Homework 3

(Canvas online submission, **due time 11:59pm EST, Wednesday, Feb 23, 2021**)

1 **(EM for factor analysis)** In the derivation of the EM algorithm for factor analysis, it is assumed that $\mu = 0$ for simplicity. Please re-derive the algorithm with μ included. Summarize the algorithm in iterative E and M steps.

2 **(EM for Probabilistic PCA)** Derive the EM algorithm for estimating the model parameters of Probabilistic PCA. Summarize the algorithm in iterative E and M steps.

3 **(EM for Factor analysis)** Generate $n = 100$ observations of the 7-dim vector Y from the following factor analysis model:

$$Y_{7\times1} = \mu^*_{7\times1} + \Lambda^*_{7\times2} X_{2\times1} + W_{7\times1}$$

where X and W are independent, $\mu^* = 0_{7\times1}$ and

$$\Lambda^* = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 1 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{pmatrix}_{7\times2}, \quad X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}_{2\times1} \sim N_2(0, I_2), \quad and \quad W_{7\times1} \sim N_7(0, 0.4 \times I_7).$$

(a) Suppose we know that there are two factors in the factor analysis model. Implement your EM algorithm in Question 1 for estimating Λ, μ and Ψ. To evaluate the convergence of this algorithm, initialize Λ to be close to be the true value Λ^*: $\Lambda^{(0)} \sim N(\Lambda^*, \sigma^2 I)$, for varying choices of small σ's. Repeat the EM procedure multiple times to obtain the best possible optimization result. Report your findings.

(b) Since Λ can only be identified up to an orthornormal transformation, one need to use a suitable notion of distance to evaluate the convergence of the *column space* of Λ to the true column space (of Λ^*). Propose one such suitable distance measure (as did in Hw2), re-run the algorithm and report the behavior of your algorithm. (Now, you need to initialize Λ randomly without using the knowledge of Λ^* as in part (a))

(c) Implement your EM algorithm in Question 2. Suppose we know that there are two principal components in the Probabilistic PCA model. Compare your results with those obtained in part (a) and (b).

4 (**MDS)** The figure below depicts airline distances among 10 US cities. Please implement the classical MDS program to produce a 2-dimensional map for these cities.

CITIES	ATLA.	CHIC.	DENV.	HOUS.	L.A.	MIAMI	N.Y.	S.F.	SEAT.	WASH D.C.
ATLANTA		587	1212	701	1936	604	748	2139	2182	543
CHICAGO	587		920	940	1745	1188	713	1858	1737	597
DENVER	1212	920		879	831	1726	1631	949	1021	1494
HOUSTON	701	940	879		1374	968	1420	1645	1891	1220
LOS ANGELES	1936	1745	831	1374		2339	2451	347	959	2300
MIAMI	604	1188	1726	968	2339		1092	2594	2734	923
NEW YORK	748	713	1631	1420	2451	1092		2571	2408	205
SAN FRANCISCO	2139	1858	949	1645	347	2594	2571		678	2442
SEATTLE	2182	1737	1021	1891	959	2734	2408	678		2329
WASHINGTON D.C.	543	597	1494	1220	2300	923	205	2442	2329	

(B) AIRLINE DISTANCES BETWEEN TEN U.S. CITIES

5 **Kernel PCA.** On Canvas site there is an R file `nytimes.RData`. Load this into R workspace, and look for data frame `nyt.frame`. This data frame has 102 rows and 4432 columns. Each row corresponds to a New York Times article. The first column contains the class label of the article (art or music), the remaining columns contains the normalized word counts, one column for each word. You will see a lot of zeroes, because many words do not appear in a given article.

Perform Kernel PCA with polynomial and Gaussian kernels. Project each article onto a subspace with one, two or three dimensions. Be sure to include the class label (art or music) for each article. How many dimensions do you need to visually separate the two classes of articles reasonably well? How do the results compare with the PCA results in HW2? Report your findings.