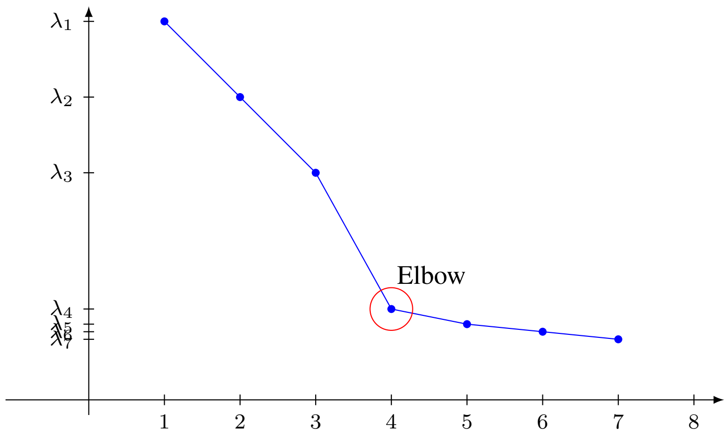

Figure 4.1: Scree Plot

4.1.3 How to carry out PCA

Given $\mathbf{x}_1, \ldots, \mathbf{x}_n \mathbb{R}^p$, fix $k \ll p$.

- Compute $\mathbf{S}_n = \frac{1}{n-1} \sum_{i=1}^{n} (\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x}_i - \overline{\mathbf{x}})^T$. Find its spectral decomposition as $\mathbf{S}_n = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^T$ where $\mathbf{\Lambda} = \operatorname{diag}(\lambda_1, \ldots, \lambda_p)$ with $\lambda_1 \geq \cdots \geq \lambda_p$ and $\mathbf{V} = (\mathbf{v}_1, \ldots, \mathbf{v}_p) \in \mathcal{O}(p)$.
- $\mathbf{v}_1, \ldots, \mathbf{v}_k$ are called the k Principal eigenvectors to the principal eigenvalues $\lambda_1 \geq \cdots \geq \lambda_k$.
- Projected points are found by:

$$\hat{\mathbf{x}}_i = \begin{bmatrix} \mathbf{v}_1^T \\ \vdots \\ \mathbf{v}_k^T \end{bmatrix} \mathbf{x}_i, \qquad i = 1, \ldots, n$$

Let us discuss computational complexity of PCA. Using the conventional method, discussed above, the complexity of constructing $\mathbf{S}_n$ is $O(np^2)$ [1] and the complexity of spectral decomposition is $O(p^3)$ [Ban08]. Therefore the computational complexity of both steps together are $O(\max\{np^2, p^3\})$.

However this can be improved. Assume $p < n$. Write:

$$\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n] \quad \text{and} \quad \mathbf{S}_n = \frac{1}{n-1}(\mathbf{X} - \overline{\mathbf{x}}\mathbf{1}_n^T)(\mathbf{X} - \overline{\mathbf{x}}\mathbf{1}_n^T)^T.$$

[1] This is called Big-O notation or Bachmann-Landau notation. A function $f(n)$ is $O(g(n))$ if for some $n_0 > 0$ and a constant $c > 0$, $|f(n)| \leq c|g(n)|$ for $n \geq n_0$. For example, if an algorithm over n objects takes at most $n^2 + n$ time to run, then its complexity is $O(n^2)$.

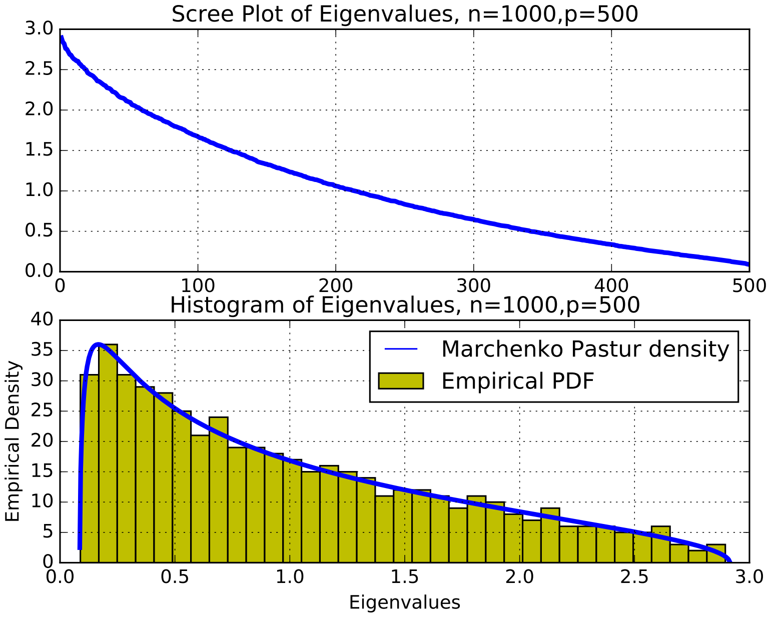

Figure 4.2: Eigenvalues of $\mathbf{S}_n$ and its scree plot

Consider singular value decomposition (SVD) of $\mathbf{X} - \overline{\mathbf{x}}\mathbf{1}_n^T = \mathbf{U}_{p\times p}\mathbf{D}\mathbf{V}_{p\times n}^T$ where $\mathbf{U} \in \mathcal{O}(p)$, $\mathbf{V}^T\mathbf{V} = \mathbf{I}_p$, $\mathbf{D} = \mathrm{diag}(\sigma_1, \dots, \sigma_p)$. Using this decomposition, we have:

$$\mathbf{S}_n = \frac{1}{n-1}\mathbf{U}\mathbf{D}\mathbf{V}^T\mathbf{V}\mathbf{D}\mathbf{U}^T = \frac{1}{n-1}\mathbf{U}\mathbf{D}^2\mathbf{U}^T.$$

Hence $\mathbf{U} = [\mathbf{u}_1, \dots, \mathbf{u}_p]$ contains the eigenvectors of $\mathbf{S}_n$. Computational complexity of finding SVD for $\mathbf{X} - \overline{\mathbf{x}}\mathbf{1}_n^T$ is given by $O(\min\{n^2p, p^2n\})$. However if one is only interested in top k eigenvectors, the cost reduces to $O(dnp)$.

Another issue is about the choice of k. If the goal of PCA is data visualization, then $k = 2$ or $k = 3$ are reasonable choices. But PCA is also used for dimensionality reduction. In application, it can happen that the data lies in a low dimensional subspace but it is corrupted by a high dimensional noise. Also, it is possible that some algorithms are computationally expensive to run on high dimensions and it makes sense to bring the data to lower dimensions and run the algorithm more efficiently on lower dimensional space.

To choose proper k, one heuristic is to look at the scree plot or scree graph. The scree plot is the plot of ordered eigenvalues of $\mathbf{S}_n$.

The scree graph was introduced by Raymond B. Cattell [Cat66]. It is a very subjective way of determining k. The idea is to find k from the plot such that the line through the points to the left of k is steep and the line through the points to the right of k is not steep. This looks like an elbow in the scree plot. In Figure 4.1, a scree plot is

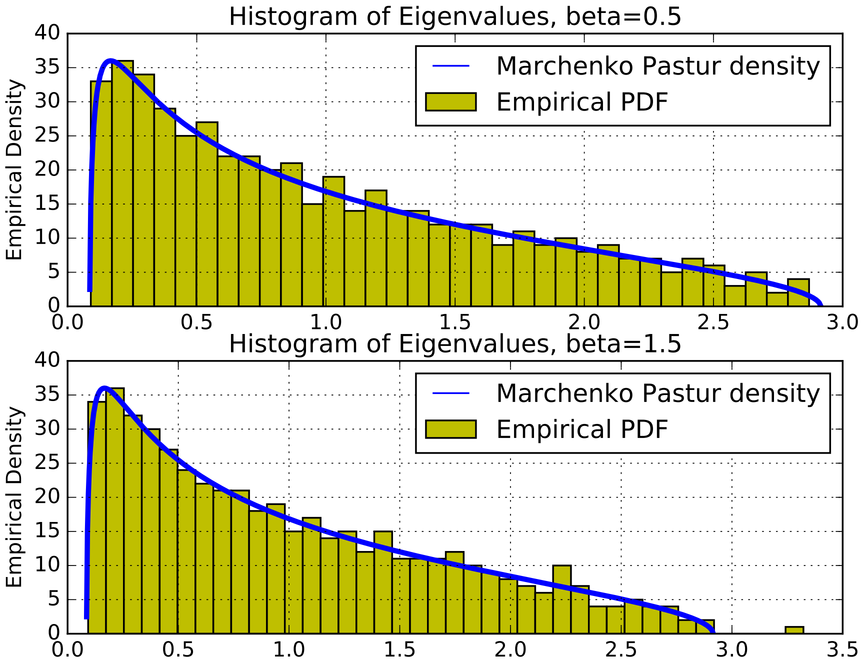

Figure 4.3: Eigenvalues of $\mathbf{S}_n$ for Spike model with $\beta = 1.5, 0.5$

shown. The value of k can be chosen by recognizing an elbow in the graph of ordered eigenvalues.

4.1.4 Eigenvalue structure of S_n in high dimensions

Suppose that $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^p$ independent samples of a Gaussian random variable $\mathbf{X} \sim N_p(\mathbf{0}, \mathbf{\Sigma})$. Let $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]$. Estimate Σ by $\mathbf{S}_n = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i\mathbf{x}_i^T = \frac{1}{n}\mathbf{X}\mathbf{X}^T$.

If p is fixed, from law of large numbers, $\mathbf{S}_n$ will tend to $\mathbf{\Sigma}$ as $n \to \infty$ almost everywhere. However if both n and p are large, then it is not clear anymore what the relation between $\mathbf{S}_n$ and $\mathbf{\Sigma}$ is. To see this, consider the case where $\mathbf{\Sigma} = \mathbf{I}$ [Ban08]. Figure 4.2 shows the scree plot and histogram of the eigenvalues for $n = 1000$ and $p = 500$. The plot shows that there are many eigenvalues bigger than 1 unlike $\mathbf{\Sigma} = \mathbf{I}$ which has all eigenvalues equal to one. Scree plot also implies that data lies on a low dimensional space which is also not true.

Following theorem is about distribution of eigenvalues of $\mathbf{S}_n$ when p and n are comparable.

Theorem 4.1 (Marchenko-Pastur, 1967)**.** *Let $\mathbf{X}_1, \ldots, \mathbf{X}_n$ be i.i.d. random vectors on $\mathbb{R}^p$ with $\mathbb{E}(\mathbf{X}_i) = 0$ and $\mathrm{Cov}(\mathbf{X}_i) = \sigma^2\mathbf{I}_p$. Let $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_n) \in \mathbb{R}^{p\times n}$ and $\mathbf{S}_n = \frac{1}{n}\mathbf{X}\mathbf{X}^T \in \mathbb{R}^{p\times p}$. Let $\lambda_1, \ldots, \lambda_p$ be the eigenvalues of $\mathbf{S}_n$. Suppose that $p, n \to \infty$ such* *that $\frac{p}{n} \to \gamma \in (0,1]$ as $n \to \infty$. Then the sample distribution of $\lambda_1, \dots, \lambda_p$ converges almost surely to the following density:*

$$f_\gamma(u) = \frac{1}{2\pi\sigma^2 u\gamma}\sqrt{(b-u)(u-a)}, \;\; a \le u \le b$$

with $a(\gamma) = \sigma^2(1-\sqrt{\gamma})^2$ and $b(\gamma) = \sigma^2(1+\sqrt{\gamma})^2$.

Proof. Refer to [Bai99] for various proofs. □

Marchenko-Pastur distribution is presented in Figure 4.2 by the blue curve.

Remark 1. If $\gamma > 1$, there will be a mass point at zero with probability $1 - \frac{1}{\gamma}$. Since $\gamma > 1$, then $n < p$. Moreover the rank of $\mathbf{S}_n = \frac{1}{n}\mathbf{X}\mathbf{X}^T$ will be at most $\min(p, n)$ which is $n < p$ in this case. This means that $\mathbf{S}_n$ is not full rank and zero is definitely one of the eigenvalues.

The theorem shows that there is a wide spread of spectrum of eigenvalues even in the case i.i.d. distributed random variables. The main question is to what degree PCA can recover low dimensional structure from the data. Is PCA useful at all?

4.1.5 Spike Models

Suppose that there is a low dimensional structure in data. Let us say that each sample results from a point on a one dimensional space with an additional high dimensional noise perturbation. The one dimensional part is modeled by $\sqrt{\beta}G\mathbf{v}$ where $\mathbf{v}$ is a unit norm vector in $\mathbb{R}^p$, β is a non-negative constant and G is the standard normal random variable. The high dimensional noise is modeled by $\mathbf{U} \sim N_p(0, \mathbf{I}_p)$. Therefore the samples are $\mathbf{X}_i = \mathbf{U}_i + \sqrt{\beta}G_i\mathbf{v}$ with $\mathbb{E}(\mathbf{X}_i) = 0$. Since G_i and $\mathbf{U}_i$ are independent, using Theorem 3.3, we have:

$$\mathrm{Cov}(\mathbf{X}_i) = \mathrm{Cov}(\mathbf{U}_i) + \mathrm{Cov}(\sqrt{\beta}G_i\mathbf{v}) = \mathbf{I}_p + \mathbf{v}\mathrm{Cov}(\sqrt{\beta}G_i)\mathbf{v}^T = \mathbf{I}_p + \beta\mathbf{v}\mathbf{v}^T.$$

Suppose that $\mathbf{X}_1, \dots, \mathbf{X}_n$ are i.i.d. distributed with $\mathrm{Cov}(\mathbf{X}_i) = \mathbf{I}_p + \beta\mathbf{v}\mathbf{v}^T$. Let us look at distribution of eigenvalues for some numerical examples. Figure 4.3 shows the distribution of eigenvalues for $\beta = 1.5$ and $\beta = 0.5$ and $p = 500$ and $n = 1000$, and $\mathbf{v} = \mathbf{e}_1$. It can be seen that all eigenvalues appear inside the interval proposed by Marchenko-Pastur distribution when $\beta = 0.5$. However, the situation is different when $\beta = 1.5$. One eigenvalue pops out of the interval in this case. Note that in general the maximum eigenvalue of $\mathbf{I}_p + \beta\mathbf{e}_1\mathbf{e}_1^T$ is 1+1.5 which is 2.5, and all other eigenvalues are 1.

The question is whether there is a threshold for β above which we will see one eigenvalue popping out. The following theorem provides the transition point known as BPP (Baik, Ben Arous and Péché) transition.

Theorem 4.2 ([BAP05]). *Let $\mathbf{X}_1, \dots, \mathbf{X}_n$ be i.i.d. random vectors on $\mathbb{R}^p$ with $\mathbb{E}(\mathbf{X}_i) = 0$ and $\mathrm{Cov}(\mathbf{X}_i) = \mathbf{I}_p + \beta\mathbf{v}\mathbf{v}^T$, $\beta \ge 0$, $\mathbf{v} \in \mathbb{R}^p$, $\|\mathbf{v}\| = 1$. Let $\mathbf{X} = (\mathbf{X}_1, \dots, \mathbf{X}_n) \in \mathbb{R}^{p\times n}$ and $\mathbf{S}_n = \frac{1}{n}\mathbf{X}\mathbf{X}^T \in \mathbb{R}^{p\times p}$. Suppose that $p, n \to \infty$ such that $\frac{p}{n} \to \gamma \in (0,1]$ as $n \to \infty$.*

- *If* $\beta \leq \sqrt{\gamma}$ *then* $\lambda_{\max}(\mathbf{S}_n) \to (1+\sqrt{\gamma})^2$ *and* $|\langle \mathbf{v}_{\max}, \mathbf{v} \rangle| \to 0$.
- *If* $\beta > \sqrt{\gamma}$ *then* $\lambda_{\max}(\mathbf{S}_n) \to (1+\beta)(1+\frac{\gamma}{\beta}) > (1+\sqrt{\gamma})^2$ *and* $|\langle \mathbf{v}_{\max}, \mathbf{v} \rangle| \to \frac{1-\gamma/\beta^2}{1-\gamma/\beta}$.

The interpretation of this theorem is that, only if $\beta > \sqrt{\gamma}$, the largest eigenvalue exceeds the upper asymptotic bound of the asymptotic support and the corresponding eigenvector has a non-trivial correlation with the eigenvector $\mathbf{v}$.