Prof. Dr. Rudolf Mathar, Dr. Arash Behboodi, Emilio Balda

Exercise 4

Friday, November 10, 2017

Problem 1. *(PCA in 2-dimensional space)* Suppose that for n samples, the sample covariance matrix $\mathbf{S}_n$ is given by

$$\mathbf{S}_n = \begin{pmatrix} 14 & -14 \\ -14 & 110 \end{pmatrix}.$$

a) Calculate the spectral decomposition $\mathbf{V}\mathbf{\Lambda}\mathbf{V}^{\mathrm{T}}$ of $\mathbf{S}_n$ by determining the matrices $\mathbf{V}$ and $\mathbf{\Lambda}$.

b) Determine the best projection matrix $\mathbf{Q}$ to transform the two-dimensional samples to a one-dimensional data.

c) Determine the residuum $\frac{1}{n-1} \max_{\mathbf{Q}} \sum_{i=1}^{n} \|\mathbf{Q}\mathbf{x}_i - \mathbf{Q}\bar{\mathbf{x}}_n\|^2$ for the above choice of $\mathbf{Q}$.

Problem 2. *(PCA in 2-dimensional space)* Consider four vectors given as follows

$$\mathbf{x}_1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \quad \mathbf{x}_2 = \begin{pmatrix} -1 \\ 0 \end{pmatrix}, \quad \mathbf{x}_3 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \quad \mathbf{x}_4 = \begin{pmatrix} 0 \\ -1 \end{pmatrix}$$

a) Calculate the sample covariance matrix $\mathbf{S}_n$ and the spectral decomposition $\mathbf{V}\mathbf{\Lambda}\mathbf{V}^{\mathrm{T}}$ of $\mathbf{S}_n$ by determining the matrices $\mathbf{V}$ and $\mathbf{\Lambda}$.

b) Determine the best projection matrix $\mathbf{Q}$ to transform the two-dimensional samples to a one-dimensional data and calculate the projection of $\mathbf{x}_1$ and $\mathbf{x}_2$.

Problem 3. *(Spike model)* Fix $p = 500$ as the dimension of the space $\mathbb{R}^p$. Suppose that the data is generated from two one dimensional subspaces modeled by $\sqrt{0.2}G_1\boldsymbol{v}_1$ and $\sqrt{0.5}G_2\boldsymbol{v}_2$, where $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$ are orthogonal unit norm vectors in $\mathbb{R}^p$, and G_1 and G_2 are independent standard normal random variables. The high dimensional noise $\boldsymbol{U} \in \mathbb{R}^p$ is independent of both G_1 and G_2 and is modeled as a standard normal random vector. The covariance matrix of this model $\boldsymbol{X} = \boldsymbol{U} + \sqrt{0.2}G_1\boldsymbol{v}_1 + \sqrt{0.5}G_2\boldsymbol{v}_2$ is described by:

$$\mathrm{Cov}(\boldsymbol{X}) = \boldsymbol{I}_p + 0.2\boldsymbol{v}_1\boldsymbol{v}_1^{\mathrm{T}} + 0.5\boldsymbol{v}_2\boldsymbol{v}_2^{\mathrm{T}}.$$

Suppose that $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ are i.i.d. distributed with $\mathrm{Cov}(\boldsymbol{X}_i) = \mathrm{Cov}(\boldsymbol{X})$.

a) Find the minimum number n_2 of samples such that only the dominant eigenvalue is visible. Calculate the distance $\langle \boldsymbol{v}_2, \boldsymbol{v}_{\mathrm{dom}} \rangle$ for this case.

b) Find the minimum number n_1 of samples such that both dominant eigenvalues are visible. Calculate the distance $\langle \boldsymbol{v}_2, \boldsymbol{v}_{\mathrm{dom}} \rangle$ for this case. Sketch the Marchenko-Pastur density for the latter case along with both dominant eigenvalues of the sample covariance matrix $\mathbf{S}_n$.