Homework set 4 — APPM4720/5720, Spring 2016

Problem 1: Suppose that A is an $m \times n$ matrix of *approximate* rank k, and that we have identified two index sets $I_{\rm s}$ and $J_{\rm s}$ such that the matrices

$$\mathsf{C} := \mathsf{A}(:, J_{\rm s}) \tag{1}$$

$$\mathsf{R} := \mathsf{A}(I_{\rm s}, :) \tag{2}$$

hold k columns/rows that approximately span the column/row space of A. You may assume that C and R both have rank k (in other words, the index vectors $J_{\rm s}$ and $I_{\rm s}$ are not *very* bad). Then

$$\mathsf{A} \approx \mathsf{C}\mathsf{C}^{\dagger}\,\mathsf{A}\,\mathsf{R}^{\dagger}\mathsf{R},$$

and the optimal choice for the "U" factor in the CUR decomposition is

$$\mathsf{U} := \mathsf{C}^{\dagger}\,\mathsf{A}\,\mathsf{R}^{\dagger}.$$

Set $\mathsf{X} = \mathsf{C}\mathsf{C}^{\dagger}$.

(a) Suppose that C has the SVD

$$\underset{m \times k}{\mathsf{C}} = \underset{m \times k}{\mathsf{U}}\;\underset{k \times k}{\mathsf{D}}\;\underset{k \times k}{\mathsf{V}^{*}}.$$

Prove that $\mathsf{X} = \mathsf{U}\mathsf{U}^{*}$.

(b) Suppose that C has the QR factorization

$$\underset{m \times k}{\mathsf{C}}\;\underset{k \times k}{\mathsf{P}} = \underset{m \times k}{\mathsf{Q}}\;\underset{k \times k}{\mathsf{S}}.$$

Prove that $\mathsf{X} = \mathsf{Q}\mathsf{Q}^{*}$. (Observe that S is necessarily invertible, since C has rank k. You can then prove that $\mathsf{C}^{\dagger} = \mathsf{P}\mathsf{S}^{-1}\mathsf{Q}^{*}$.)

(c) Prove that X is the orthogonal projection onto $\mathrm{Col}(\mathsf{C})$.

(d) Suppose that A has precisely rank k and that C and R are both of rank k. Prove that then

$$\mathsf{C}^{\dagger}\,\mathsf{A}\,\mathsf{R}^{\dagger} = \big(\mathsf{A}(I_{\rm s}, J_{\rm s})\big)^{-1}.$$

Problem 2: Let A be an $n \times n$ matrix and suppose (i) that $\mathsf{A}(i,j) > 0$ for every i, j and (ii) that $\sum_{i=1}^{n} \mathsf{A}(i,j) = 1$ for every j (each column sums to one).

(a) Let p be a vector of non-negative numbers such that $\sum_{j=1}^{n} \mathsf{p}(j) = 1$. Set $\mathsf{p}' = \mathsf{A}\mathsf{p}$. Prove that $\sum_{j=1}^{n} \mathsf{p}'(j) = 1$. (In other words, the matrix A maps every probability distribution on the set $\{1, 2, \dots, n\}$ to another probability distribution.)

(b) Prove that A has an eigenvector with corresponding eigenvalue 1.

Problem 3: On the course webpage, you can download a file `testmatrices.mat` that holds three test matrices A, B, and C. Each matrix is of size $m \times 1\,000$ and contains a thousand samples from a multivariate normal distribution on $\mathbb{R}^m$. Use PCA to estimate the mean and the co-variance matrices of these distributions. It is sufficient to hand in your numerical answers. (The person doing the reference homework should also hand in code, and a brief description of the solution process.)