \log(6ed/\delta)/n} \leq 1\}$ and we define a set of parameters on $(d, n, k)$ such that it is indeed 'hard'.

The main idea is as follows: $G = (V, E) \sim G_{2m}$ with planted clique of size $X$. Then we choose $V_l$ as an $m$-sized subgraph and $V_r$ as an $n$-sized subgraph. Take a random intersection of $G$ such that this intersects both $V_l$ and $V_r$. This creates an induced subgraph. We have added $d - m$ dummy vertices to $V_l$ and connect them with probability $\frac{1}{2}$ to $V_r$. There will be a complete bipartite subgraph of some random size if you can distinguish them.

So we have a reduction from some graph $G$ to a matrix of $n \times d$ size. where we write $\eta_i(2B_i - 1) = X_i$, where $B_i$ is an encoded element of $V_r$. Then, we can write $X_G$ as the $n \times d$ matrix. This gets us back to the sparse PCA setting. Thus we have some reduction from planted clique to sparse PCA.

Let's say you have the artificial way from the graph you can generate a distribution like $P_1^{bl(G)}$, and the natural way is the distribution $P_v^{\otimes n}$. At a very high level, you need to show that $\exists v$ such that $P_1^{bl(G)}$ and $P_v^{\otimes n}$ are 'close'. Thus if you can solve these guys are close, you can solve $P_0^{\otimes n} = P_0^{bl(G)}$, and so you can solve planted clique.

What is the key issue in showing that there exists a $v$ such that the two distributions are close. $P_v^{\otimes n}$ is independent draws from a distribution, but $P_1^{bl(G)}$ is not independent. The whole issue is that $P_1^{bl(G)}$ is not 'independent' while $P_v^{\otimes n}$ has to be by definition. You can make $P_1^{bl(G)}$ independent if you set $X_i$ 1 with probability $m/n$ - translates to some conditioning on drawing edges?

# 3 Worst-case Hardness

Here are the results from the second paper. If you assume NP-hard, then it is hard to distinguish between $v^T Av$ and $(1 - \Omega(1))v^T Av$. If you assume SSE-hard, then it's hard to distinguish between $cv^T Av$ and $sv^T Av$ constants $c, s$. The scaling matters because you don't know if you can rule out certain things. . SSE implies unique conjecture, but not vice-versa. Both reductions are straight-forward. One surprising result is that the sparsity $k = Theta(n)$ for both cases. You can't hope for sub-polynomial $k$.

Question for next time: So what about sparse PCA on Gaussians with some covariance matrix (not uniform)? Is it hard to do sparse PCA on that? Gaussians are probably the most natural thing - look at Tengyu's manuscript on arXiv, posted about a month ago.