Lecture 4

3.1.3 Computing the principal components $\mathcal{V}$:

The above calculations show that the principle components $\mathcal{V} = \{v_1, \ldots, v_d\}$ are computed by solving:

$$\max_{V^T V = I} \operatorname{Tr}(V^T \Sigma_n V).$$

This is a particular case of the following optimization problem:

$$\max_{V^T V = I} \operatorname{Tr}(V^T M V),$$

where M is any symmetric $p \times p$ matrix. This problem in equivalent to:

$$\max_{\substack{v_1, \ldots, v_d \in \mathbb{R}^p \\ v_i^T v_j = \delta(i-j)}} \sum_{k=1}^{d} v_k^T M v_k, \tag{6}$$

where $\delta(0) = 1$ and $\delta(x) = 0$ if $x \neq 0$. When $d = 1$ this reduces to:

$$\max_{\substack{v \in \mathbb{R}^p \\ \|v\| = 1}} v^T M v.$$

Since M is symmetric, it has p orthonormal eigenvectors $\varphi_1, \ldots, \varphi_p \in \mathbb{R}^p$ with corresponding (not necessarily distinct) eigenvalues $\lambda_1, \ldots, \lambda_p$ such that

$$M\varphi_i = \lambda_i \varphi_i.$$

Suppose the eigenvalues are ordered so that $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$. Then, it is not hard to see that (verify for yourself!):

$$\max_{\substack{v \in \mathbb{R}^p \\ \|v\| = 1}} v^T M v = \lambda_1, \tag{7}$$

which is achieved by taking $v = \varphi_1$. In fact, the more general (6) is maximized by taking $v_i = \varphi_i$ for $i = 1, \ldots, d$, so that

$$\max_{\substack{v_1, \ldots, v_d \in \mathbb{R}^p \\ v_i^T v_j = \delta(i-j)}} \sum_{k=1}^{d} v_k^T M v_k = \sum_{k=1}^{d} \lambda_k.$$

Now let us return to computing the principle components. The above shows that we need to compute the d eigenvectors of Σ_n with the d largest eigenvalues (note the eigenvalues of Σ_n will be nonnegative). We can do this using the singular value decomposition (SVD). Recall that the SVD of an $p \times n$ matrix A decomposes A as:

$$A = U_L D U_R,$$

where U_L is a $p \times p$ matrix, D is a nonnegative diagonal $p \times n$ matrix, and U_R is an $n \times n$ matrix. The matrix U_L and U_R satisfy:

$$U_L^T U_L = U_L U_L^T = I \qquad U_R^T U_R = U_R U_R^T = I.$$

The columns of U_L consist of eigenvectors of AA^T and the columns of U_R are the eigenvectors of $A^T A$. The nonzero eigenvalues of AA^T and $A^T A$ are identical and positive; denote them as $\lambda_1, \ldots, \lambda_m$. The squareroots of these eigenvalues are the singular values of A, and they are the nonzero values on the diagonal of D.

Now let's apply the SVD to the matrix $X_n - \mu_n \mathbf{1}^T$; we get:

$$X_n - \mu_n \mathbf{1}^T = U_L D U_R^T.$$

But by the above the columns of U_L are the eigenvectors of Σ_n and the diagonal of D^2 gives the eigenvalues of Σ_n. Note that the eigenvectors of U_R are the eigenvectors of the dot product kernel defined in Section 2.1, if $\mu_n = 0$. Regardless, the SVD can be computed efficiently, especially if one only wants to compute the top d singular values and corresponding eigenvectors.

Exercises

Exercise 8. Prove (7) (hint: one way is to use the Spectral Theorem applied to M).

Exercise 9. Write a script or function to compute the principal components and associated eigenvalues of a data matrix X_n. You may use an existing eigensolver or SVD (e.g., in MATLAB `eig` or `svd`, etc.).

3.1.4 How do we pick the dimension d?

PCA can be used to visualize data in two or three dimensions; however, these dimensions may not always be sufficient to capture all of the variability within the data. It is still useful, though, to reduce the dimension of the data. Indeed, the PCA coordinates can sometimes be interpretable; they can be used to denoise the data; and it may be the case that an algorithm you want to run on the data is too expensive in high dimensions, but is tractable in low dimensions.

The question of how to pick the number d of principal components then arises. One heuristic is to pick d so that $\{V^T x_i\}_{i \leq n}$ captures a certain percentage of the total variance of $\mathcal{X}_n$ (e.g., 95%). This can be done easily by observing if $\lambda_1, \ldots, \lambda_n$ are the eigenvalues of Σ_n, then (check this on your own!):

$$\text{Total Variance}(\mathcal{X}_n) = \sum_{i=1}^{n} \lambda_i. \tag{8}$$

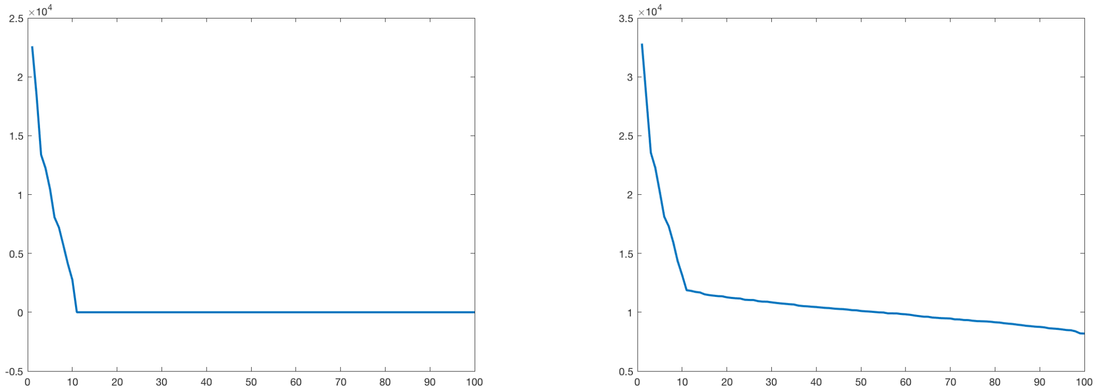
Thus the value of d that yields, say 95% of the variance, is the minimum such d so that:

$$\frac{\sum_{i=1}^{d} \lambda_i}{\sum_{i=1}^{n} \lambda_i} > 0.95.$$

Another related heuristic is to plot the eigenvalues of Σ_n in decreasing order, and look for when they decay rapidly or when the plot has "kink". For example, consider a 100 dimensional data set $\mathcal{X}_n$ drawn from the normal distribution $\mathcal{N}(\mu, \Sigma)$, with $\mu \in \mathbb{R}^{100}$ and $\Sigma \in \mathbb{R}^{100 \times 100}$. Suppose further that $\operatorname{rank}(\Sigma) = 10$. Thus even though a priori $\mathcal{X}_n$ is 100 dimensional, its intrinsic dimension is 10. If we take $n = 10000$ and plot the eigenvalues, we get the plot in Figure 3(a). From this plot, it is clear the intrinsic dimension of the data is 10, since indeed all eigenvalues after the tenth one are numerically zero. If we add a (relatively) small amount of noise to the data, we can still discern that it's 10 dimensional; see, Figure 3(b). Notice the eigenvalue beyond the tenth one are no longer zero, but they are significantly smaller than the ten corresponding to the principal directions of the data. These other 90 eigenvalues are capturing the directions of the noise, which are relatively small and isotropic.

Exercises

Exercise 10. Download the Yale faces data set from D2L (`Yale_64x64.mat`) and load it into MATLAB. It has two variables, `fea` and `gnd`. The variable `fea` contains 165 64 × 64 pixel faces taken from 15 people, which is stored as a 4096 × 165 matrix. The variable `gnd` indexes which faces



(a) No noise added to the data

(b) With noise added to the data

Figure 3: PCA eigenvalues of 100 dimensional data drawn from a 10 dimensional distribution, plotted in descending order.

belong to which people. Extract some of the columns of this matrix, reshape them into 64×64 matrices, and visualize the faces. Then apply your PCA function to the matrix `fea`, considering it as $n = 165$ data samples in dimension $p = 4096$. You should get principal components which are $4096x1$ vectors. Reshape these principal components into 64×64 matrices. Visualize the top principal components (i.e., those whose corresponding eigenvalues are the largest) - however many you like. What do you observe? Can you interpret them? How many dimensions do you need to capture most of the variance within the data?

3.2 PCA in high dimensions

This is a shortened version of [2, Chapter 1.2 & Chapter 1.3]

3.2.1 Brief review of the Law of Large Numbers

Let $Y_1, Y_2, \ldots \in \mathbb{R}$ be an infinite sequence of independently and identically distributed (i.i.d.) random variables, such that they all have the same expected value:

$$\forall\, i > 0, \quad \mathbb{E}(Y_i) = \mu.$$

As before, define the sample mean as:

$$\mu_n = \frac{1}{n}\sum_{i=1}^{n} Y_i.$$

Then the *Law of Large Numbers* states that, in some suitable sense that can be made precise,

$$\mu_n \to \mu, \quad \text{as } n \to \infty.$$

There is a weak and strong law, the content of which differs in the nature of the convergence $\mu_n \to \mu$.

3.2.2 Marchenko-Pastur distribution

Assume now that $\mathcal{X}_n = \{x_1, \ldots, x_n\} \subset \mathbb{R}^p$ are independent draws from a normal distribution with mean zero:

$$x_i \sim \mathcal{N}(0, \Sigma),$$

where Σ is the $p \times p$ covariance matrix. Recall the sample covariance:

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

Since $\mathbb{E}\mu_n = 0$ and $\mathbb{E}\Sigma_n = \Sigma$, an application of the law of large numbers shows that $\Sigma_n \to \Sigma$ as $n \to \infty$. Thus if p is fixed and n is large, when we use PCA we should find a low dimensional representation of the distribution, which corresponds to the large eigenvalues of Σ and their corresponding eigenvectors.

However, it is often the case that p is on the order of n. In this case the law of large numbers does not hold and the analysis is more delicate - this leads to the field of high dimensional statistics. For simplicity, we consider the following matrix S_n as opposed to Σ_n:

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

Indeed, we know $\mu_n \to 0$ and $n/(n-1) \to 1$, so the two matrices are quite similar and have essentially the same spectral properties.

It is illuminating to look at the simple example of $\Sigma = I$. The distribution in this case has no low dimensional structure, since all eigenvalues of the covariance matrix Σ are equal to one. However, if we take $p = 500$ and $n = 1000$, the histogram of the 1000 eigenvalues of S_n is given by the blue bars in Figure 4(a); the same eigenvalues are plotted in descending order in 4(b). Notice that many eigenvalues are much smaller than one, and many other eigenvalues are much larger than one. If we computed these eigenvalues without knowing that the underlying covariance $\Sigma = I$, we might suspect that the data has an intrinsic dimension smaller than p. However this is not the case! In fact, this distribution of eigenvalues is predicted by the Marchenko-Pastur distribution [5].

In [5], Marchenko and Pastur proved that if $p \to \infty$ and $n \to \infty$ and $p/n = \gamma \leq 1$, then the eigenvalues of S_n (for $\Sigma = I$) in the limit will be distributed as:

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

where $\gamma_{+} = (1 + \sqrt{\gamma})^2$ and $\gamma_{-} = (1 - \sqrt{\gamma})^2$; this is the red curve in Figure 4(a)!

3.2.3 Spike models

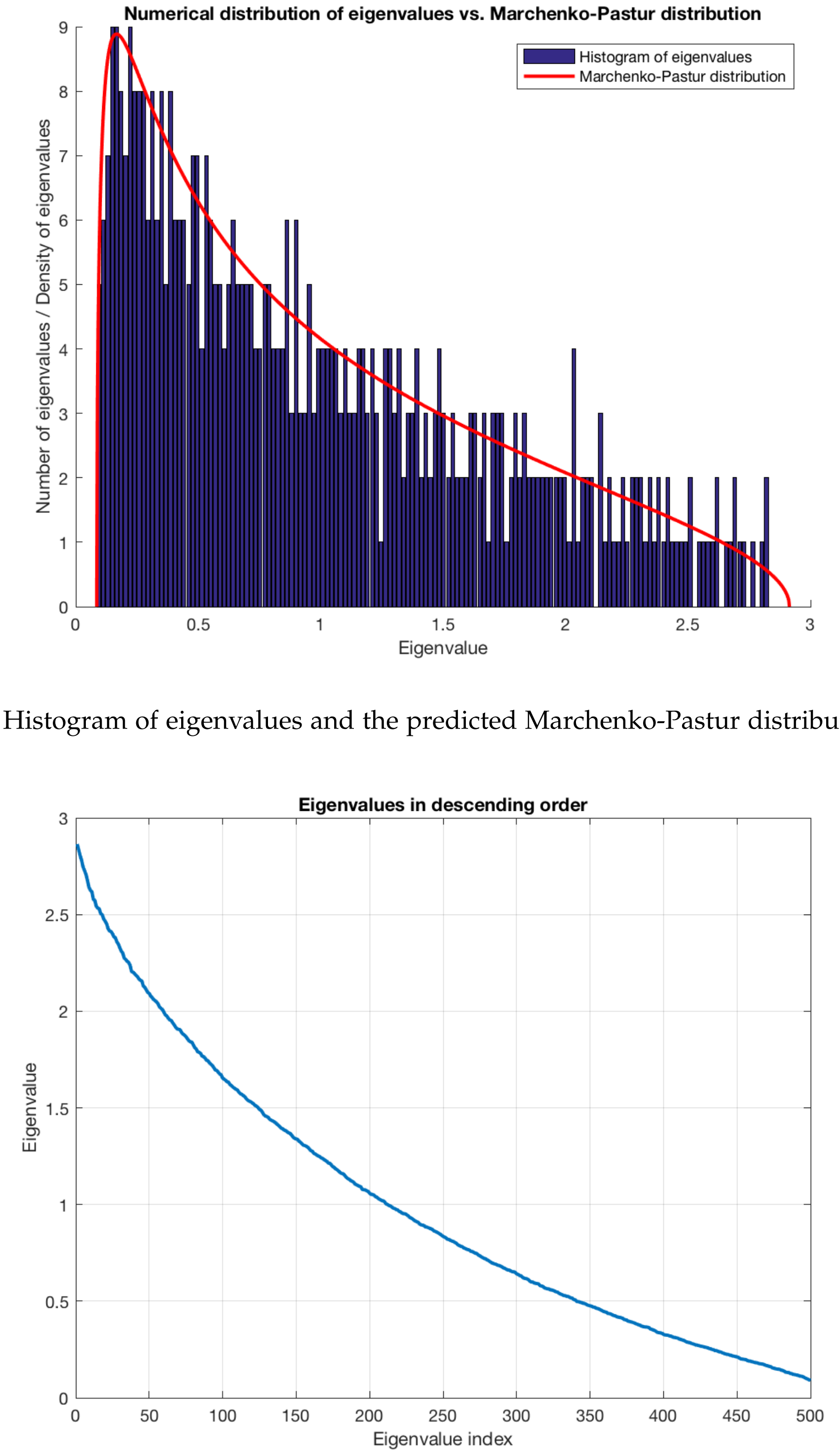
Let us now consider the case where there in fact is a (linear) one-dimensional structure in the data, and see when we can capture it with PCA. To that end, suppose that $x_i \sim \mathcal{N}(0, \Sigma)$, and that Σ satisfies the “spike” model:

$$\Sigma = I + \beta v v^T, \quad \beta \geq 0, \;\; v \in \mathbb{R}^p. \tag{9}$$

Note that $\beta v v^T$ is a rank 1 matrix; if $\beta > 0$, then it defines a single dimension along which the data has variance $1 + \beta$, which is more than the “noisey” part in the other $p - 1$ dimensions, for which the magnitude of the variance is 1.

We want to know if we can see this direction from S_n (i.e., PCA). Let’s check it numerically for $v = (1, 0, \ldots, 0)$ and $\beta = 2$; we get the histogram in Figure 5(a). Notice the presence of an isolated eigenvalue between 3.5 and 4.0. This is not an accident! Now let’s try $\beta = 1/2$; we get the eigenvalue distribution in Figure 5(b). Notice that there is no longer an isolated eigenvalue, and we would have no idea that there is a single dimension along which the variance of the data is larger!

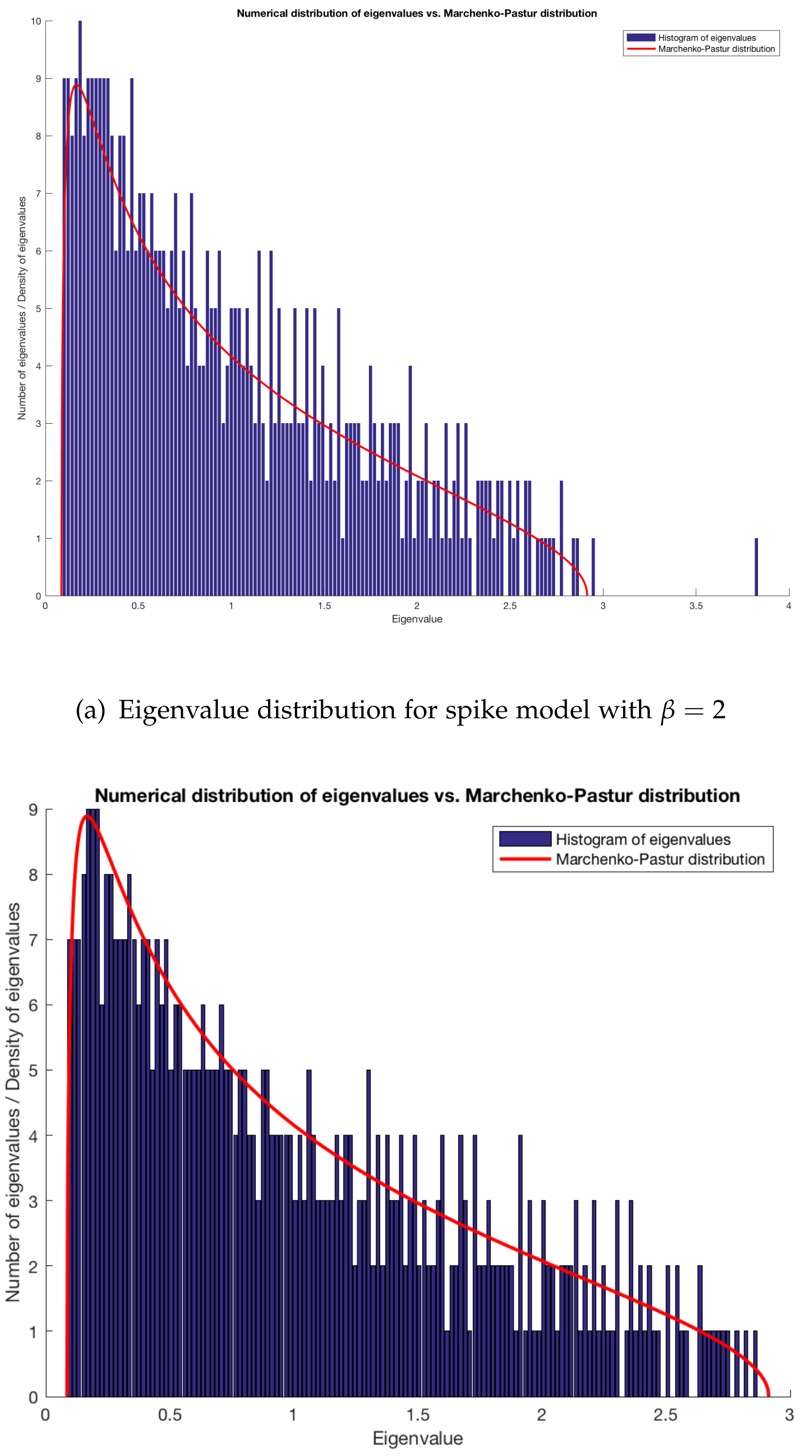
This raises the question of when can we find this linear one dimensional structure. The answer depends on $\gamma = p/n$ and β. As the Theorem



(a) Histogram of eigenvalues and the predicted Marchenko-Pastur distribution.

(b) Plot of the eigenvalues in descending order.

Figure 4: Marchenko-Pastur distribution



(a) Eigenvalue distribution for spike model with $\beta = 2$

(b) Eigenvalue distribution for spike model with $\beta = 1/2$

Figure 5: The spike model

1 shows, there is a critical value of β (depending on γ) below which the eigenvalue does not "pop out." This phenomenon is known as the BBP transition, after [6].

Theorem 1 ([7]). *In the spike model* (9)*, suppose* $v = e_1 = (1, 0, \ldots, 0)$*. Let* $\lambda_{\max}$ *denote the largest eigenvalue of* S_n*, and let* $v_{\max}$ *be its associated eigenvector.*

- *If* $\beta \leq \sqrt{\gamma}$*, then*
$$\lambda_{\max} \to \gamma_+ \tag{10}$$
and
$$|\langle v_{\max}, e_1 \rangle|^2 \to 0. \tag{11}$$

- *If* $\beta > \sqrt{\gamma}$*, then*
$$\lambda_{\max} \to (1+\beta)\left(1 + \frac{\gamma}{\beta}\right) > \gamma_+ \tag{12}$$
and
$$|\langle v_{\max}, e_1 \rangle|^2 \to \frac{1 - \gamma/\beta^2}{1 - \gamma/\beta}. \tag{13}$$

Equations (10) and (12) give theoretical justification to the numerical behavior we saw in Figures 5(a) and 5(b), and furthermore give the exact critical value of β and the expected value of $\lambda_{\max}$. Equations (11) and (13) show how correlated the PCA direction of maximum variance $v_{\max}$ is with the actual direction of maximum variance e_1. Below the critical value, there is no correlation (in the limit). Above the critical value, there is positive correlation. Notice that this correlation goes to 1 as $\beta \to \infty$ (for fixed γ); in other words, as the variance of the single dimension becomes increasingly large, it is easier to find, which makes intuitive sense. Also, as $\gamma \to 0$ (for β fixed), the correlation also goes to 1; this is essentially the transition back to "standard" statistics, in which the dimension p is fixed and the number of samples $n \to \infty$.

References

[1] Bernhard Schölkopf and Alexander J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. Adaptive Computation and Machine Learning. The MIT Press, 2002.

[2] Afonso S. Bandeira. Ten lectures and forty-two open problems in the mathematics of data science. MIT course *Topics in Mathematics of Data Science*, 2015.

[3] Jon Shlens. A tutorial on principal component analysis. arXiv:1404.1100, 2014.

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

[5] 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.

[6] 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.

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