CH 5440 Multivariate Data Analysis in Process Monitoring and Diagnosis

Assignment 3

PCA for Model Identification

1. A flow process is given in Fig. 1. The flow rates of all streams are measured with identical flow meters which have the same accuracies. 1000 samples corresponding to different steady states are obtained and given in file flowdata.mat

(a) Based on Fig. 1, determine the number of linear constraints that exists among the flow variables for this process.

(b) Apply PCA to determine the eigenvalues and eigenvectors of the covariance matrix of the data. Apply a sequential hypothesis test procedure for testing equality of smallest k eigenvalues starting from the maximum possible number for k and graduating decrementing it until $k = 2$, in order to determine the number of constraints. Report the degrees of freedom, test statistic, and test criterion in the form of a table, and the estimated number of constraints. Are you able to determine the correct number of constraints based on the hypothesis test?

(c) Develop a variant of sequential hypothesis testing strategy as suggested by one of the students in the class. At each stage of the sequential procedure apply a hypothesis test to test only whether two of the neighbouring eigenvalues are equal or not. Report the degrees of freedom, test statistic, and test criterion in the form of a table, and estimated number of constraints. Are you able to determine the correct number of constraints based on this modified sequential hypothesis test strategy?

(d) Obtain an estimate of the steady state flow constraint matrix using PCA. Compare the estimated constraint matrix with the true constraint matrix using subspace angle. Estimate the variance of error in the measurements.

(e) Based on the estimated constraint matrix, identify all sets of variables that are poor choices of dependent variables. Explain how you identified these sets.

(f) From the estimated constraint matrix obtain the regression matrix relating flows of streams 1, 2 and 3 (chosen as dependent variables) to flows of streams 4, 5 and 6 (chosen as independent variables). Obtain the maximum absolute difference between estimated coefficients of regression and true regression matrix obtained from flow balances.

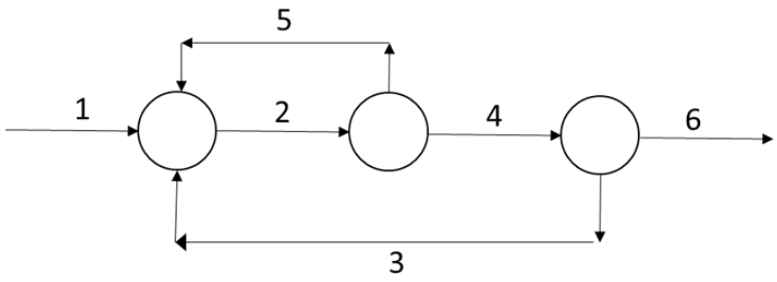

Figure 1

Principal Component Analysis for Data Compression/Denoising:

2. A zoologist obtained measurements of the mass (in grams), the snout-vent length (SVL) and hind limb span (HLS) in mm of 25 lizards. **The mean and covariance matrix of the data about the mean** are given by

$$\bar{x} = \begin{bmatrix} 9 \\ 68 \\ 129 \end{bmatrix} \qquad S = \begin{bmatrix} 7 & 21 & 34 \\ 21 & 64 & 102 \\ 34 & 102 & 186 \end{bmatrix}$$

(a) The largest eigenvalue of the above covariance matrix is 250.4009. Determine the normalized eigenvector corresponding to this eigenvalue. Also determine the remaining eigenvalues and corresponding mutually orthogonal eigenvectors.

(b) How many principal components should be retained, if at least 95% of the variance in the data has to be captured?

(c) Assuming that there are two linear relationships among the three variables, determine one possible set of these linear relations.

(d) Using the PCA model, determine the scores for a female lizard with the following measurements: mass = 10.1 gms, SVL = 73mm and HLS = 135.5mm.

(e) Using the PCA model, estimate the mass of a lizard whose measured SVL is 73mm

(f) Using the PCA model, estimate the mass of a lizard whose measured SVL is 73mm and measured HLS is 135.5 mm.

Note: This problem has to be solved manually (you can verify the results using a computer code

Application of PCA in image processing

3. Yale faces data set is an image data set of 15 subjects that is used for testing facial recognition methods. There are six different grayscale images of each subject with different facial expressions. The resolution of each image is 243 x 320 pixels.

(i) Convert each image into a column vector of size 77760 x 1 by stacking the columns of the digital image one below the other (also known as the vec operation on a matrix). Using the normal images of each subject as the standard, determine how many of the other images you are able to correctly identify by using Euclidean distance between the image and the standard images.

(ii) For each subject determine a weighted combination of the six images that best represents the subject (Hint: Use the first PC). The representative images for each subject can be stored in database. Determine how many of the original images you are able to correctly classify based on comparison with the representative images stored in database.

(iii) Instead of storing just one representative image, you can store two representative images for each subject to improve identification accuracy. Determine how many original images you ae able to correctly classify using two representative images for each subject.