Advanced Computational Statistics (STAT 7730)
Homework 2

This homework is due by **11:30 AM on Friday, September 20**. Be sure to explain your reasoning in your solutions, label your figures and attach the code at the end. When you submit your homework, you must bring it (hard copy) to our class.

1. (16 pts) Following the notation and the assumption in the lecture of the Principal Component Analysis, show
   (a) $\text{argmin}_{U\in R^{n\times d}:U'U=I_d} E(||X-UU'X||_F^2) = \text{argmax}_{U\in R^{n\times d}:U'U=I_d} E(||U'X||_F^2)$;
   (b) $E(||U'X||_F^2) = \sum_{i=1}^d u_i'Cu_i$ where u_i is i-th column of U and C is the covariance of X;
   (c) $\sum_{i=1}^d u_i'Cu_i = \sum_{j=1}^n \lambda_j \sum_{i=1}^d (\tilde{u}_j'u_i)^2$ where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$ are the eigenvalues of C and $\tilde{u}_j$ is the corresponding eigenvector (n-by-1) of λ_j. Therefore, $E(||U'X||_F^2) = \sum_{j=1}^d \lambda_j$ when we take $u_i = \tilde{u}_i$ for $i = 1, \cdots, d$;
   (d) $\sum_{j=1}^d \lambda_j = \max_{U\in R^{n\times d}:U'U=I_d} E(||U'X||_F^2)$.

2. (35 pts) **PCA for image patches.** Download the img.zip file, which contains 100 different 600×600 grayscale images of outdoor scenes. For this homework we will be working with 12×12 image patches. There are $100(600-11)^2$ different 12×12 image patches in this dataset. For $i_1, i_2 = 1, \ldots, 589$ and $k = 1, \ldots, 100$, you can get the patch, say V, starting at pixel (i_1, i_2) in k-th image I_k:

$$V = [V_{i,j}]_{i,j=1}^{12}, \quad \text{where } V_{i,j} = I_k(i_1 - 1 + i, i_2 - 1 + j)$$

   and you can convert it into a 144×1 vector, say $\vec{X}$, by connecting columns:

$$\vec{X} = (V_{1,1}, V_{2,1}, \cdots, V_{12,1},\ V_{1,2}, V_{2,2}, \cdots, V_{12,2},\ \cdots\cdots,\ V_{1,12}, V_{2,12}, \cdots, V_{12,12})^T.$$

   Estimate the mean vector $\vec{\mu}$ and covariance matrix C of converted image patches $\vec{X}$ using $n = 10^5$ randomly chosen image patches from this database. By randomly chosen, I mean that each of the $100(600-11)^2$ patches is equally likely to be chosen. Compute all 144 eigenvectors of C, say $\vec{e}(1), \ldots, \vec{e}(144)$, sorted so that the corresponding eigenvalues are decreasing, say $\lambda_1 \geq \cdots \geq \lambda_{144}$.

   (a) What are the range of values in the data (i.e., maximum and minimum values of pixel intensities of the 100 images)? What are the range of values in $\vec{\mu}$? So, in words, what does the mean image patch look like?
   (b) Create a figure with the first 36 eigenvectors (in order, from left-to-right, then top-to-bottom) displayed as 12×12 image patches. Note that the eigenvector entries are not from 0 to 255. When plotting each of the eigenvectors, rescale it by a linear map so that the minimum and maximum entries become 0 and 255, respectively. Include this plot with your homework. Do this for the next 36 eigenvectors, and so on, until you have seen all 144 eigenvectors. You do not have to attach these additional plots to your homework. In words, what are the dominant modes of variation in image patches? What are the least dominant?
   (c) Pick an image from the database. Break it into 50^2 disjoint 12×12 image patches. For $d = 20$, compute the d-dimensional PCA representation of each patch. Now you have a $20(50^2)$-dimensional version of the image which is about 7 times smaller than the original image. To visualize your PCA image, project the PCA representation of each patch back into 144-dimensional space (Don't forget to add the mean!), convert each (144-dimensional vector) back into 12×12, and glue the patches back into a 600×600 image. Plot the original image beside the PCA reconstruction. Comment on any differences.

(d) Repeat (c) for $d = 1, 3, 5, 10, 144$. Include plots for the cases $d = 1$ and $d = 144$ with your homework. Comment on your results for each choice of d.

3. (10 pts) Let $X_1, X_2, \cdots, X_n \sim$ i.i.d. $f(x|\theta)$, where

$$f(x|\theta) = p \cdot \phi(x; \mu_1, \sigma_1) + (1 - p) \cdot \phi(x; \mu_2, \sigma_2),$$

$\theta = (p, \mu_1, \mu_2, \sigma_1, \sigma_2)$, $p \in (0, 1)$ and $\phi(x; \mu_i, \sigma_i)$ is the density function of the normal distribution with mean $\mu_i \in \mathbb{R}$ and standard deviation $\sigma_i > 0$. [i.e. $f(x|\theta)$ is the density of the Gaussian mixture distribution.] Then, show

(a) θ is not identifiable for the family of distributions

$$\{f(x|\theta) \mid \theta = (p, \mu_1, \mu_2, \sigma_1, \sigma_2), p \in (0, 1), \mu_1, \mu_2 \in \mathbb{R}, \sigma_1 > 0, \sigma_2 > 0\}.$$

(b) The likelihood function $L(\tilde{\theta}) = \prod_{i=1}^{n} f(X_i|\tilde{\theta})$ can be arbitrarily large.