Mathematical Methods in Data Science and Signal Processing

Homework Assignment 1

November 19, 2022

General instructions: Upload your solution text and code to Moodle via the dedicated submission box.

In the following, let $A = U\Sigma V^T \in \mathbb{R}^{m\times n}$. The ith columns of U and V, respectively, are u_i and v_i. The ith largest singular value of A is σ_i. The number of nonzero singular values of A is r.

1. Suppose we need to solve the linear system of equations $Ax \approx b$ in the least squares sense. Namely, we wish to minimize $\|b - Ax\|_2^2$. Prove that $x_{LS} = V\Sigma^\dagger U^T b = A^\dagger b$, where $Q^\dagger$ denotes the pseudo-inverse of Q.

2. Prove that the rank of a matrix equals to the number of its nonzero singular values.

3. Prove:

   a) range(A) = span$(u_1, \ldots, u_r)$;

   b) null(A) = span$(v_{r+1}, \ldots, v_n)$.

4. Prove that

$$\inf_{\substack{B\in\mathbb{R}^{m\times n}\\ \text{rank}(B)\le s}} ||A - B||_2 = \sigma_{s+1}.$$

5. (Computing PCA) Set $n = 1000$ and $p = 10$ and draw two random signals $x_1, x_2 \in \mathbb{R}^p$. Generate n observations of the form

$$y_i = Sa_i x_1 + b_i x_2,$$

for $a_i, b_i \sim \mathcal{N}(0,1)$ and fixed (scalar) S. Compute the ratio between:

a) the leading eigenvalue and second eigenvalue of the empirical covariance matrix;

b) the leading singular value and second singular value of the data matrix $[y_1, \ldots, y_n]$.

Repeat this experiment for S ranging from 1 to 10^6 and plot the curves of these two ratios against S in log-log scale. What are the slopes? What are the implications for PCA?

6. (High-dimensional PCA) Set $n = 2000$ and $p = 500$.

   a) Generate n vectors $x_1, \ldots, x_n \in \mathbb{R}^p$ drawn from $x_i \sim \mathcal{N}(0, I)$. Compute and plot the eigenvalues of the empirical covariance matrix. Graphically compare their distribution with the Marčenko-Pastur distribution.

   b) Simulate the spiked model for high-dimensional PCA with rank-one perturbation. What is the minimal SNR value β for which you see the signal popping out the noise bulk? Does it fit the theory?