Methods of Moments for Recovering Spectrum from Samples

Zhili Feng
UIUC
zfeng6@illinois.edu

Abstract

In this project we look into how to recover spectrum from a noisy sample of the true population covariance. Two algorithms are given, one for estimating moments, and the other for recovering spectrum from the estimated moments. Upperbound of the estimated spectrum is also derived in L1 and Wasserstein sense.

1 Introduction

Matrix completion and estimation from samples are very important tasks nowadays. For example, in a recommendation system, like the one used by Netflix, we have a large matrix with rows representing the users, columns representing the movies, and each (i, j) entry represents the score user i gives to movie j. Nonetheless, these matrices are usually very large to directly estimate, or in some scenarios, we can only observe a noisy version of the matrices. Therefore, researchers want to exploit the low-dimensional nature of such matrices (if exists), or reconstruct some features of the original matrix from the observed one. The reconstructed features are usually eigenvalues, or spectrum, which presents the vector of eigenvalues.

1.1 Setup and Goals

Consider dealing with multivariate distribution over $\mathbb{R}^d$, represented by $X(0, 1)$ with mean 0 and variance 1. We assume X has boudned fourth moment β.
Define a real $d \times d$ matrix $\mathbf{S}$, which is unobserved. Instead, we observe a noisy version $\mathbf{Y} = \mathbf{XS}$, where $\mathbf{X} \in \mathbb{R}^{\mathbf{n}\times\mathbf{d}}$, we draw each column $\mathbf{X}_{(\cdot,\mathbf{i})}$, $\forall i \in n$ from distribution $X(0, 1)$.
We want to estimate $\mathbf{S}^\top\mathbf{S}$, which is also known as the *population covariance.*

1.2 Overview

To achieve the goal of accurately estimating the spectrum of the original matrix by only observing the noisy version, the researchers propose a method of moments, where we first estimate the first k moments of the matrix through a cycle-counting approach, then we create a distribution based on the estimated moments. Then we recover the spectrum from the hand-crafted distribution.
We would first see this recovered spectrum is close to the original spectrum in Wasserstein distance, then we would see those two spectrums are also closed in L1 distance.

2 Estimating Moments

Let A be a $n \times n$ matrix, a k-cycle σ on A is defined to be a sequence of k integers bounded by $[0.n]$ and $\sigma = (\sigma_1 \dots \sigma_k)$. We consider closed walk here so let $\sigma_{k+1} = \sigma_1$.
Define a product $A_\sigma = \prod_{i=1}^{k} A_{\sigma_i,\sigma_{i+1}}$. Then we have the following fact:
Let σ be a k-cycle, X be a $n \times d$ matrix, where the column of X is drawn from a distribution with

mean 0 and variance 1. Let T be a $d \times d$ real matrix, but unobserved. We want to estimate k^{th} moment of T, defined as $Tr(T^k)$, and $\mathbb{E}\left[(X^\top TX)_\sigma\right] = Tr(T^k)$ is an unbiased estimator of the k^{th} spectral moment of T.
To prove, simply expand the formula:

$$\begin{aligned}
\mathbb{E}\left[(X^\top TX)_\sigma\right] &= \mathbb{E}\left[\prod_{i=1}^{k} \sum_{\delta_i, \gamma_i \in [d]} X_{\delta_i, \sigma_i} T_{\delta_i, \gamma_{i+1}} X_{\gamma_{i+1}, \sigma_{i+1}}\right] \\
&= \sum_{\delta_1^k, \gamma_1^k} \mathbb{E}\left[\prod_{i=1}^{k} X_{\delta_i, \sigma_i} T_{\delta_i, \gamma_{i+1}} X_{\gamma_{i+1}, \sigma_{i+1}}\right] \\
&= \sum_{\delta_1^k} \prod_{i=1}^{k} T_{\delta_i, \delta_{i+1}} \\
&= Tr(T^k)
\end{aligned}$$

As an example, let X, T be two 2×2 matrices. Let $\sigma = (1, 2)$ be a 2-cycle. Expanding the formula above we have:

$$\begin{aligned}
\mathbb{E}\left[(X^\top TX)_\sigma\right] &= \mathbb{E}\left[(X^\top TX)_{1,2} \cdot (X^\top TX)_{2,1}\right] \\
&= \mathbb{E}[(X_{11}T_{11}X_{12} + X_{21}T_{21}X_{12} + X_{11}T_{12}X_{22} + X_{21}T_{22}X_{22}) \\
&\cdot (X_{12}T_{11}X_{11} + X_{22}T_{21}X_{11} + X_{12}T_{12}X_{21} + X_{22}T_{22}X_{21})] \\
&= T_{11}^2 + T_{12}T_{21} + T_{22}^2 \\
&= Tr(T^2)
\end{aligned}$$

Even though this is an unbiased estimator, the variance of each k-cycle could be very large. To reduce the variance, we could average over all k-cycles, but counting all k-cycles would be NP-hard. Therefore, the researchers propose to only consider the increasing cycles.

2.1 Increasing Cycles

Define increasing k-cycle to be $\sigma = (\sigma_1 \ldots \sigma_k)$ such that $\sigma_1 < \sigma_2 < \ldots < \sigma_k$. Since the product A_σ is a multiplication of $A_{\sigma_i, \sigma_{i+1}}$, an increasing cycle essentially means the element we choose to multiply in A has bigger row index than column index. In other words, we only consider the upper triangular entries of A. Hence to calculate the k^{th} moments from the increasing cycle, we pad the lower-triangular and diagonal entries of matrix A in the first $k-1$ copies:

```
Data: $Y \in \mathbb{R}^{n \times d}$
Result: kth spectral moment
Let $A = YY^\top$, $G = A_{up}$ be the matrix in which we pad lower-triangular and diagonal entries
  with 0.
Output $\frac{Tr(G^{k-1}A)}{d\binom{n}{k}}$
```

Algorithm 1: Estimating the kth moment

The guarantee of variance is given by

Lemma 2.1

$$Var\left(\frac{1}{|U|} \sum_{\sigma \in U} (X^\top TX)_\sigma\right) = f(k) \frac{\max(d^{k-2}, 1)}{n^k} Tr(T^k)^2$$

where $f(k) = 2^{12k} k^{6k} \beta^k$, β is the fourth moment of X, U is the set of all k-increasing-cycle.

3 Recover Spectrum from Moments

We would first create a discrete distribution based on the first k moments we calculate, via a linear programming algorithm:

```
Data: $\hat{\alpha}$: vector of k estimated moments; d: dimensionality of population variance; partition
    $x = (x_0 \dots x_t)$ on $[0, b]$; b: upperbound of population eigenvalue. Notice we can make
    $x_i = i\epsilon$ where $\epsilon \leq \frac{1}{\max(n,d)}$
Result: Estimated spectrum $\hat{\lambda}_1 \dots \hat{\lambda}_d$
1. Let $p^+$ be the distribution that solves the following linear programming:
       $\begin{aligned} \underset{p}{\text{minimize}} \quad & |Vp - \hat{\alpha}|_1 \\ \text{subject to} \quad & \mathbf{1}^\top p = 1, \\ & p > 0 \end{aligned} \qquad (1)$
    Where $V \in \mathbb{R}^{k \times t}$, $V_{ij} = x_j^i$.
2. Output spectrum $\hat{\lambda}_1 \dots \hat{\lambda}_d$ where $\hat{\lambda}_i = \min(x_j : \sum_{l \leq j} p_l^+ \geq \frac{i}{d+1})$.
```

Algorithm 2: Recover Spectrum from Moments

This algorithm first generate a discrete distribution, with a t partition. Notice we can equally partition $[0, b]$ with mesh $\epsilon \leq \frac{1}{\max(n,d)}$, then the number of partition $t \geq \max(n, d)$.
Now given a fixed partition x, we need to assign each x_i a probability mass. Notice the matrix V represent the moment of x from 1 to k. In other words, the first row of V represents the first moment of partition x, the second row represents the second moment, so on and so forth. p in the linear programming is a probability vector, it's straightforward to see that the minimizer p^+ minimizes the L1 distance between the first k moments of p and the estimated first k moments $\hat{\alpha}$ of the true distribution.
As the last step, output the i^{th}-quantile of the estimated distribution p^+ as the estimated spectrum.

4 Upperbound on the Recovered Spectrum

We would first show that the estimated distribution p^+, and the true distribution p, is closed in Wasserstein distance. Then we show that the L1 distance is also close.

4.1 Wassserstein Distance

Wasserstein distance, also known as the earthmover distance, measures the minimum cost of moving one probability distribution p to match another distribution q. It is defined as following:

$$W_1(p, q) = \sup_{f:1-Lipshitz} \int f(x)(p(x) - q(x))dx$$

An important lemma for the derivation of upperbound is that

$$W_1(p, q) \leq C\frac{b-a}{k} + g(k)(b-a)\|\alpha - \beta\|_2$$

where α and β are the first k moments of distribution p and q respectively, C is a constant, and $g(k) = c'3^k$ for another constant C'.
The outline of the proof is:

1. Show that the L1 distance between α^+ and α is bounded.
2. Show that the L2 distance between α^+ and α is bounded.
3. Show that the Wasserstein distance between α^+ and α is bounded use the above lemma.

4.1.1 Upperbound on L1 Distance

Recall Lemma 2.1, we have an upperbound on $Var(\hat{\alpha})$. Notice that α is a fixed constant vector without randomness, hence $Var(\hat{\alpha}) = Var(\hat{\alpha} - \alpha)$.
Meanwhile, as shown in section 2, each k-cycle is an unbiased estimator of the k^{th} moment α_k.

Hence $\mathbb{E}\left[\|\hat{\alpha}-\alpha\|^2\right] = Var(\hat{\alpha}-\alpha)$. We also have inequality $\mathbb{E}\left[X\right]^2 \le \mathbb{E}\left[X^2\right]$, for any random variable X. This gives us the following bound:

$$\begin{aligned}
\mathbb{E}\left[\|\hat{\alpha}-\alpha\|_1\right] &= \sum_{i=1}^{k} \mathbb{E}\left[|\hat{\alpha}_i-\alpha_i|\right] \\
&\le \sum_{i=1}^{k} \mathbb{E}\left[(\hat{\alpha}_i-\alpha_i)^2\right] \\
&= \sum_{i=1}^{k} Var(\hat{\alpha}_i-\alpha_i) \\
&= \sum_{i=1}^{k} f(i)\frac{\max(d^{i/2-1},1)}{n^{i/2}} b^i
\end{aligned}$$

where $f(i) = 2^{6i}i^{3i}\beta^{i/2}$, b is the upperbound on eigenvalue of matrix S in our setup. Compare this upperbound with the inequality in lemma 2.1. You can see it is simply a linear combination of the square root of the variance.

Notice in the linear programming 2, our partition x may not align with $\hat{\alpha}$ perfectly. Instead, x is constructed by making $x_i = i\epsilon$, meaning we round $\hat{\alpha}$ to its nearest integer multiple of ϵ. Also notice partition x covers $[0,b]$, so the largest deviation for the rounding of the i^{th} moment is $(b+\epsilon)^i - b^i$, since $(\cdot)^i$ is a monotone increasing function with increasing derivative as the parameter increasing (when i is positive).

Hence by triangle inequality:

$$\mathbb{E}\left[\|\alpha^+-\alpha\|_1\right] \le \mathbb{E}\left[\|\alpha^+-\hat{\alpha}\|_1\right] + \mathbb{E}\left[\|\hat{\alpha}-\alpha\|_1\right] \le 2\mathbb{E}\left[\|\hat{\alpha}-\alpha\|_1\right] + \sum_{i=1}^{k}\left((b+\epsilon)^i - b^i\right)$$

4.1.2 Upperbound on L2 Distance

Notice the inequality:

$$\sqrt{a^2+b^2} \le |a+b|$$

Hence we can upperbound L2 distance by L1 distance:

$$\begin{aligned}
\mathbb{E}\left[\|\alpha^+-\alpha\|_2\right] &\le \mathbb{E}\left[\|\alpha^+-\alpha\|_1\right] \\
&\le 2\mathbb{E}\left[\|\hat{\alpha}-\alpha\|_1\right] + \sum_{i=1}^{k}\left((b+\epsilon)^i - b^i\right) \\
&\le 2\left(\sum_{i=1}^{k} f(i)\frac{\max(d^{i/2-1},1)}{n^{i/2}} b^i + (b+\epsilon)^i - b^i\right) \\
&\le 2k\left(f(k)\frac{2^{k/2-1}}{n^{k/2}} b^k + f(1)\frac{b}{n^{1/2}} + 2^k(b^{k-1}\epsilon)\right)
\end{aligned}$$

We use triangle inequality to expand the L1 distance. In the last inequality, $f(k)\frac{2^{k/2-1}}{n^{k/2}}b^k$ accounts for the case when $d^{i/2-1} > 1$ and $f(1)\frac{b}{n^{1/2}}$ accounts for the case when $d^{i/2-1} \le 1$. $2^k(b^{k-1}$ is attained by expanding $(b+\epsilon)^k - b^k$, there are in total 2^k terms after the expansion, and the largest one is b^k, which cancels with $-b^k$. The second largest on is $b^{k-1}\epsilon$, which upperbounds all other terms.

4.1.3 Upperbound on Wasserstein Distance

We need the following two facts:

Fact 4.1 *Let two sorted vectors $a = (a_1, \ldots, a_d)$ and $b = (b_1, \ldots, b_d)$. Let distribution p_a represent a discrete distribution with weight $1/d$ on each $a_i \in a$, similarly for p_b. Then:*

$$|a-b|_1 = d \cdot W_1(p_a, p_b)$$

Fact 4.2 *Let distribution p be supported on $[a,b]$. Let distribution p' be a discrete distribution with mass $1/d$ on each d of $(d+1)$-quantile of distribution p. Then*

$$W_1(p,p') \leq \frac{b-a}{d}$$

Hence we can bound the Wasserstein distance, let p^+ be the distribution produced by Algorithm 2, let p^+_{quant} be the distribution with equal pointed mass $1/d$ on each of $(d+1)$-quantile of p^+:

$$\begin{aligned} W_1(p^+_{quant}, p) &\leq W_1(p^+_{quant}, p^+) + W_1(p^+, p) \\ &\leq \frac{b}{d} + b(\frac{C}{k} + g(k)\|\alpha^+ - \alpha\|_2) \end{aligned}$$

$W_1(p^+_{quant}, p^+)$ is bounded according to fact 4.2. $W_1(p^+, p)$ is bounded in Section 4.1.2. Using fact 4.1, we can then bound the spectrum.

5 Wasserstein Distance and L2 Distance

In Section 4.1 we mention an important bound on Wasserstein distance:

$$W_1(p,q) \leq C\frac{b-a}{k} + g(k)(b-a)\|\alpha - \beta\|_2$$

where α and β are the first k moments of distribution p and q respectively, C is a constant, and $g(k) = c'3^k$ for another constant C'.
We would prove this inequality (relaxed version) via a polynomial approximation technique.
Suppose distribution p and q have matched first k moments. Then any polynomial P with degree at most k, we have

$$\int P(x)(p(x) - q(x))dx = 0$$

Let P_f be a polynomial with degree at most k. Use P_k to approximate 1-Lipschitz function f, we have

$$\begin{aligned} \int f(x)(p(x)-q(x))dx &= \int |f(x) + P_f(x) - P_f(x)|(p(x)-q(x))dx \\ &\leq \int |P_f(x) - f(x)|(p(x)-q(x))dx + \int P_f(x)(p(x)-q(x))dx \\ &\leq \int |P_f(x) - f(x)|p(x)dx + \int |P_f(x) - f(x)|q(x)dx \\ &\leq 2\|f - P_f\|_\infty \end{aligned}$$

The first inequality is just triangular inequality. Notice $\int P_f(x)(p(x)-q(x))dx = 0$ since p and q have matched first k moments. The second inequality comes from the fact that:

$$\int |f+g|dx \leq \int |f|dx + \int |g|dx$$

and the last inequality is derived from the definition of infinity norm.

Fact 5.1 *Let g be a k+1 differentiable function on [a,b], there exists P_g with at most degree k such that*

$$\|g(x) - P_g(x)\|_\infty \leq (\frac{b-a}{2})^{k+1}\frac{\max(g^{k+1}(x))}{2^k(k+1)!}$$

Although our function f is Lipschitz, this doesn't guarantee the existence or boundedness of higher derivative. To compensate for this, we need another function f_s such that $\|f - f_s\|_\infty$ is small, while f_s has bounded higher derivatives. Naturally, we can find another function h that has small higher derivatives, and convolutes with f: $f_s = f * h$. By the definition of convolution, we have:

$$(f_s)^{(k+1)}(x) = (f * h^{(k+1)})(x)$$

Consider function

$$b(y) = \begin{cases} e^{-\frac{y^2}{1-y^2}} & |y| < 1 \\ 0 & otherwise \end{cases}$$

We will let the convoluted function $h = c\hat{b}(cy)$, where $\hat{b}(y)$ is the Fourier transform of $b(y)$, and c is a constant of our choice. Properties of $\hat{b}$ include $\|\hat{b}^{(k)}\|_1 = O(1/k)$ and $b^{(k)}\|_\infty = O(1)$.
Let $f_s = c\hat{b}(cy) * f$. By triangular inequality:

$$\int f(x)(p(x) - q(x))dx \leq 2\|f - P_f\|_\infty \leq 2\|f - f_s\|_\infty + 2\|f_s - P_f\|_\infty$$

For the first term:

$$\begin{aligned} |f(x) - f_s(x)| &= |f(x) - \in f(x-t)\hat{b}_c(t)dt| \\ &= |f(x)(1 - \int \hat{b}_c(t)dt) + \int (f(x) - f(x-t))\hat{b}_c(t)| \\ &\leq \int |t\hat{b}_c(t)dt| \\ &= O(1/c) \end{aligned}$$

Notice that $\int \hat{b}_c(t)dt = 1.$, so the first term in the second line is 0. Since f is 1-Lipschitz:

$$|f(x) - f(x-t)| \leq |x - x + t| = t$$

The last inequality is given by [1].
For the second term, bound the $k+1$ derivative of f_s:

$$\begin{aligned} |(f_s)^{(k+1)}|_\infty &= c^{k+1}|(f * (\hat{b}^{(k+1)})_c)(x)|_\infty \\ &\leq c^{k+1}|f|_\infty|\hat{b}^{(k+1)}|_1 \\ &= O(\frac{b-a}{2}c^{k+1}) \end{aligned}$$

where the first line is the definition of derivative of convolutions, the second line is Cauchy-Schwartz, and the third line is derived from the properties of f and $\hat{b}$.
Then by Fact 5.1:

$$\begin{aligned} \|f_s - P_f\|_\infty &\leq (\frac{b-a}{2})^{k+1}\frac{\max(g^{k+1}(x))}{2^k(k+1)!} \\ &= O\left((\frac{b-a}{2})^{k+2}\frac{c^{k+1}}{2^k(k+1)!}\right) \end{aligned}$$

Choose $c = \frac{k}{b-a}$, we will get the desired bound on Wasserstein distance.

References

[1] Kane, Daniel M., Jelani Nelson, and David P. Woodruff. ”On the exact space complexity of sketching and streaming small norms.” Proceedings of the twenty-first annual ACM-SIAM symposium on Discrete Algorithms. Society for Industrial and Applied Mathematics, 2010.