DEPT. OF CHEMICAL ENGINEERING

CH 5440 MULTIVARIATE DATA ANALYSIS

TAKE HOME END SEMESTER EXAM
Issue Date 09/05/14 at 12:00 noon
Due Date 10/05/14 before12:00 noon

INSTRUCTIONS

1. This is a take home exam. **YOU ARE EXPECTED TO WORK ON THIS ON YOUR OWN WITHOUT CONSULTING ANY OTHER LIVING BEING**. You are free to consult your notes, text books, research papers, internet for solving the problems.
2. You are encouraged to use MATLAB or SCILAB to solve the problems. Submit your answers summarizing the results along with explanations. If you are using a package downloaded from Internet, mention the package and website. Explain how you used the package along with the parameter choices you used. If you have written you own MATLAB/SCILAB code, attach a print-out of the code (should be well documented) along with your submission. The solutions should be uploaded on moodle no later than **12 noon on Saturday, May 10, 2014**.
3. Sign the declaration below and hand over a hard copy of the declaration (first page) at my office.

DECLARATION

I HAVE NEITHER ASSISTED NOT TAKEN ANY ASSISTANCE FROM ANYONE TO SOLVE THIS TAKE HOME EXAM PAPER.

Name and Roll No:

Signature:

Date:

1. Figure 1 shows the flow network of a simple water distribution system. A sample of 1000 measurements corresponding to water flow rates in all streams have been generated and stored in file *WDNdata.mat*.

(a) Obtain the constraint matrix corresponding to the steady state flow balances for the given network using first principles.

(b) Apply PCA to obtain an estimate of the constraint matrix relating the measured variables. For this network, what is the actual number of PCs you should retain? Determine whether the number of PCs to be retained is estimated correctly using (i) 95% of variance explained (ii) SCREE plot (iii) maximum incremental reduction in RMSE obtained as you systematically increase the number of PCs to be retained. RMSE is defined as

$$\sqrt{\frac{1}{N}\sum_{i=1}^{N}(\mathbf{z}_i - \hat{\mathbf{z}}_i)^T(\mathbf{z}_i - \hat{\mathbf{z}}_i)}$$

where *N* is the number of samples. The incremental reduction in RMSE is the reduction in RMSE obtained when you increase the number of retained PCs from k to k + 1.

(c) Suppose we choose the flow variables 1, 2 and 3 to be the independent flow variables. From your estimated constraint matrix determine the regression matrix relating the dependent flows to the independent flow variables. Also determine how good the estimated regression matrix is as compared to the true regression matrix by finding the maximum absolute difference among all elements of the estimated and true regression matrix.

(d) If you are not given the flow network and asked to choose a set of independent flows based on your estimated constraint matrix only, what is the best choice of independent flow variables you would recommend and why?

(e) Using the PCA model, determine the de-noised estimates of variables for the last sample in the data set *WDNdata.mat*.

2. For the process in Fig. 1, assume that the flows of streams 2 and 6 are not measured while all the rest are measured. Delete these measurements of these variables from the given data set *WDNdata.mat* and apply PCA to the remaining measurements to estimate the constraint matrix relating the measured variables. Justify based on physical arguments how many PCs you will retain. For the given process, determine the true constraint matrix relating the measured variables and determine how well the constraint matrix is estimated using subspace angle between the row spaces of estimated and actual constraint matrices.

3. For the process in Fig. 1, assume that the measurements of stream x is perfect (noise free). The standard (vanilla) PCA assumes all measurements are noisy and are equally inaccurate. Develop a modified PCA approach, which can be used when measurements

of one or more variables are specified as perfect. Describe the steps of your algorithm clearly. Use this method to determine the de-noised estimates of the last sample in the data set *WDNdata.mat*.

4. For the process in Fig. 1, it is known that PCA can be used to obtain a constraint matrix which will, in general, be a linear combination of the rows of the actual constraint matrix obtained from first principles using flow balances around individual nodes. If we wish to obtain a desired form of the constraint matrix, we can specify the zero elements of the constraint matrix.

(a) For the given process identify which elements of the constraint matrix can be specified as zero.

(b) What is the minimum number of elements of the constraint matrix that must be to be specified as zero so that the desired form of the constraint matrix can be estimated uniquely from measurements upto a scale factor?

(c) Develop a modified PCA approach for determining the desired form of the constraint matrix, which will ensure that estimates of certain specified elements of the constraint matrix are zero (assuming that necessary conditions for uniqueness are satisfied).

(d) Apply your method to estimate the constraint matrix from the given measurements *WDNdata.mat* so that it matches with the constraint matrix derived from first principles upto a scale factor (that is the rows of the estimated constraint matrix can be a scale of the actual constraint matrix).

5. The specific heat capacity of pure species is generally described by a polynomial function of temperature. The specific heat of gaseous carbon dioxide obtained from experiments for a temperature range between 175 K and 3000K is given in Table 1. Using MATLAB or SCILAB write a code for developing a nonlinear model between specific heat and temperature using Kernel PCR. For this purpose scale the temperature data using the standard deviation of temperature measurements and use a polynomial Kernel of order 10. Use leave one sample out cross validation to pick the optimum number of PCs in feature space. Use the developed model to estimate the specific heat capacities at the following temperatures: 250 K, 500K, 1000K, 2000K, and 6000K, and report the results in the form of a Table.

6. The Fisher iris data set is a standard data for testing classification algorithms. It consists of data for three different types of iris flowers. The data contains type, petal width (PW), petal length (PL), sepal width (SW), and sepal length (SL) for 150 iris samples. The lengths are measured in millimeters. Type 0 is *Setosa*; type 1 is *Verginica*; and type 2 is *Versicolor*.

(a) Write a MATLAB/SCILAB code for K means classification and apply it to the Fisher raw measurements of iris data set. Assume that the number of clusters is three and

determine the number of misclassifications. Choose the first 100 samples for classification and the rest as a test set for determining the number of misclassifications.

(b) Apply PCA to first obtain the scores and apply the K-means algorithm on the scores. Determine the number of misclassifications and determine the optimum number of PCs for minimizing the number of misclassifications on the test set.

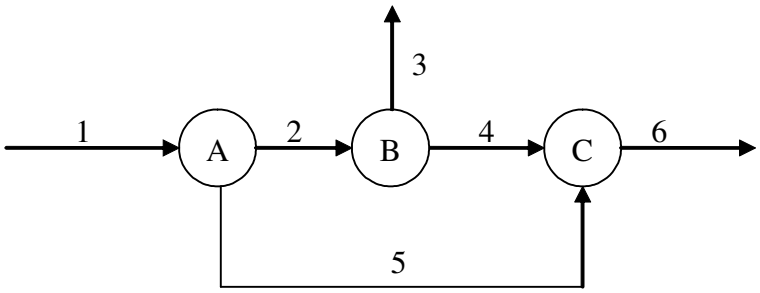

Figure 1 Flow process of simple water distribution network

Data for problems 5 and 6 are given in Excel file endsem14data.xlsx.

Note: Problems 2, 3 and 4 are comparatively more difficult. Also for problems 3 and 4, the modified PCA algorithm can consist of (a) a data pre-processing step, (b) iterative use of PCA and (3) a post-processing step or a combination of these. Essentially, PCA should be a component of the proposed algorithm.