M532 Mathematical Modeling of Large Data Sets
Problem Set One

Due Friday, February 9, 2018

1. Consider the vector whose components are given w.r.t. the standard basis are

$$x = \begin{bmatrix} 0 \\ 0 \\ 1 \\ -1 \end{bmatrix}$$

Determine the coordinates of this point w.r.t. the Walsh basis

$$U_1 = \frac{1}{4} \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 \\ 1 & -1 & 1 & -1 \end{bmatrix}$$

and the Haar basis

$$U_2 = \frac{1}{2} \begin{bmatrix} 1 & 1 & \sqrt{2} & 0 \\ 1 & 1 & -\sqrt{2} & 0 \\ 1 & -1 & 0 & \sqrt{2} \\ 1 & -1 & 0 & -\sqrt{2} \end{bmatrix}$$

Which representation do you think is better for this point x and why? You may use MATLAB or another computer program to assist in your calculations.

2. Consider the two-dimensional subspace spanned by the vectors

$$U = \frac{1}{4} \begin{bmatrix} 1 & 1 \\ 1 & 1 \\ 1 & -1 \\ 1 & -1 \end{bmatrix}$$

Compute the projection of the point

$$x = \begin{bmatrix} 2 \\ 3 \\ 1 \\ -1 \end{bmatrix}$$

onto the subspace spanned by U. What is the novelty of this point? What is the representation of x in the 2D subspace w.r.t. the basis for the subspace?

3. Use the QR algorithm to compute an orthonormal basis from the vectors

$$\begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}, \begin{bmatrix} 1 \\ -2 \\ 3 \end{bmatrix}, \begin{bmatrix} 0 \\ -1 \\ -1 \end{bmatrix}$$

Verify that $Q^T Q = I$. Work through this problem by hand and verify via matlab.

4. Recall that the eigenvectors of symmetric matrices associated with distinct eigenvalues are orthogonal and that eigenvalues are real. Prove these facts. (Try without looking at your old linear algebra book.)

5. In class we showed that the optimization problem for the first principal vector is given by

$$\max \phi^T C \phi - \lambda(\phi^T \phi - 1)$$

leads to the necessary condition

$$C\phi = \lambda\phi$$

for the best eigenvector. Provide the details of this calculation.

6. Let $\{x^{(\mu)}\}$ be a data set in $\mathbb{R}^n$ and let $\beta = \{u^{(1)}, \ldots, u^{(N)}\}$ be a basis. Further, for any pattern $x^{(\mu)}$ write

$$x^{(\mu)} = \sum_{i=1}^{N} \alpha_i^{(\mu)} u^{(i)}$$

a) Show that if the data set has been mean subtracted then, for each i,

$$\sum_{\mu=1}^{N} \alpha_i^{(\mu)} = 0$$

b) Show that the statistical variance of $\alpha_i^{(\mu)}$ over μ for fixed i is the eigenvalue λ_i of the covariance matrix C.

1 Computing

Part B: Computing

Download the pumpkin data set. Compute a data matrix with each pumpkin being reshaped into a column vector to form a data matrix X. Note that the red, green and blue sheets of each pumpkin should all be stacked into a single vector. The resulting matrix X should have size 1036800 by 200.

Problem 1.

a) Compute the mean pumpkin and subtract it from the data. Does this average look like you would expect?

b) Compute the best basis using PCA on the mean subtracted data. Include pictures of eigenpumpkins 1-4 and 101-104. (Note that these eigenvectors need to be reassembled into RGB 3-way arrays with values scaled to be in the same form as the initial data, i.e., integers $\{0, \ldots, 255\}$.) Plot the eigenvalues. Comment. (Note that you will need to use the snapshot method since XX^T is too large.)

c) Repeat part b) using the thin svd, i.e., svd(X,0) and compare the results left singular vectors and squared singular values with the eigenvectors and eigenvalues from PCA.

d) Plot the coefficients of the pumpkin $(\alpha_1^{(\mu)}, \alpha_2^{(\mu)})$ with respect to the PCA basis.

Problem 2.

This project concerns the application of the KL procedure for incomplete data. Let the complete data set be translationally invariant:

$$f(x_m, t_\mu) = \frac{1}{N} \sum_{k=1}^{N} \frac{1}{k} \sin[k(x_m - t_\mu)],$$

where $m = 1, \ldots, M$, with M the dimension of the ambient space (size of the spatial grid), and $\mu = 1, \ldots, P$, with P the number of points in the ensemble. Let $x_m = (m-1)2\pi/M$ and $t_\mu = (\mu - 1)2\pi/P$. Each pattern in the incomplete ensemble may be written

$$\tilde{x}^{(\mu)} = m^{(\mu)}.f^{(\mu)},$$

where $(f^{(\mu)})_m = f(x_m, t_\mu)$. Let $P = M = 64$ and $N = 3$.

Part A. What is the rank of the data matrix? Provide a theoretical argument followed by an actual computation in Matlab.

Part B. Compute a best basis for this data. Does it look familiar?

Part C. Write a code to repair a single pattern with 50% missing entries using this basis. If you increase the prtecentage of missing data, how far can you go before you fail to reconstruct the pattern?

Part D. Write a code to repair the entire ensemble of corrupted patterns. Assume no good basis is available for the repair.

Part E. Comment on the rate of convergence of the algorithm as you vary the amount of missing data. Use the measure

$$E(k) = (\sum_{i=1}^{r} (\lambda_i(k) - \lambda_i(k-1)))^{1/2}$$

where r is the rank of the approximation of the gappy data, k is the iteration and the λ are the eigenvalues of the best basis. The algorithm should terminate for the smallest k such that

$$E(k) \leq 0.01$$