Prof. Dr. Rudolf Mathar, Dr. Arash Behboodi, Markus Rothe

Exercise 6

Friday, November 23, 2018

Problem 1. *(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 $\boldsymbol{X}_1, \dots, \boldsymbol{X}_n$ 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 $\mathbf{S}_n$.

Problem 2. *(Centering Matrix)*

For a set of column vectors $\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)} \in \mathbb{R}^k$ and the centering matrix $\mathbf{E}_k = \mathbf{I}_k - \frac{1}{k}\mathbf{1}_k\mathbf{1}_k^{\mathrm{T}} \in \mathbb{R}^{k\times k}$, let $\mathbf{X} \in \mathbb{R}^{n\times k}$ be $\mathbf{X} = [\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)}]^{\mathrm{T}}$ and $\overline{x}^{(j)} = \frac{1}{k}\sum_{i=1}^{k} x_i^{(j)}$, where $x_i^{(j)}$ is the i-th entry of $\mathbf{x}^{(i)}$.

a) Show that $\mathbf{E}_k\mathbf{x}^{(j)} = \mathbf{x}^{(j)} - \overline{x}^{(j)}\mathbf{1}_k$.

b) Show that the (i, j)-th entry of $\mathbf{E}_k\mathbf{X}^{\mathrm{T}}$ is given by $x_i^{(j)} - \overline{x}^{(j)}$.

c) Show that $\sum_{i=1}^{k}\left(\mathbf{E}_k\mathbf{X}^{\mathrm{T}}\right)_{i,j} = 0$ for any $j \in \{1, 2, \dots, n\}$.

Problem 3. *(Spike model II)*

Show that if $\beta > 0$ then $(1 + \beta)(1 + \frac{\gamma}{\beta}) > (1 + \sqrt{\gamma})^2$.