Chapter 3
Singular Value Decomposition and Principal Component Analysis

3.1 Brief review of some linear algebra tools

In this Section we will briefly review a few basic linear algebra tools that will be important throughout the book. If you need a refresh on any of these concepts, we recommend taking a look at [81] and/or [72].

3.1.1 Singular Value Decomposition

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

$$M = U\Sigma V^T, \tag{3.1}$$

where $U \in O(m)$, $V \in O(n)$ are orthogonal matrices (meaning that $U^TU = UU^T = \mathrm{I}_{m\times m}$ and $V^TV = VV^T = \mathrm{I}_{n\times n}$) and $\Sigma \in \mathbb{R}^{m\times n}$ is a matrix with non-negative entries on 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. Through the SVD, any matrix can be written as a sum of rank-1 matrices

$$M = \sum_{k=1}^{r} \sigma_k u_k v_k^T, \tag{3.2}$$

where $\sigma_1 \geq \sigma_2 \geq \sigma_r > 0$ are the non-zero singular values of M, and u_k and v_k are the corresponding left and right singular vectors. In particular, $\mathrm{rank}(M) = r$, that is, the number of non-zero singular values r is the rank of M.

Remark 3.1. 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 $V \in O(n)$.

3.1.2 Matrix norms and low rank matrix approximation

Just like with vectors, the size of a matrix can be measured using a suitable norm. One popular norm is the Frobenius norm, or the Hilbert-Schmidt norm, defined as

$$\|M\|_F = \sqrt{\sum_{i,j} M_{ij}^2}, \tag{3.3}$$

which is simply the Euclidean norm of a vector of length mn of the matrix elements. The Frobenius norm can also be expressed in terms of the singular values. To see this, first express the Frobenius norm in terms of the trace of $M^T M$ as

$$\|M\|_F^2 = \sum_{i,j} M_{ij}^2 = \mathrm{Tr}(M^T M), \tag{3.4}$$

where we recall that the trace of a square matrix A is defined as

$$\mathrm{Tr}(A) = \sum_i A_{ii}. \tag{3.5}$$

A particularly important property of the trace is that for any A of size $m \times n$ and B of size $n \times m$

$$\mathrm{Tr}(AB) = \mathrm{Tr}(BA). \tag{3.6}$$

Note that this implies that, e.g., $\mathrm{Tr}(ABC) = \mathrm{Tr}(CAB)$, but it does not imply that, e.g., $\mathrm{Tr}(ABC) = \mathrm{Tr}(ACB)$ which is not true in general. Now, plugging the SVD (3.1) into (3.4) gives

$$\|M\|_F^2 = \mathrm{Tr}(M^T M) = \mathrm{Tr}(V\Sigma^T U^T U \Sigma V^T) = \mathrm{Tr}(\Sigma^T \Sigma) = \sum_{k=1}^{r} \sigma_k^2, \tag{3.7}$$

where we used the orthogonality of U and V and the trace property (3.6). We conclude that the Frobenius norm equals the Euclidean norm of the vector of singular values.

A different way to define the size of a matrix is by viewing it as an operator and measuring by how much it can dilate vectors. For example, the operator 2-norm is defined as

$$\|M\|_2 = \sup_{\|x\|=1} \|Mx\|. \tag{3.8}$$

Again, this operator norm can be succinctly expressed in terms of the singular values. Indeed, for any $x \in \mathbb{R}^n$

$$Mx = \sum_{k=1}^{r} \sigma_k u_k (v_k^T x). \tag{3.9}$$

Using the orthogonality of the left singular vectors u_k we get

$$\|Mx\|^2 = \sum_{k=1}^{r} \sigma_k^2 \langle v_k, x\rangle^2 \leq \sigma_1^2 \sum_{k=1}^{r} \langle v_k, x\rangle^2 \leq \sigma_1^2 \sum_{k=1}^{n} \langle v_k, x\rangle^2 = \sigma_1^2 \|x\|^2, \tag{3.10}$$

where the last equality is due to the orthogonality of the right singular vectors v_k. Moreover, we get equality by choosing $x = v_1$. We conclude that the 2-norm is simply the largest singular value

$$\|M\|_2 = \sigma_1. \tag{3.11}$$

A very important property of the SVD is that it provides the best low rank approximation of a matrix, when the approximation error is measured in terms of the Frobenius norm. Specifically, for any $0 \leq s \leq r$ consider the rank-s matrix $M_s = \sum_{k=1}^{s} \sigma_k u_k v_k^T$. Then, among all matrices of rank s, M_s best approximates M in terms of the Frobenius norm error. Moreover, the approximation error is given in terms of the remaining $r-s$ smallest singular values as

$$\|M - M_s\|_F = \inf_{B \in \mathbb{R}^{m\times n}, \text{rank}(B) \leq s} \|M - B\|_F = \sqrt{\sum_{k=s+1}^{r} \sigma_k^2} \tag{3.12}$$

A similar result holds for the best low rank approximation in the 2-norm

$$\|M - M_s\|_2 = \inf_{B \in \mathbb{R}^{m\times n}, \text{rank}(B) \leq s} \|M - B\|_2 = \sigma_{s+1} \tag{3.13}$$

In fact, M_s is the best low rank approximation for any univariate matrix norm satisfying $\|UMV\| = \|M\|$ for any $U \in O(m), V \in O(n)$, that is, norms that are invariant to multiplication by orthogonal matrices.

The low rank approximation property has a wide ranging implication on data compression. The storage size of an $m \times n$ data matrix is mn. If that matrix is of rank r, then storage size reduces from mn to $(n+m+1)r$ (for storing r left and right singular vectors and values). For $r \ll \min\{n,m\}$ this reduction can be quite dramatic. For example, if $r = 10$ and $n = m = 10^6$, then storage reduces from 10^{12} entries to just $2 \cdot 10^7$. But even if the matrix is not precisely of rank r, but only approximately, in the sense that $\sigma_{r+1} \ll \sigma_1$, then we are guaranteed by the above approximation results to incur only a small approximation due to compression using the top r singular vectors and values. In many cases, the singular values of large data matrices decrease very quickly, motivating this type of low rank approximation which oftentimes is the only way to handle massive data sets that otherwise cannot be stored and/or manipulated efficiently.

Remark 3.2. The computational complexity of computing the SVD of a matrix of size $m \times n$ with $m \geq n$ is $\mathcal{O}(mn^2)$. This cubic scaling could be prohibitive for massive data matrices, and in Chapter **??** we discuss numerical algorithms that use randomization for efficient computation the low rank approximation of such large matrices.

3.1.3 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.

For symmetric matrices, the operator 2-norm is also known as the spectral norm, given by

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

3.1.4 Quadratic Forms

Later on 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

$$\max_{\substack{v_1,\ldots,v_d\in\mathbb{R}^n\\ v_i^Tv_j=\delta_{ij}}} \sum_{k=1}^{d} v_k^TMv_k, \tag{3.14}$$

where δ is the Kronecker delta ($\delta_{ij}=1$ for $i=j$ and $\delta_{ij}=0$ otherwise).

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

$$\max_{\substack{v\in\mathbb{R}^n\\ \|v\|_2=1}} v^TMv. \tag{3.15}$$

It is easy to see (for example, using the spectral decomposition of M) that (3.15) 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 [109]) that (3.14) 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 (3.14) can be computed sequentially: we can first solve for $d = 1$, computing v_1, then v_2, and so on.

Remark 3.3. All of the tools and results above have natural analogues when the matrices have complex entries (and are Hermitian instead of symmetric).

3.2 Principal Component Analysis

3.2.1 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 dimensional space or by finding a better representation for the data using a small number of meaningful features. Beyond data compression and visualization, dimension reduction facilitates downstream analysis such as clustering and regression that perform significantly better in lower dimensions. Throughout this book we will explore a few different ways of reducing the dimension, both linearly and non-linearly.

We will start with the classical Principal Component Analysis (PCA). PCA continues to be one of the most effective and simplest tools for exploratory data analysis. Remarkably, it dates back to a 1901 paper by Karl Pearson [122].

Suppose 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 ($d = 2,3$). There are a couple of seemingly different criteria 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 preserves 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

$$\mu_n = \frac{1}{n} \sum_{k=1}^{n} x_k, \tag{3.16}$$

and its sample covariance as

$$\Sigma_n = \frac{1}{n-1} \sum_{k=1}^{n} (x_k - \mu_n)(x_k - \mu_n)^T. \tag{3.17}$$

Remark 3.4. 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.

3.2.1.1 PCA as the best d-dimensional affine fit

We are trying to approximate each x_k by

$$x_k \approx \mu + \sum_{i=1}^{d} (\beta_k)_i v_i, \tag{3.18}$$

where $v_1, \ldots, v_d$ is an orthonormal basis for the d-dimensional subspace, $\mu \in \mathbb{R}^p$ represents the translation, and $\beta_k \in \mathbb{R}^d$ 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 (3.19) as

$$x_k \approx \mu + V\beta_k, \tag{3.19}$$

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

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

$$\min_{\substack{\mu,\, V,\, \beta_k \\ V^T V = \mathrm{I}}} \sum_{k=1}^{n} \|x_k - (\mu + V\beta_k)\|_2^2 \tag{3.20}$$

We start by optimizing for μ. It is easy to see that the first order condition for μ corresponds 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 we can assume, without loss of generality, that $\sum_{k=1}^{n} \beta_k = 0$, 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 to finding the solution for (3.20) by solving

$$\min_{\substack{V,\ \beta_k \\ V^T V = \mathrm{I}}} \sum_{k=1}^{n} \|x_k - \mu_n - V\beta_k\|_2^2. \tag{3.21}$$

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

$$\min_{\beta_k} \|x_k - \mu_n - V\beta_k\|_2^2 = \min_{\beta_k} \left\| x_k - \mu_n - \sum_{i=1}^{d} (\beta_k)_i v_i \right\|_2^2. \tag{3.22}$$

Since $v_1, \ldots, v_d$ are orthonormal, it is easy to see that the solution is given by $\left(\beta_k^*\right)_i = v_i^T (x_k - \mu_n)$ which can be succinctly written as $\beta_k = V^T (x_k - \mu_n)$. Thus, (3.21) is equivalent to

$$\min_{V^T V = \mathrm{I}} \sum_{k=1}^{n} \left\| (x_k - \mu_n) - VV^T (x_k - \mu_n) \right\|_2^2. \tag{3.23}$$

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 (3.23) is equivalent to

$$\max_{V^T V = \mathrm{I}} \sum_{k=1}^{n} (x_k - \mu_n)^T VV^T (x_k - \mu_n). \tag{3.24}$$

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 (3.24) is given by

$$\max_{V^TV=\mathrm{I}}\mathrm{Tr}\left[V^T\Sigma_n V\right]. \tag{3.25}$$

As we saw above (recall (3.14)) 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 next show that alternative interpretation of PCA as finding the d-dimensional projection of $x_1,\ldots,x_n$ that preserves the most variance also arrives to the optimization problem (3.25).

3.2.1.2 PCA as the 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 $x_1,\ldots,x_n$ onto this subspace has the most variance. Equivalently we can ask for the points

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

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

$$\max_{V^TV=\mathrm{I}}\sum_{k=1}^{n}\left\|V^Tx_k-\frac{1}{n}\sum_{r=1}^{n}V^Tx_r\right\|^2. \tag{3.26}$$

Note that

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

showing that (3.26) is equivalent to (3.25) and that the two interpretations of PCA are indeed equivalent.

3.2.1.3 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 compute the leading eigenvectors of

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

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 [81] and/or [72]).

An alternative is to use the Singular Value Decomposition (3.1). Let $X = [x_1 \cdots x_n]$ recall that,

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

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,

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

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, [79, 133, 112]).

Numerical stability is another important reason why computing the principal components using the SVD is preferable. Since the eigenvalues of Σ_n are proportional to the squares of the singular values of $X - \mu_n \mathbf{1}^T$, problems arise when the ratio of singular values exceeds 10^8, causing the ratio of the corresponding eigenvalues of Σ_n to be larger than 10^{16}. In this case, the smaller eigenvalue would be rounded to zero (due to machine precision), which is certainly not desirable.

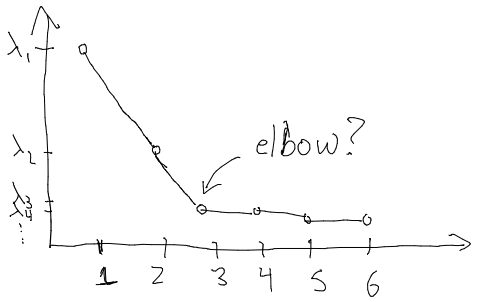
3.2.1.4 Which d should we pick?

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. Denoising: often times the data belongs to a lower dimensional space but is corrupted by high dimensional noise. When using PCA it is oftentimes possible to reduce the noise while keeping the signal.

2. Downstream analysis: One may be interested in running an algorithm (clustering, regression, etc.) that would be too computationally expensive or too statistically insignificant to run in high dimensions. Dimension reduction using PCA may help there.

In these applications (and many others) it is not clear how to pick d. 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. Since the total variance is $\mathrm{Tr}(\Sigma_n) = \sum_{k=1}^{p} \lambda_k$, the proportion of variance in the i'th component is nothing but $\frac{\lambda_i}{\mathrm{Tr}(\Sigma_n)}$. A plot of the values of the ordered eigenvalues, also known as a scree plot, helps identify a reasonable choice of d. Here is an example:



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 choose the cut-off value.

3.2.2 PCA in high dimensions and the Marčenko-Pastur law

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

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

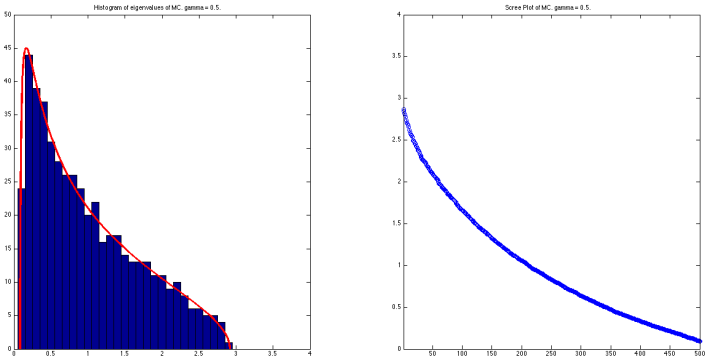
have p in the order of n (or, sometimes, even larger). For example, if our dataset is composed by images then n is the number of images and p the number of pixels per image; it is conceivable that the number of pixels be on the order of the number of images in a set. Unfortunately, in that case, it is 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 try to understand the spectral properties of

$$S_n = \frac{1}{n} XX^T,$$

where $x_1, \ldots, x_n$ are the columns of X. Since $x \sim \mathcal{N}(0, \Sigma)$ we know that $\mu_n \to 0$ (and, clearly, $\frac{n}{n-1} \to 1$), hence the spectral properties of S_n will be essentially the same as Σ_n.[1]

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 Marčenko-Pastur distribution (3.27), that we will discuss below.



As one can see in the image, there are many eigenvalues considerably larger than 1, as well as many eigenvalues significantly smaller than 1. 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. [149] and [12]). This particular limiting distribution was first established in 1967 by Marčenko and Pastur [100] and is now referred to as the Marčenko-Pastur distribution. They showed that, if p and n are both going to ∞ with their ratio fixed

[1] In this case, S_n is actually the maximum likelihood estimator for Σ; we will discuss maximum likelihood estimation later in Chapter ???.

$p/n = \gamma \leq 1$, the sample distribution of the eigenvalues of S_n (like the histogram above), in the limit, will be

$$dF_\gamma(\lambda) = \frac{1}{2\pi}\frac{\sqrt{(\gamma_+ - \lambda)(\lambda - \gamma_-)}}{\gamma\lambda} 1_{[\gamma_-,\gamma_+]}(\lambda)d\lambda, \tag{3.27}$$

with support $[\gamma_-, \gamma_+]$, where $\gamma_- = (1-\gamma)^2$ and $\gamma_+ = (1+\gamma)^2$. This is plotted as the red line in the figure above.

Remark 3.5. We will not provide the proof of the Marčenko-Pastur law here (you can see, for example, [14] for several different proofs of it), but an approach to a proof is using the so-called moment method. The central 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,

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

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

3.2.3 Spike Models and BBP phase transition

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

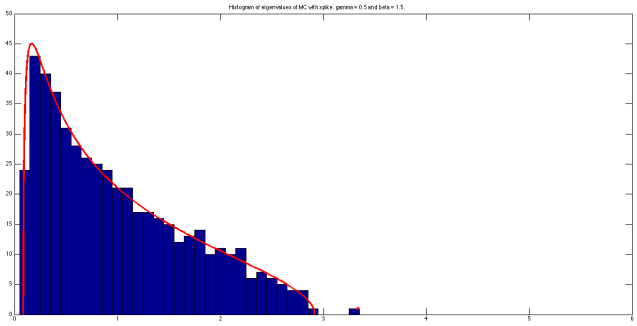
One way to think about this instance is as each data point x consisting of a signal part $\sqrt{\beta}g_0 u$ where g_0 is a one-dimensional standard Gaussian $\mathcal{N}(0,1)$ (i.e. a normally distributed multiple of a fixed vector $\sqrt{\beta}u$) and a noise part $g \sim \mathcal{N}(0,I)$ (independent of g_0). Then $x = g + \sqrt{\beta}g_0 u$ is a Gaussian random variable

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

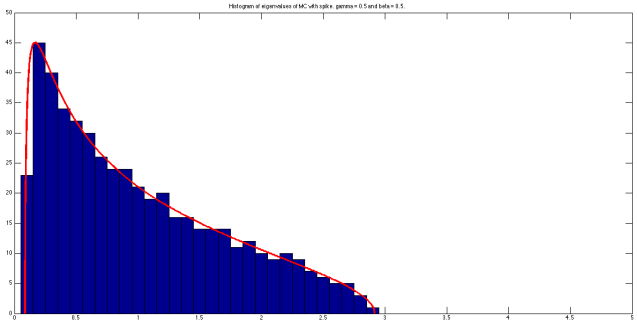
Whereas the signal part $\sqrt{\beta}g_0 u$ resides on a central line in the direction of u, the noise part is high dimensional and isotropic. We therefore refer to β as the signal-to-noise ratio (SNR). Indeed, β is the ratio of the signal variance (in the u-direction) to the noise variance (in each direction).

A natural question is whether this rank-1 perturbation can be seen in S_n. Or equivalently, can one detect the direction of the line u from corrupted measurements in high dimension? Let us build some intuition with an example. The following is the

histogram of the eigenvalues of a sample of S_n for $p = 500$, $n = 1000$, u is the first element of the canonical basis $u = e_1$, and $\beta = 1.5$:



The histogram suggests that there is an eigenvalue of S_n that "pops out" of the support of the Marčenko-Pastur distribution (below we will estimate the location of this eigenvalue, and that estimate corresponds to the red "x"). It is worth noting 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 $\beta = 0.5$:



It appears that, for $\beta = 0.5$, the histogram of the eigenvalues is indistinguishable from when $\Sigma = \mathrm{I}$. In particular, no eigenvalue is separated from the Marčenko-Pastur distribution.

This motivates the following question:

Question 3.1. For which values of γ and β do we expect to see an eigenvalue of S_n popping out of the support of the Marčenko-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, denoted β_c, below which we do not expect to see a change in the distribution of eigenvalues and above which we expect one of the eigenvalues to pop outside of the support. This phenomenon is

known as the BBP phase transition (after Baik, Ben Arous, and Péché [15]). There are many very nice papers about this and similar phenomena, including [120, 83, 15, 121, 16, 84, 32, 33]. [2]

In what follows we will find the critical value β_c and estimate the location of the largest eigenvalue of S_n for any β. While the argument we will use can be made precise (and is borrowed from [120]) we will be ignoring a few details for the sake of exposition. In other words, the argument below can be transformed into a rigorous proof, but it is not one at the present form.

We want to understand the behavior of the leading eigenvalue of the sample covariance matrix

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

Since $x \sim \mathcal{N}(0, I+\beta uu^T)$ we can write $x = (I+\beta uu^T)^{1/2} z$ where $z \sim \mathcal{N}(0,I)$ is an isotropic Gaussian. Then,

$$S_n = \frac{1}{n}\sum_{i=1}^{n} (I+\beta uu^T)^{1/2} z_i z_i^T (I+\beta uu^T)^{1/2} = (I+\beta uu^T)^{1/2} Z_n (I+\beta uu^T)^{1/2},$$

where $Z_n = \frac{1}{n}\sum_{i=1}^{n} z_i z_i^T$ is the sample covariance matrix of independent isotropic Gaussians. The matrices $S_n = (I+\beta uu^T)^{1/2} Z_n (I+\beta uu^T)^{1/2}$ and $Z_n(I+\beta uu^T)$ are related by a similarity transformation, and therefore have exactly the same eigenvalues. Hence, it suffices to find the leading eigenvalue of the matrix $Z_n(I+\beta uu^T)$, which is a rank-1 perturbation of Z_n (indeed, $Z_n(I+\beta uu^T) = Z_n + \beta Z_n uu^T$). We already know that the eigenvalues of Z_n follow the Marčenko-Pastur distribution, so we are left to understand the effect of a rank-1 perturbation on its eigenvalues.

To find the leading eigenvalue λ of $Z_n(I+\beta uu^T)$, let v be the corresponding eigenvector, that is,

$$Z_n(I+\beta uu^T)v = \lambda v.$$

Subtract $Z_n v$ from both sides to get

$$\beta Z_n uu^T v = (\lambda I - Z_n)v.$$

Assuming λ is not an eigenvalue of Z_n, we can multiply by $(\lambda I - Z_n)^{-1}$ to get[3]

$$\beta(\lambda I - Z_n)^{-1} Z_n uu^T v = v.$$

Our assumption also implies that $u^T v \neq 0$, for otherwise $v = 0$. Multiplying by u^T gives

[2] Notice that the Marčenko-Pastur theorem does not imply that all eigenvalues are actually in the support of the Marčenko-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, [120]).

[3] Intuitively, λ is larger than all the eigenvalues of Z_n, because it corresponds to a perturbation of Z_n by a positive definite matrix βuu^T; yet, a formal justification is beyond the present discussion.

$$\beta u^T(\lambda I - Z_n)^{-1} Z_n u (u^T v) = u^T v.$$

Dividing by $\beta u^T v$ (which is not 0 as explained above) yields

$$u^T(\lambda I - Z_n)^{-1} Z_n u = \frac{1}{\beta}. \tag{3.28}$$

Suppose $w_1, \ldots, w_p$ are orthonormal eigenvectors of Z_n (with corresponding eigenvalues $\lambda_1, \ldots, \lambda_p$), and expand u in that basis:

$$u = \sum_{i=1}^{p} \alpha_i w_i.$$

Plugging this expansion in (3.28) gives

$$\sum_{i=1}^{p} \frac{\lambda_i}{\lambda - \lambda_i} \alpha_i^2 = \frac{1}{\beta} \tag{3.29}$$

For large p, each α_i^2 concentrates around its mean value $\mathbb{E}[\alpha_i^2] = \frac{1}{p}$ (again, this statement can be made rigorous), and (3.29) becomes

$$\lim_{p\to\infty} \frac{1}{p} \sum_{i=1}^{p} \frac{\lambda_i}{\lambda - \lambda_i} = \frac{1}{\beta} \tag{3.30}$$

Since the eigenvalues λ_1, λ_p follow the Marčenko-Pastur distribution, the limit on the left hand side can be replaced by the integral

$$\int_{\gamma_-}^{\gamma_+} \frac{t}{\lambda - t}\, dF_\gamma(t) = \frac{1}{\beta} \tag{3.31}$$

Using an integral table (or an integral software), we find that

$$\frac{1}{\beta} = \int_{\gamma_-}^{\gamma_+} \frac{t}{\lambda - t}\, dF_\gamma(t) = \frac{1}{4\gamma}\left[2\lambda - (\gamma_- + \gamma_+) - 2\sqrt{(\lambda - \gamma_-)(\lambda - \gamma_+)}\right]. \tag{3.32}$$

For $\lambda = \gamma_+$, that is, when the top eigenvalue touches the right edge of the Marčenko-Pastur distribution, (3.32) becomes $\frac{1}{4\gamma}(\gamma_+ - \gamma_-)$. This is the critical point that one gets the pop out of the top eigenvalue from the bulk of the Marčenko-Pastur distribution. To calculate the critical value β_c, we recall that $\gamma_- = (1 - \sqrt{\gamma})^2$ and $\gamma_+ = (1 + \sqrt{\gamma})^2$, hence

$$\frac{1}{\beta_c} = \frac{1}{4\gamma}\left((1 + \sqrt{\gamma})^2 - (1 - \sqrt{\gamma})^2\right). \tag{3.33}$$

Therefore, the critical SNR is

$$\beta_c = \sqrt{\gamma} = \sqrt{\frac{p}{n}}. \tag{3.34}$$

When $\beta > \sqrt{\frac{p}{n}}$ one can observe the pop out of the top eigenvalue from the bulk.

Eq. (3.34) illustrates the interplay of the SNR β, the number of samples n, and the dimension p. Low SNR, small sample size, and high dimensionality are all obstacles for detecting linear structure in noisy high dimensional data.

More generally, inverting the relationship between β and λ given by (3.32) (which simply amounts to solving a quadratic), we find that the largest eigenvalue λ of the sample covariance matrix S_n has the limiting value

$$\lambda \to \begin{cases} (\beta+1)\left(1+\frac{\gamma}{\beta}\right) & \text{for} \quad \beta \geq \sqrt{\gamma}, \\ (1+\sqrt{\gamma})^2 & \text{for} \quad \beta < \sqrt{\gamma}. \end{cases} \tag{3.35}$$

In the finite sample case λ will be fluctuating around that value.

Notice that the critical SNR value, $\beta_c = \sqrt{\gamma}$ is buried deep inside the support of the Marčenko-Pastur distribution, because $\sqrt{\gamma} < \gamma_+ = (1+\sqrt{\gamma})^2$. In other words, the SNR does not have to be greater than the operator norm of the noise matrix in order for it to pop out. We see that the noise effectively pushes the eigenvalue to the right (indeed, $\lambda > \beta$).

The asymptotic squared correlation $|\langle u, v\rangle|^2$ between the top eigenvector v of the sample covariance matrix and true signal vector u can be calculated in a similar fashion. The limiting correlation value turns out to be

$$|\langle v, u\rangle|^2 \to \begin{cases} \frac{1-\frac{\gamma}{\beta^2}}{1+\frac{\gamma}{\beta^2}} & \text{for} \quad \beta \geq \sqrt{\gamma} \\ 0 & \text{for} \quad \beta < \sqrt{\gamma} \end{cases} \tag{3.36}$$

Notice that the correlation value tends to 1 as $\beta \to \infty$, but is strictly less than 1 for any finite SNR.

3.2.3.1 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 BBP like transition for this matrix ensemble [66]. 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
$$\lambda_{\max}\left(\frac{1}{\sqrt{n}}W + \xi vv^T\right) \to \xi + \frac{1}{\xi}. \tag{3.37}$$

The typical correlation, with v, of the leading eigenvector $v_{\max}$ of $\frac{1}{\sqrt{n}}W + \xi vv^T$ is also known:

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

Recent work addresses the problem of when it is possible to statistically detect a spike in a random matrix, for different distributions on the spike and the underlying matrix [123].

3.2.4 Rank and covariance estimation

The spike model and random matrix theory thus offers a principled way for determining the number of principal components, or equivalently of the rank of the hidden linear structure: simply count the number of eigenvalues to the right of the Marčenko-Pastur distribution. In practice, this approach for rank estimation is often too simplistic for several reasons. First, for actual datasets, n and p are finite, and one needs to take into account non-asymptotic corrections and finite sample fluctuations [87, 88]. Second, the noise may be heteroskedastic (that is, noise variance is different in different directions). Moreover, the noise statistics could also be unknown and it can be non-Gaussian [95]. In some situations it might be possible to estimate the noise statistics directly from the data and to homogenize the noise (a procedure sometimes known as "whitening") [94]. These situations call for careful analysis, and many open problems remain in the field.

Another popular method for rank estimation is using permutation methods. In permutation methods, each column of the data matrix is randomly permutated, so that the low-rank linear structure in the data is destroyed through scrambling, while only the noise is preserved. The process can be repeated multiple times, and the statistics of the singular values of the scrambled data matrices are then used to determine the rank. In particular, only singular values of the original (unscrambled) data matrix that are larger than the largest singular value of the scrambled matrices (taking fluctuations into account of course) are considered as corresponding to signal and are counted towards the rank. The mathematical analysis of permutation methods is another active field of research [58, 59].

In some applications, the objective is to estimate the low rank covariance matrix of the clean signal Σ from the noisy measurements. We saw that in the spike model, the eigenvalues of the sample covariance matrix are inflated due to noise. It is therefore required to shrink the computed eigenvalues of S_n in order to obtain a better estimate of the eigenvalues of Σ. That is, if

$$S_n = \sum_{i=1}^{p} \lambda_i v_i v_i^T$$

is the spectral decomposition of S_n, then we seek an estimator of Σ, denoted $\hat{\Sigma}$ of the form

$$\hat{\Sigma} = \sum_{i=1}^{p} \eta(\lambda_i) v_i v_i^T .$$

The scalar nonlinearity $\eta : \mathbb{R}^+ \to \mathbb{R}^+$ is known as the shrinkage function. An obvious shrinkage procedure is to estimate $\beta = \eta(\lambda)$ from the computed λ by inverting (3.35) (and setting $\beta = 0$ for $\lambda < \gamma_+$). It turns out that this particular shrinker is optimal in terms of the operator norm loss. However, for other loss functions (such as the Frobenius norm loss), the optimal shrinkage function takes a different form [62]. The reason why the shrinker depends on the loss function is that the eigenvectors of S_n are not perfectly correlated with those of Σ but rather make some non-trivial angle, as in (3.36). In other words, the eigenvectors are noisy, and it may require more aggressive shrinkage to account for that error in the eigenvector. It can be shown that the eigenvector v of the sample covariance is uniformly distributed in a cone around u whose opening angle is given by (3.36). While we can improve the estimation of the eigenvalue via shrinkage, it is however unclear how to improve the estimation of the eigenvector (without any a priori knowledge about it). Finally, we remark that eigenvalue shrinkage also plays an important role in denoising, as will be discussed in Chapter ??.