18.S096: Principal Component Analysis in High Dimensions and the Spike Model

Topics in Mathematics of Data Science (Fall 2015)

Afonso S. Bandeira
bandeira@mit.edu
http://math.mit.edu/~bandeira

September 18, 2015

These are lecture notes not in final form and will be continuously edited and/or corrected (as I am sure it contains many typos). Please let me know if you find any typo/mistake. Also, I am posting shorter versions of these notes (together with the open problems) on my Blog, see [Ban15b].

0.1 Brief Review of some linear algebra tools

In this Section we'll briefly review a few linear algebra tools that will be important during the course. If you need a refresh on any of these concepts, I recommend taking a look at [HJ85] and/or [Gol96].

0.1.1 Singular Value Decomposition

The Singular Value Decomposition (SVD) is one of the most useful tools for this course! Given a matrix $M \in \mathbb{R}^{m\times n}$, the SVD of M is given by

$$M = U\Sigma V^T, \tag{1}$$

where $U \in O(m)$, $V \in O(n)$ are orthogonal matrices (meaning that $U^TU = UU^T = \mathrm{I}$ and $V^TV = VV^T = \mathrm{I}$) and $\Sigma \in \mathbb{R}^{m\times n}$ is a matrix with non-negative entries in its diagonal and otherwise zero entries.

The columns of U and V are referred to, respectively, as left and right singular vectors of M and the diagonal elements of Σ as singular values of M.

Remark 0.1 *Say $m \leq n$, it is easy to see that we can also think of the SVD as having $U \in \mathbb{R}^{m\times n}$ where $UU^T = \mathrm{I}$, $\Sigma \in \mathbb{R}^{n\times n}$ a diagonal matrix with non-negative entries and $V \in O(n)$.*

0.1.2 Spectral Decomposition

If $M \in \mathbb{R}^{n\times n}$ is symmetric then it admits a spectral decomposition

$$M = V\Lambda V^T,$$

where $V \in O(n)$ is a matrix whose columns v_k are the eigenvectors of M and Λ is a diagonal matrix whose diagonal elements λ_k are the eigenvalues of M. Similarly, we can write

$$M = \sum_{k=1}^{n} \lambda_k v_k v_k^T.$$

When all of the eigenvalues of M are non-negative we say that M is positive semidefinite and write $M \succeq 0$. In that case we can write

$$M = \left(V\Lambda^{1/2}\right)\left(V\Lambda^{1/2}\right)^T.$$

A decomposition of M of the form $M = UU^T$ (such as the one above) is called a Cholesky decomposition.

The spectral norm of M is defined as

$$\|M\| = \max_k |\lambda_k(M)|.$$

0.1.3 Trace and norm

Given a matrix $M \in \mathbb{R}^{n\times n}$, its trace is given by

$$\mathrm{Tr}(M) = \sum_{k=1}^{n} M_{kk} = \sum_{k=1}^{n} \lambda_k(M).$$

Its Frobeniues norm is given by

$$\|M\|_F = \sqrt{\sum_{ij} M_{ij}^2} = \mathrm{Tr}(M^T M)$$

A particularly important property of the trace is that:

$$\mathrm{Tr}(AB) = \sum_{i,j=1}^{n} A_{ij}B_{ji} = \mathrm{Tr}(BA).$$

Note that this implies that, e.g., $\mathrm{Tr}(ABC) = \mathrm{Tr}(CAB)$, it does not imply that, e.g., $\mathrm{Tr}(ABC) = \mathrm{Tr}(ACB)$ which is not true in general!

0.2 Quadratic Forms

During the course we will be interested in solving problems of the type

$$\max_{\substack{V\in\mathbb{R}^{n\times d}\\ V^TV=\mathrm{I}_{d\times d}}} \mathrm{Tr}\left(V^TMV\right),$$

where M is a symmetric $n \times n$ matrix.

Note that this is equivalent to

$$\max_{\substack{v_1,\ldots,v_d\in\mathbb{R}^n \\ v_i^T v_j=\delta_{ij}}} \sum_{k=1}^{d} v_k^T M v_k, \tag{2}$$

where δ_{ij} is the Kronecker delta (is 1 is $i=j$ and 0 otherwise).

When $d=1$ this reduces to the more familiar

$$\max_{\substack{v\in\mathbb{R}^n \\ \|v\|_2=1}} v^T M v. \tag{3}$$

It is easy to see (for example, using the spectral decomposition of M) that (3) is maximized by the leading eigenvector of M and

$$\max_{\substack{v\in\mathbb{R}^n \\ \|v\|_2=1}} v^T M v = \lambda_{\max}(M).$$

It is also not very difficult to see (it follows for example from a Theorem of Fan (see, for example, page 3 of [Mos11]) that (2) is maximized by taking $v_1,\ldots,v_d$ to be the k leading eigenvectors of M and that its value is simply the sum of the k largest eigenvalues of M. The nice consequence of this is that the solution to (2) can be computed sequentially: we can first solve for $d=1$, computing v_1, then v_2, and so on.

Remark 0.2 *All of the tools and results above have natural analogues when the matrices have complex entries (and are Hermitian instead of symmetric).*

1.1 Dimension Reduction and PCA

When faced with a high dimensional dataset, a natural approach is to try to reduce its dimension, either by projecting it to a lower dimension space or by finding a better representation for the data. During this course we will see a few different ways of doing dimension reduction.

We will start with Principal Component Analysis (PCA). In fact, PCA continues to be one of the best (and simplest) tools for exploratory data analysis. Remarkably, it dates back to a 1901 paper by Karl Pearson [Pea01]!

Let's say we have n data points $x_1,\ldots,x_n$ in $\mathbb{R}^p$, for some p, and we are interested in (linearly) projecting the data to $d<p$ dimensions. This is particularly useful if, say, one wants to visualize the data in two or three dimensions. There are a couple of different ways we can try to choose this projection:

1. Finding the d-dimensional affine subspace for which the projections of $x_1,\ldots,x_n$ on it best approximate the original points $x_1,\ldots,x_n$.
2. Finding the d dimensional projection of $x_1,\ldots,x_n$ that preserved as much variance of the data as possible.

As we will see below, these two approaches are equivalent and they correspond to Principal Component Analysis.

Before proceeding, we recall a couple of simple statistical quantities associated with $x_1, \ldots, x_n$, that will reappear below.

Given $x_1, \ldots, x_n$ we define its sample mean as

$$\mu_n = \frac{1}{n} \sum_{k=1}^{n} x_k, \tag{4}$$

and its sample covariance as

$$\Sigma_n = \frac{1}{n-1} \sum_{k=1}^{n} (x_k - \mu_n)(x_k - \mu_n)^T. \tag{5}$$

Remark 1.3 *If $x_1, \ldots, x_n$ are independently sampled from a distribution, μ_n and Σ_n are unbiased estimators for, respectively, the mean and covariance of the distribution.*

We will start with the first interpretation of PCA and then show that it is equivalent to the second.

1.1.1 PCA as best d-dimensional affine fit

We are trying to approximate each x_k by

$$x_k \approx \mu + \sum_{i=1}^{d} (\beta_k)_i \, v_i, \tag{6}$$

where $v_1, \ldots, v_d$ is an orthonormal basis for the d-dimensional subspace, $\mu \in \mathbb{R}^p$ represents the translation, and β_k corresponds to the coefficients of x_k. If we represent the subspace by $V = [v_1 \cdots v_d] \in \mathbb{R}^{p \times d}$ then we can rewrite (7) as

$$x_k \approx \mu + V\beta_k, \tag{7}$$

where $V^T V = \mathrm{I}_{d \times d}$ as the vectors v_i are orthonormal.

We will measure goodness of fit in terms of least squares and attempt to solve

$$\min_{\substack{\mu,\, V,\, \beta_k \\ V^T V = \mathrm{I}}} \sum_{k=1}^{n} \|x_k - (\mu + V\beta_k)\|_2^2 \tag{8}$$

We start by optimizing for μ. It is easy to see that the first order conditions for μ correspond to

$$\nabla_\mu \sum_{k=1}^{n} \|x_k - (\mu + V\beta_k)\|_2^2 = 0 \Leftrightarrow \sum_{k=1}^{n} (x_k - (\mu + V\beta_k)) = 0.$$

Thus, the optimal value μ^* of μ satisfies

$$\left(\sum_{k=1}^{n} x_k \right) - n\mu^* - V \left(\sum_{k=1}^{n} \beta_k \right) = 0.$$

Because $\sum_{k=1}^{n} \beta_k = 0$ we have that the optimal μ is given by

$$\mu^* = \frac{1}{n} \sum_{k=1}^{n} x_k = \mu_n,$$

the sample mean.

We can then proceed on finding the solution for (9) by solving

$$\min_{\substack{V,\ \beta_k \\ V^T V = \mathrm{I}}} \sum_{k=1}^{n} \|x_k - \mu_n - V\beta_k\|_2^2 . \tag{9}$$

Let us proceed by optimizing for β_k. Since the problem decouples for each k, we can focus on, for each k,

$$\min_{\beta_k} \|x_k - \mu_n - V\beta_k\|_2^2 = \min_{\beta_k} \left\| x_k - \mu_n - \sum_{i=1}^{d} (\beta_k)_i \, v_i \right\|_2^2 . \tag{10}$$

Since $v_1, \ldots, v_d$ are orthonormal, it is easy to see that the solution is given by $(\beta_k^*)_i = v_i^T (x_k - \mu_n)$ which can be succinctly written as $\beta_k = V^T (x_k - \mu_n)$. Thus, (9) is equivalent to

$$\min_{V^T V = \mathrm{I}} \sum_{k=1}^{n} \left\| (x_k - \mu_n) - VV^T (x_k - \mu_n) \right\|_2^2 . \tag{11}$$

Note that

$$\begin{aligned} \left\| (x_k - \mu_n) - VV^T (x_k - \mu_n) \right\|_2^2 &= (x_k - \mu_n)^T (x_k - \mu_n) \\ &\quad -2 (x_k - \mu_n)^T VV^T (x_k - \mu_n) \\ &\quad + (x_k - \mu_n)^T V \left(V^T V \right) V^T (x_k - \mu_n) \\ &= (x_k - \mu_n)^T (x_k - \mu_n) \\ &\quad - (x_k - \mu_n)^T VV^T (x_k - \mu_n) . \end{aligned}$$

Since $(x_k - \mu_n)^T (x_k - \mu_n)$ does not depend on V, minimizing (9) is equivalent to

$$\max_{V^T V = \mathrm{I}} \sum_{k=1}^{n} (x_k - \mu_n)^T VV^T (x_k - \mu_n) . \tag{12}$$

A few more simple algebraic manipulations using properties of the trace:

$$\begin{aligned} \sum_{k=1}^{n} (x_k - \mu_n)^T VV^T (x_k - \mu_n) &= \sum_{k=1}^{n} \mathrm{Tr} \left[(x_k - \mu_n)^T VV^T (x_k - \mu_n) \right] \\ &= \sum_{k=1}^{n} \mathrm{Tr} \left[V^T (x_k - \mu_n) (x_k - \mu_n)^T V \right] \\ &= \mathrm{Tr} \left[V^T \sum_{k=1}^{n} (x_k - \mu_n) (x_k - \mu_n)^T V \right] \\ &= (n-1) \, \mathrm{Tr} \left[V^T \Sigma_n V \right] . \end{aligned}$$

This means that the solution to (13) is given by

$$\max_{V^TV=\mathrm{I}} \mathrm{Tr}\left[V^T\Sigma_n V\right]. \tag{13}$$

As we saw above (recall (2)) the solution is given by $V = [v_1, \cdots, v_d]$ where $v_1, \ldots, v_d$ correspond to the d leading eigenvectors of Σ_n.

Let us first show that interpretation (2) of finding the d-dimensional projection of $x_1, \ldots, x_n$ that preserves the most variance also arrives to the optimization problem (13).

1.1.2 PCA as d-dimensional projection that preserves the most variance

We aim to find an orthonormal basis $v_1, \ldots, v_d$ (organized as $V = [v_1, \ldots, v_d]$ with $V^TV = \mathrm{I}_{d\times d}$) of a d-dimensional space such that the projection of $x_1, \ldots, x_n$ projected on this subspace has the most variance. Equivalently we can ask for the points

$$\left\{ \begin{bmatrix} v_1^T x_k \\ \vdots \\ v_d^T x_k \end{bmatrix} \right\}_{k=1}^{n},$$

to have as much variance as possible. Hence, we are interested in solving

$$\max_{V^TV=\mathrm{I}} \sum_{k=1}^{n} \left\| V^T x_k - \frac{1}{n}\sum_{r=1}^{n} V^T x_r \right\|^2. \tag{14}$$

Note that

$$\sum_{k=1}^{n} \left\| V^T x_k - \frac{1}{n}\sum_{r=1}^{n} V^T x_r \right\|^2 = \sum_{k=1}^{n} \left\| V^T\left(x_k - \mu_n\right)\right\|^2 = \mathrm{Tr}\left(V^T\Sigma_n V\right),$$

showing that (14) is equivalent to (13) and that the two interpretations of PCA are indeed equivalent.

1.1.3 Finding the Principal Components

When given a dataset $x_1, \ldots, x_n \in \mathbb{R}^p$, in order to compute the Principal Components one needs to find the leading eigenvectors of

$$\Sigma_n = \frac{1}{n-1}\sum_{k=1}^{n}\left(x_k - \mu_n\right)\left(x_k - \mu_n\right)^T.$$

A naive way of doing this would be to construct Σ_n (which takes $\mathcal{O}(np^2)$ work) and then finding its spectral decomposition (which takes $\mathcal{O}(p^3)$ work). This means that the computational complexity of this procedure is $\mathcal{O}\left(\max\left\{np^2, p^3\right\}\right)$ (see [HJ85] and/or [Gol96]).

An alternative is to use the Singular Value Decomposition (1). Let $X = [x_1 \cdots x_n]$ recall that,

$$\Sigma_n = \frac{1}{n}\left(X - \mu_n \mathbf{1}^T\right)\left(X - \mu_n \mathbf{1}^T\right)^T.$$

Let us take the SVD of $X - \mu_n \mathbf{1}^T = U_L D U_R^T$ with $U_L \in O(p)$, D diagonal, and $U_R^T U_R = \mathrm{I}$. Then,

$$\Sigma_n = \frac{1}{n}\left(X - \mu_n \mathbf{1}^T\right)\left(X - \mu_n \mathbf{1}^T\right)^T = U_L D U_R^T U_R D U_L^T = U_L D^2 U_L^T,$$

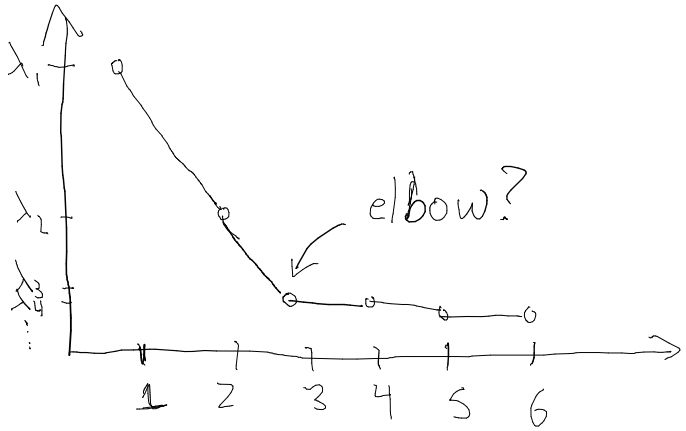
meaning that U_L correspond to the eigenvectors of Σ_n. Computing the SVD of $X - \mu_n \mathbf{1}^T$ takes $\mathcal{O}(\min n^2 p, p^2 n)$ but if one is interested in simply computing the top d eigenvectors then this computational costs reduces to $\mathcal{O}(dnp)$. This can be further improved with randomized algorithms. There are randomized algorithms that compute an approximate solution in $\mathcal{O}\left(pn \log d + (p+n)d^2\right)$ time (see for example [HMT09, RST09, MM15]).[1]

1.1.4 Which d should we pick?

Given a dataset, if the objective is to visualize it then picking $d = 2$ or $d = 3$ might make the most sense. However, PCA is useful for many other purposes, for example: (1) often times the data belongs to a lower dimensional space but is corrupted by high dimensional noise. When using PCA it is oftentimess possible to reduce the noise while keeping the signal. (2) One may be interested in running an algorithm that would be too computationally expensive to run in high dimensions, dimension reduction may help there, etc. In these applications (and many others) it is not clear how to pick d.

If we denote the k-th largest eigenvalue of Σ_n as $\lambda_k^{(+)}(\Sigma_n)$, then the k-th principal component has a $\frac{\lambda_k^{(+)}(\Sigma_n)}{\mathrm{Tr}(\Sigma_n)}$ proportion of the variance. [2]

A fairly popular heuristic is to try to choose the cut-off at a component that has significantly more variance than the one immediately after. This is usually visualized by a scree plot: a plot of the values of the ordered eigenvalues. Here is an example:



[1] If there is time, we might discuss some of these methods later in the course.

[2] Note that $\mathrm{Tr}(\Sigma_n) = \sum_{k=1}^{p} \lambda_k(\Sigma_n)$.

It is common to then try to identify an "elbow" on the scree plot to choose the cut-off. In the next Section we will look into random matrix theory to try to understand better the behavior of the eigenvalues of Σ_n and it will help us understand when to cut-off.

1.1.5 A related open problem

We now show an interesting open problem posed by Mallat and Zeitouni at [MZ11]

Open Problem 1.1 (Mallat and Zeitouni [MZ11]) *Let $g \sim \mathcal{N}(0, \Sigma)$ be a gaussian random vector in $\mathbb{R}^p$ with a known covariance matrix Σ and $d < p$. Now, for any orthonormal basis $V = [v_1, \ldots, v_p]$ of $\mathbb{R}^p$, consider the following random variable Γ_V: Given a draw of the random vector g, Γ_V is the squared ℓ_2 norm of the largest projection of g on a subspace generated by d elements of the basis V. The question is:*

What is the basis V for which $\mathbb{E}[\Gamma_V]$ is maximized?

The conjecture in [MZ11] is that the optimal basis is the eigendecomposition of Σ. It is known that this is the case for $d = 1$ (see [MZ11]) but the question remains open for $d > 1$. It is not very difficult to see that one can assume, without loss of generality, that Σ is diagonal.

A particularly intuitive way of stating the problem is:

1. Given $\Sigma \in \mathbb{R}^{p \times p}$ and d
2. Pick an orthonormal basis $v_1, \ldots, v_p$
3. Given $g \sim \mathcal{N}(0, \Sigma)$
4. Pick d elements $\tilde{v}_1, \ldots, \tilde{v}_d$ of the basis
5. `Score`: $\sum_{i=1}^{d} \left(\tilde{v}_i^T g\right)^2$

The objective is to pick the basis in order to maximize the expected value of the `Score`.

Notice that if the steps of the procedure were taken in a slightly different order on which step 4 would take place before having access to the draw of g (step 3) then the best basis is indeed the eigenbasis of Σ and the best subset of the basis is simply the leading eigenvectors (notice the resemblance with PCA, as described above).

More formally, we can write the problem as finding

$$\underset{\substack{V \in \mathbb{R}^{p \times p} \\ V^T V = \mathrm{I}}}{\operatorname{argmax}} \left(\mathbb{E} \left[\max_{\substack{S \subset [p] \\ |S| = d}} \sum_{i \in S} \left(v_i^T g\right)^2 \right] \right),$$

where $g \sim \mathcal{N}(0, \Sigma)$. The observation regarding the different ordering of the steps amounts to saying that the eigenbasis of Σ is the optimal solution for

$$\underset{\substack{V \in \mathbb{R}^{p \times p} \\ V^T V = \mathrm{I}}}{\operatorname{argmax}} \left(\max_{\substack{S \subset [p] \\ |S| = d}} \mathbb{E} \left[\sum_{i \in S} \left(v_i^T g\right)^2 \right] \right).$$

1.2 PCA in high dimensions and Marcenko-Pastur

Let us assume that the data points $x_1, \ldots, x_n \in \mathbb{R}^p$ are independent draws of a gaussian random variable $g \sim \mathcal{N}(0, \Sigma)$ for some covariance $\Sigma \in \mathbb{R}^{p \times p}$. In this case when we use PCA we are hoping to find low dimensional structure in the distribution, which should correspond to large eigenvalues of Σ (and their corresponding eigenvectors). For this reason (and since PCA depends on the spectral properties of Σ_n) we would like to understand whether the spectral properties of Σ_n (eigenvalues and eigenvectors) are close to the ones of Σ.

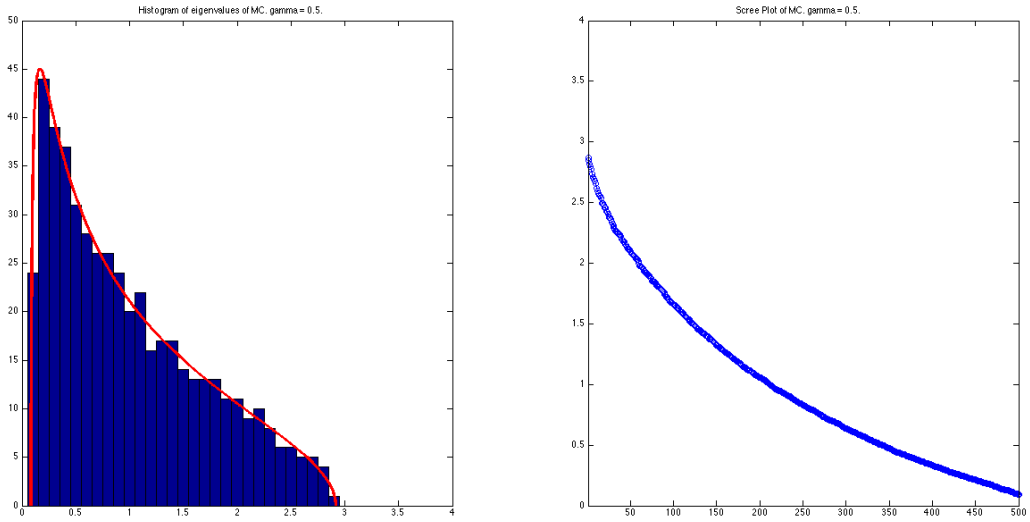
Since $\mathbb{E}\Sigma_n = \Sigma$, if p is fixed and $n \to \infty$ the law of large numbers guarantees that indeed $\Sigma_n \to \Sigma$. However, in many modern applications it is not uncommon to have p in the order of n (or, sometimes, even larger!). For example, if our dataset is composed by images then n is the number of images and p the number of pixels per image; it is conceivable that the number of pixels be on the order of the number of images in a set. Unfortunately, in that case, it is no longer clear that $\Sigma_n \to \Sigma$. Dealing with this type of difficulties is the realm of high dimensional statistics.

For simplicity we will instead try to understand the spectral properties of

$$S_n = \frac{1}{n} X X^T.$$

Since $x \sim \mathcal{N}(0, \Sigma)$ we know that $\mu_n \to 0$ (and, clearly, $\frac{n}{n-1} \to 1$) the spectral properties of S_n will be essentially the same as Σ_n.[3]

Let us start by looking into a simple example, $\Sigma = \mathrm{I}$. In that case, the distribution has no low dimensional structure, as the distribution is rotation invariant. The following is a histogram (left) and a scree plot of the eigenvalues of a sample of S_n (when $\Sigma = \mathrm{I}$) for $p = 500$ and $n = 1000$. The red line is the eigenvalue distribution predicted by the Marchenko-Pastur distribution (15), that we will discuss below.



[3] In this case, S_n is actually the Maximum likelihood estimator for Σ, we'll talk about Maximum likelihood estimation later in the course.

As one can see in the image, there are many eigenvalues considerably larger than 1 (and some considerably larger than others). Notice that , if given this profile of eigenvalues of Σ_n one could potentially be led to believe that the data has low dimensional structure, when in truth the distribution it was drawn from is isotropic.

Understanding the distribution of eigenvalues of random matrices is in the core of Random Matrix Theory (there are many good books on Random Matrix Theory, e.g. [Tao12] and [AGZ10]). This particular limiting distribution was first established in 1967 by Marchenko and Pastur [MP67] and is now referred to as the Marchenko-Pastur distribution. They showed that, if p and n are both going to ∞ with their ratio fixed $p/n = \gamma \leq 1$, the sample distribution of the eigenvalues of S_n (like the histogram above), in the limit, will be

$$dF_\gamma(\lambda) = \frac{1}{2\pi}\frac{\sqrt{(\gamma_+ - \lambda)(\lambda - \gamma_-)}}{\gamma\lambda}1_{[\gamma_-,\gamma_+]}(\lambda)d\lambda, \tag{15}$$

with support $[\gamma_-, \gamma_+]$. This is plotted as the red line in the figure above.

Remark 1.4 *We will not show the proof of the Marchenko-Pastur Theorem here (you can see, for example, [Bai99] for several different proofs of it), but an approach to a proof is using the so-called moment method. The core of the idea is to note that one can compute moments of the eigenvalue distribution in two ways and note that (in the limit) for any k,*

$$\frac{1}{p}\mathbb{E}\operatorname{Tr}\left[\left(\frac{1}{n}XX^T\right)^k\right] = \frac{1}{p}\mathbb{E}\operatorname{Tr}\left(S_n^k\right) = \mathbb{E}\frac{1}{p}\sum_{i=1}^{p}\lambda_i^k(S_n) = \int_{\gamma_-}^{\gamma_+}\lambda^k dF_\gamma(\lambda),$$

and that the quantities $\frac{1}{p}\mathbb{E}\operatorname{Tr}\left[\left(\frac{1}{n}XX^T\right)^k\right]$ can be estimated (these estimates rely essentially in combinatorics). The distribution $dF_\gamma(\lambda)$ can then be computed from its moments.

1.2.1 A related open problem

Open Problem 1.2 (Monotonicity of singular values [BKS13]) *Consider the setting above but with $p = n$, then $X \in \mathbb{R}^{n\times n}$ is a matrix with iid $\mathcal{N}(0,1)$ entries. Let*

$$\sigma_i\left(\frac{1}{\sqrt{n}}X\right),$$

denote the i-th singular value[4] of $\frac{1}{\sqrt{n}}X$, and define

$$\alpha_{\mathbb{R}}(n) := \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\sigma_i\left(\frac{1}{\sqrt{n}}X\right)\right],$$

as the expected value of the average singular value of $\frac{1}{\sqrt{n}}X$.

The conjecture is that, for every $n \geq 1$,

$$\alpha_{\mathbb{R}}(n+1) \geq \alpha_{\mathbb{R}}(n).$$

[4]The i-th diagonal element of Σ in the SVD $\frac{1}{\sqrt{n}}X = U\Sigma V$.

Moreover, for the analogous quantity $\alpha_{\mathbb{C}}(n)$ defined over the complex numbers, meaning simply that each entry of X is an iid complex valued standard gaussian $\mathbb{C}\mathcal{N}(0,1)$ the reverse inequality is conjectured for all $n \geq 1$:

$$\alpha_{\mathbb{C}}(n+1) \leq \alpha_{\mathbb{C}}(n).$$

Notice that the singular values of $\frac{1}{\sqrt{n}}X$ are simply the square roots of the eigenvalues of S_n,

$$\sigma_i\left(\frac{1}{\sqrt{n}}X\right) = \sqrt{\lambda_i\left(S_n\right)}.$$

This means that we can compute $\alpha_{\mathbb{R}}$ in the limit (since we know the limiting distribution of $\lambda_i\left(S_n\right)$) and get (since $p = n$ we have $\gamma = 1$, $\gamma_- = 0$, and $\gamma_+ = 2$)

$$\lim_{n\to\infty} \alpha_{\mathbb{R}}(n) = \int_0^2 \lambda^{\frac{1}{2}} dF_1(\lambda) = \frac{1}{2\pi}\int_0^2 \lambda^{\frac{1}{2}} \frac{\sqrt{(2-\lambda)\,\lambda}}{\lambda} = \frac{8}{3\pi} \approx 0.8488.$$

Also, $\alpha_{\mathbb{R}}(1)$ simply corresponds to the expected value of the absolute value of a standard gaussian g

$$\alpha_{\mathbb{R}}(1) = \mathbb{E}|g| = \sqrt{\frac{2}{\pi}} \approx 0.7990,$$

which is compatible with the conjecture.

On the complex valued side, the Marchenko-Pastur distribution also holds for the complex valued case and so $\lim_{n\to\infty} \alpha_{\mathbb{C}}(n) = \lim_{n\to\infty} \alpha_{\mathbb{R}}(n)$ and $\alpha_{\mathbb{C}}(1)$ can also be easily calculated and seen to be larger than the limit.

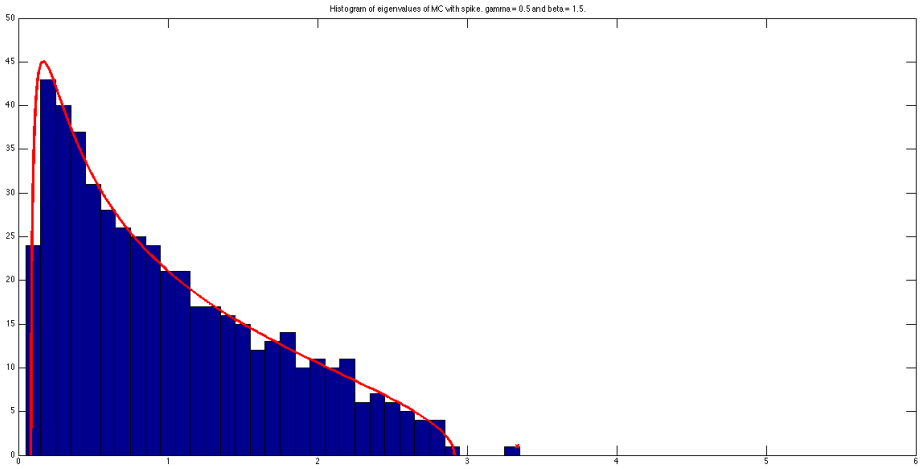
2 Spike Models and BPP transition

What if there actually is some (linear) low dimensional structure on the data? When can we expect to capture it with PCA? A particularly simple, yet relevant, example to analyse is when the covariance matrix Σ is an identity with a rank 1 perturbation, which we refer to as a spike model $\Sigma = I + \beta v v^T$, for v a unit norm vector and $\beta \geq 0$.

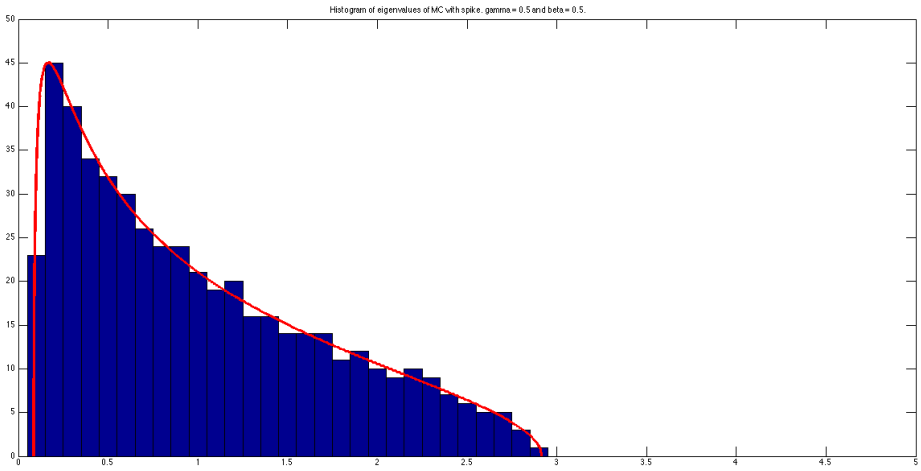
One way to think about this instance is as each data point x consisting of a signal part $\sqrt{\beta} g_0 v$ where g_0 is a one-dimensional standard gaussian (a gaussian multiple of a fixed vector $\sqrt{\beta} v$ and a noise part $g \sim \mathcal{N}(0, I)$ (independent of g_0. Then $x = g + \sqrt{\beta} g_0 v$ is a gaussian random variable

$$x \sim \mathcal{N}(0, I + \beta v v^T).$$

A natural question is whether this rank 1 perturbation can be seen in S_n. Let us build some intuition with an example, the following is the histogram of the eigenvalues of a sample of S_n for $p = 500$, $n = 1000$, v is the first element of the canonical basis $v = e_1$, and $\beta = 1.5$:



The images suggests that there is an eigenvalue of S_n that "pops out" of the support of the Marchenko-Pastur distribution (below we will estimate the location of this eigenvalue, and that estimate corresponds to the red "x"). It is worth noticing that the largest eigenvalues of Σ is simply $1 + \beta = 2.5$ while the largest eigenvalue of S_n appears considerably larger than that. Let us try now the same experiment for $\beta = 0.5$:



and it appears that, for $\beta = 0.5$, the distribution of the eigenvalues appears to be undistinguishable from when $\Sigma = \mathrm{I}$.

This motivates the following question:

Question 2.1 *For which values of γ and β do we expect to see an eigenvalue of S_n popping out of the support of the Marchenko-Pastur distribution, and what is the limit value that we expect it to take?*

As we will see below, there is a critical value of β below which we don't expect to see a change in the distribution of eivenalues and above which we expect one of the eigenvalues to pop out of the support, this is known as BPP transition (after Baik, Ben Arous, and Péché [BBAP05]). There are many very nice papers about this phenomenon, including [Pau, Joh01, BBAP05, Pau07, BS05, Kar05].[5]

In what follows we will find the critical value of β and estimate the location of the largest eigenvalue of S_n. While the argument we will use can be made precise (and is borrowed from [Pau]) we will be ignoring a few details for the sake of exposition. **In short, the argument below can be transformed into a rigorous proof, but it is not one at the present form!**

First of all, it is not difficult to see that we can assume that $v = e_1$ (since everything else is rotation invariant). We want to understand the behavior of the leading eigenvalue of

$$S_n = \frac{1}{n}\sum_{i=1}^{n} x_i x_i^T = \frac{1}{n} X X^T,$$

where

$$X = [x_1, \ldots, x_n] \in \mathbb{R}^{p\times n}.$$

We can write X as

$$X = \begin{bmatrix} \sqrt{1+\beta} Z_1^T \\ Z_2^T \end{bmatrix},$$

where $Z_1 \in \mathbb{R}^{n\times 1}$ and $Z_2 \in \mathbb{R}^{n\times(p-1)}$, both populated with i.i.d. standard gaussian entries ($\mathcal{N}(0,1)$). Then,

$$S_n = \frac{1}{n} X X^T = \frac{1}{n}\begin{bmatrix} (1+\beta) Z_1^T Z_1 & \sqrt{1+\beta} Z_1^T Z_2 \\ \sqrt{1+\beta} Z_2^T Z_1 & Z_2^T Z_2 \end{bmatrix}.$$

Now, let $\hat{\lambda}$ and $v = \begin{bmatrix} v_1 \\ v_2 \end{bmatrix}$ where $v_2 \in \mathbb{R}^{p-1}$ and $v_1 \in \mathbb{R}$, denote, respectively, an eigenvalue and associated eigenvector for S_n. By the definition of eigenvalue and eigenvector we have

$$\frac{1}{n}\begin{bmatrix} (1+\beta) Z_1^T Z_1 & \sqrt{1+\beta} Z_1^T Z_2 \\ \sqrt{1+\beta} Z_2^T Z_1 & Z_2^T Z_2 \end{bmatrix}\begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = \hat{\lambda}\begin{bmatrix} v_1 \\ v_2 \end{bmatrix},$$

which can be rewritten as

$$\frac{1}{n}(1+\beta) Z_1^T Z_1 v_1 + \frac{1}{n}\sqrt{1+\beta} Z_1^T Z_2 v_2 = \hat{\lambda} v_1 \quad (16)$$

$$\frac{1}{n}\sqrt{1+\beta} Z_2^T Z_1 v_1 + \frac{1}{n} Z_2^T Z_2 v_2 = \hat{\lambda} v_2. \quad (17)$$

(17) is equivalent to

$$\frac{1}{n}\sqrt{1+\beta} Z_2^T Z_1 v_1 = \left(\hat{\lambda}\,\mathrm{I} - \frac{1}{n} Z_2^T Z_2\right) v_2.$$

[5] Notice that the Marchenko-Pastur theorem does not imply that all eigenvalues are actually in the support of the Marchenk-Pastur distribution, it just rules out that a non-vanishing proportion are. However, it is possible to show that indeed, in the limit, all eigenvalues will be in the support (see, for example, [Pau]).

If $\hat{\lambda}\,\mathrm{I} - \frac{1}{n} Z_2^T Z_2$ is invertible (this won't be justified here, but it is in [Pau]) then we can rewrite it as

$$v_2 = \left(\hat{\lambda}\,\mathrm{I} - \frac{1}{n} Z_2^T Z_2\right)^{-1} \frac{1}{n}\sqrt{1+\beta} Z_2^T Z_1 v_1,$$

which we can then plug in (16) to get

$$\frac{1}{n}(1+\beta) Z_1^T Z_1 v_1 + \frac{1}{n}\sqrt{1+\beta} Z_1^T Z_2 \left(\hat{\lambda}\,\mathrm{I} - \frac{1}{n} Z_2^T Z_2\right)^{-1} \frac{1}{n}\sqrt{1+\beta} Z_2^T Z_1 v_1 = \hat{\lambda} v_1$$

If $v_1 \neq 0$ (again, not properly justified here, see [Pau]) then this means that

$$\hat{\lambda} = \frac{1}{n}(1+\beta) Z_1^T Z_1 + \frac{1}{n}\sqrt{1+\beta} Z_1^T Z_2 \left(\hat{\lambda}\,\mathrm{I} - \frac{1}{n} Z_2^T Z_2\right)^{-1} \frac{1}{n}\sqrt{1+\beta} Z_2^T Z_1 \tag{18}$$

First observation is that because $Z_1 \in \mathbb{R}^n$ has standard gaussian entries then $\frac{1}{n} Z_1^T Z_1 \to 1$, meaning that

$$\hat{\lambda} = (1+\beta)\left[1 + \frac{1}{n} Z_1^T Z_2 \left(\hat{\lambda}\,\mathrm{I} - \frac{1}{n} Z_2^T Z_2\right)^{-1} \frac{1}{n} Z_2^T Z_1\right]. \tag{19}$$

Consider the SVD of $Z_2 = U \Sigma V^T$ where $U \in \mathbb{R}^{n\times p}$ and $V \in \mathbb{R}^{p\times p}$ have orthonormal columns (meaning that $U^T U = \mathrm{I}_{p\times p}$ and $V^T V = \mathrm{I}_{p\times p}$), and Σ is a diagonal matrix. Take $D = \frac{1}{n}\Sigma^2$ then

$$\frac{1}{n} Z_2^T Z_2 = \frac{1}{n} V \Sigma^2 V^T = V D V^T,$$

meaning that the diagonal entries of D correspond to the eigenvalues of $\frac{1}{n} Z_2^T Z_2$ which we expect to be distributed (in the limit) according to the Marchenko-Pastur distribution for $\frac{p-1}{n} \approx \gamma$. Replacing back in (19)

$$\begin{aligned}
\hat{\lambda} &= (1+\beta)\left[1 + \frac{1}{n} Z_1^T \left(\sqrt{n} U D^{1/2} V^T\right)\left(\hat{\lambda}\,\mathrm{I} - V D V^T\right)^{-1} \frac{1}{n}\left(\sqrt{n} U D^{1/2} V^T\right)^T Z_1\right] \\
&= (1+\beta)\left[1 + \frac{1}{n}\left(U^T Z_1\right)^T D^{1/2} V^T \left(\hat{\lambda}\,\mathrm{I} - V D V^T\right)^{-1} V D^{1/2}\left(U^T Z_1\right)\right] \\
&= (1+\beta)\left[1 + \frac{1}{n}\left(U^T Z_1\right)^T D^{1/2} V^T \left(V\left[\hat{\lambda}\,\mathrm{I} - D\right] V^T\right)^{-1} V D^{1/2}\left(U^T Z_1\right)\right] \\
&= (1+\beta)\left[1 + \frac{1}{n}\left(U^T Z_1\right)^T D^{1/2} \left(\left[\hat{\lambda}\,\mathrm{I} - D\right]\right)^{-1} D^{1/2}\left(U^T Z_1\right)\right].
\end{aligned}$$

Since the columns of U are orthonormal, $g := U^T Z_1 \in \mathbb{R}^{p-1}$ is an isotropic gaussian ($g \sim \mathcal{N}(0,1)$), in fact,

$$\mathbb{E} g g^T = \mathbb{E} U^T Z_1 \left(U^T Z_1\right)^T = \mathbb{E} U^T Z_1 Z_1^T U = U^T \mathbb{E}\left[Z_1 Z_1^T\right] U = U^T U = \mathrm{I}_{(p-1)\times(p-1)}.$$

We proceed

$$\begin{aligned}
\hat{\lambda} &= (1+\beta)\left[1 + \frac{1}{n} g^T D^{1/2}\left(\left[\hat{\lambda}\,\mathrm{I} - D\right]\right)^{-1} D^{1/2} g\right] \\
&= (1+\beta)\left[1 + \frac{1}{n}\sum_{j=1}^{p-1} g_j^2 \frac{D_{jj}}{\hat{\lambda} - D_{jj}}\right]
\end{aligned}$$

Because we expect the diagonal entries of D to be distributed according to the Marchenko-Pastur distribution and g to be independent to it we expect that (again, not properly justified here, see [Pau])

$$\frac{1}{p-1}\sum_{j=1}^{p-1} g_j^2 \frac{D_{jj}}{\hat{\lambda}-D_{jj}} \to \int_{\gamma_-}^{\gamma_+} \frac{x}{\hat{\lambda}-x} dF_\gamma(x).$$

We thus get an equation for $\hat{\lambda}$:

$$\hat{\lambda} = (1+\beta)\left[1+\gamma\int_{\gamma_-}^{\gamma_+} \frac{x}{\hat{\lambda}-x} dF_\gamma(x)\right],$$

which can be easily solved with the help of a program that computes integrals symbolically (such as Mathematica) to give (you can also see [Pau] for a derivation):

$$\hat{\lambda} = (1+\beta)\left(1+\frac{\gamma}{\beta}\right), \tag{20}$$

which is particularly elegant (specially considering the size of some the equations used in the derivation).

An important thing to notice is that for $\beta = \sqrt{\gamma}$ we have

$$\hat{\lambda} = (1+\sqrt{\gamma})\left(1+\frac{\gamma}{\sqrt{\gamma}}\right) = (1+\sqrt{\gamma})^2 = \gamma_+,$$

suggesting that $\beta = \sqrt{\gamma}$ is the critical point.

Indeed this is the case and it is possible to make the above argument rigorous[6] and show that in the model described above,

- If $\beta \leq \sqrt{\gamma}$ then
$$\lambda_{\max}(S_n) \to \gamma_+,$$
- and if $\beta > \sqrt{\gamma}$ then
$$\lambda_{\max}(S_n) \to (1+\beta)\left(1+\frac{\gamma}{\beta}\right) > \gamma_+.$$

Another important question is wether the leading eigenvector actually correlates with the planted perturbation (in this case e_1). Turns out that very similar techniques can answer this question as well [Pau] and show that the leading eigenvector $v_{\max}$ of S_n will be non-trivially correlated with e_1 if and only if $\beta > \sqrt{\gamma}$, more precisely:

- If $\beta \leq \sqrt{\gamma}$ then
$$|\langle v_{\max}, e_1\rangle|^2 \to 0,$$
- and if $\beta > \sqrt{\gamma}$ then
$$|\langle v_{\max}, e_1\rangle|^2 \to \frac{1-\frac{\gamma}{\beta^2}}{1-\frac{\gamma}{\beta}}.$$

[6] Note that in the argument above it wasn't even completely clear where it was used that the eigenvalue was actually the leading one. In the actual proof one first needs to make sure that there is an eigenvalue outside of the support and the proof only holds for that one, you can see [Pau]

2.0.2 A brief mention of Wigner matrices

Another very important random matrix model is the Wigner matrix (and it will show up later in this course). Given an integer n, a standard gaussian Wigner matrix $W \in \mathbb{R}^{n\times n}$ is a symmetric matrix with independent $\mathcal{N}(0,1)$ entries (except for the fact that $W_{ij} = W_{ji}$). In the limit, the eigenvalues of $\frac{1}{\sqrt{n}}W$ are distributed according to the so-called semi-circular law

$$dSC(x) = \frac{1}{2\pi}\sqrt{4-x^2}1_{[-2,2]}(x)dx,$$

and there is also a BPP like transition for this matrix ensemble [FP06]. More precisely, if v is a unit-norm vector in $\mathbb{R}^n$ and $\xi \geq 0$ then the largest eigenvalue of $\frac{1}{\sqrt{n}}W + \xi vv^T$ satisfies

- If $\xi \leq 1$ then
$$\lambda_{\max}\left(\frac{1}{\sqrt{n}}W + \xi vv^T\right) \to 2,$$
- and if $\xi > 1$ then
$$\lambda_{\max}\left(\frac{1}{\sqrt{n}}W + \xi vv^T\right) \to \xi + \frac{1}{\xi}. \tag{21}$$

2.0.3 An open problem about spike models

Open Problem 2.1 (Spike Model for cut–SDP [MS15]) *Let W denote a symmetric Wigner matrix with i.i.d. entries $W_{ij} \sim \mathcal{N}(0,1)$. Also, given $B \in \mathbb{R}^{n\times n}$ symmetric, define:*

$$Q(B) = \max\left\{\mathrm{Tr}(BX) : X \succeq 0,\ X_{ii} = 1\right\}.$$

Define $q(\xi)$ as

$$q(\xi) = \lim_{n\to\infty}\frac{1}{n}\mathbb{E}Q\left(\frac{\xi}{n}\mathbf{1}\mathbf{1}^T + \frac{1}{\sqrt{n}}W\right).$$

What is the value of ξ_*, defined as

$$\xi_* = \inf\{\xi \geq 0 : q(\xi) > 2\}.$$

It is known that, if $0 \leq \xi \leq 1$, $q(\xi) = 2$ [MS15].
One can show that $\frac{1}{n}Q(B) \leq \lambda_{\max}(B)$. In fact,

$$\max\left\{\mathrm{Tr}(BX) : X \succeq 0,\ X_{ii} = 1\right\} \leq \max\left\{\mathrm{Tr}(BX) : X \succeq 0,\ \mathrm{Tr}\, X = n\right\}.$$

It is also not difficult to show (hint: take the spectral decomposition of X) that

$$\max\left\{\mathrm{Tr}(BX) : X \succeq 0,\ \sum_{i=1}^n X_{ii} = n\right\} = \lambda_{\max}(B).$$

This means that for $\xi > 1$, $q(\xi) \leq \xi + \frac{1}{\xi}$.

Remark 2.1 *Optimization problems of the type of* $\max\{\mathrm{Tr}(BX) : X \succeq 0,\ X_{ii} = 1\}$ *are semidefinite programs, they will be a major player later in the course!*

Since $\frac{1}{n}\mathbb{E}\,\mathrm{Tr}\left[11^T\left(\frac{\xi}{n}\mathbf{1}\mathbf{1}^T + \frac{1}{\sqrt{n}}W\right)\right] \approx \xi$, by taking $X = 11^T$ we expect that $q(\xi) \geq \xi$.

These observations imply that $1 \leq \xi_* < 2$ (see [MS15]). A reasonable conjecture is that it is equal to 1. This would imply that a certain semidefinite programming based algorithm for clustering under the Stochastic Block Model on 2 clusters (we will discuss these things later in the course) is optimal for detection (see [MS15]).[7]

References

[ABH14] E. Abbe, A. S. Bandeira, and G. Hall. Exact recovery in the stochastic block model. *Available online at arXiv:1405.3267 [cs.SI]*, 2014.

[AGZ10] G. W. Anderson, A. Guionnet, and O. Zeitouni. *An introduction to random matrices*. Cambridge studies in advanced mathematics. Cambridge University Press, Cambridge, New York, Melbourne, 2010.

[Bai99] Z. D. Bai. Methodologies in spectral analysis of large dimensional random matrices, a review. *Statistics Sinica*, 9:611–677, 1999.

[Ban15a] A. S. Bandeira. Random Laplacian matrices and convex relaxations. *Available online at arXiv:1504.03987 [math.PR]*, 2015.

[Ban15b] A. S. Bandeira. Relax and Conquer BLOG: 18.S096 Principal Component Analysis in High Dimensions and the Spike Model. 2015.

[BBAP05] J. Baik, G. Ben-Arous, and S. Péché. Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices. *The Annals of Probability*, 33(5):1643–1697, 2005.

[BKS13] A. S. Bandeira, C. Kennedy, and A. Singer. Approximating the little grothendieck problem over the orthogonal group. *Available online at arXiv:1308.5207 [cs.DS]*, 2013.

[BS05] J. Baik and J. W. Silverstein. Eigenvalues of large sample covariance matrices of spiked population models. 2005.

[FP06] D. Féral and S. Péché. The largest eigenvalue of rank one deformation of large wigner matrices. *Communications in Mathematical Physics*, 272(1):185–228, 2006.

[Gol96] G. H. Golub. *Matrix Computations*. Johns Hopkins University Press, third edition, 1996.

[HJ85] R. A. Horn and C. R. Johnson. *Matrix Analysis*. Cambridge University Press, 1985.

[HMT09] N. Halko, P. G. Martinsson, and J. A. Tropp. Finding structure with randomness: probabilistic algorithms for constructing approximate matrix decompositions. *Available online at arXiv:0909.4061v2 [math.NA]*, 2009.

[7] Later in the course we will discuss clustering under the Stochastic Block Model quite thoroughly, and will see how this same SDP is known to be optimal for exact recovery [ABH14, HWX14, Ban15a].

[HWX14] B. Hajek, Y. Wu, and J. Xu. Achieving exact cluster recovery threshold via semidefinite programming. *Available online at arXiv:1412.6156*, 2014.

[Joh01] I. M. Johnston. On the distribution of the largest eigenvalue in principal components analysis. *The Annals of Statistics*, 29(2):295–327, 2001.

[Kar05] N. E. Karoui. Recent results about the largest eigenvalue of random covariance matrices and statistical application. *Acta Physica Polonica B*, 36(9), 2005.

[MM15] C. Musco and C. Musco. Stronger and faster approximate singular value decomposition via the block lanczos method. *Available at arXiv:1504.05477 [cs.DS]*, 2015.

[Mos11] M. S. Moslehian. Ky Fan inequalities. *Available online at arXiv:1108.1467 [math.FA]*, 2011.

[MP67] V. A. Marchenko and L. A. Pastur. Distribution of eigenvalues in certain sets of random matrices. *Mat. Sb. (N.S.)*, 72(114):507–536, 1967.

[MS15] A. Montanari and S. Sen. Semidefinite programs on sparse random graphs. *Available online at arXiv:1504.05910 [cs.DM]*, 2015.

[MZ11] S. Mallat and O. Zeitouni. A conjecture concerning optimality of the karhunen-loeve basis in nonlinear reconstruction. *Available online at arXiv:1109.0489 [math.PR]*, 2011.

[Pau] D. Paul. Asymptotics of the leading sample eigenvalues for a spiked covariance model. *Available online at* `http://anson.ucdavis.edu/~debashis/techrep/eigenlimit.pdf`.

[Pau07] D. Paul. Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. *Statistics Sinica*, 17:1617–1642, 2007.

[Pea01] K. Pearson. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine, Series 6*, 2(11):559–572, 1901.

[RST09] V. Rokhlin, A. Szlam, and M. Tygert. A randomized algorithm for principal component analysis. *Available at arXiv:0809.2274 [stat.CO]*, 2009.

[Tao12] T. Tao. *Topics in Random Matrix Theory*. Graduate studies in mathematics. American Mathematical Soc., 2012.