DEPT. OF CHEMICAL ENGINEERING

CH 5440 - MULTIVARIATE DATA ANALYSIS

END SEMESTER EXAM: TIME: 09:00-12:00 hrs DATE: 11/05/24

INSTRUCTIONS

1. Open book and notes exam. You are free to consult your notes, text books, research papers, internet for solving the problems. You can reuse the codes you have developed for solving the assignments.
2. You are encouraged to use MATLAB or Python to solve the problems. Attach a print-out of the code (should be well documented) along with your submission.
3. **The answers should be provided in a separate word or pdf document. The document along with the MATLAB or PYTHON codes should be uploaded as a single zip file. Name the zip file as rollnumber.zip. Name each MATLAB/PYTHON file as Qxx.m or Qxx.ipynb.**

Before you begin to answer the questions

For the three problems MATLAB codes are given which should be used to generate individualized data sets. These MATLAB codes are respectively *generate_steam_network_data.m*, *generateNonIdealVLE.m* and *generatedynamic.m*. The data files that will be generated by these codes are respectively *steamdata.mat*, *VLEdata.mat* and *arx.mat*. **Before executing each of these codes change the random number used in line 4 to your roll number. For example if your Roll number is CH20B025 then the random number to be used is 20025, that is, line 4 should be changed to *rng(20025,'twister')*.**

1. Figure 1 shows the steam distribution network for a chemical plant. A sample of measurements corresponding to a subset of the 28 streams have been generated and stored in file *steamdata.mat*. The true values of the flows are stored in variable Ftrue and the corresponding measured values in variable Fmeas. The subset of flow variables that are measured is given in variable Fmindx. The variance of noise in all measured flows are identical.

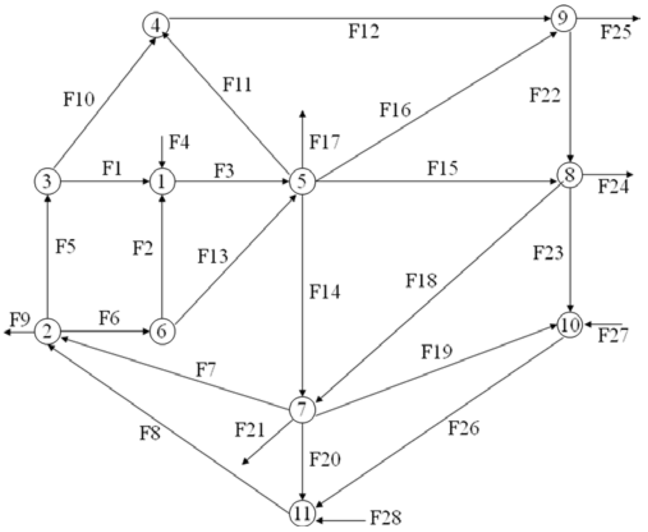

Figure 1

(a) Use the **true** values of **flows** to determine the number of constraints and the true constraint matrix relating the measured subset of flow variables. Report these giving reasons for your answers. (5 points)

(b) Use the true constraint matrix obtained in part (a) to choose a valid set of dependent variables and report them, along with an explanation for your choice. Determine the true regression matrix relating the dependent flows to the independent flow variables for your choice. (10 points)

(c) Apply PCA to the **measured flow data** and estimate the number of constraints as well as the constraint matrix. Justify your selection of the number of constraints based on hypothesis testing for equality of smallest eigenvalues for a significance level of 0.05. Report the test statistics and test criterion for different guesses of number of constraints between 12 and 2 in the form of a table. Report the estimated number of constraints and estimated error variance. (10 points)

(d) Assume that the true number of constraints is known (as obtained from part (a)). For the same choice of dependent variables as chosen in (b) estimate the regression matrix relating the dependent to the independent flow variables using the estimated constraint matrix in part (c). Report the maximum absolute difference between the estimated and true regression matrix elements. (5 points)

(e) Based on Fig. 1 and the subset of measured variables, construct the flow network which will have the same constraint matrix as obtained in part (a). Describe the procedure for obtaining this network, and a quantitative metric for verifying whether your network is correct. (10 points)

2. File VLEdata.mat contains the measured values of dew point temperature and the mole fraction of species 1 in the vapour phase for a binary system, corresponding to different input temperatures and liquid compositions. The variable temp contains the different temperatures and variable liqmf contains the different mole fraction of species 1 in liquid phase at which experimental measurements are made. Variable Pmeas(i,j) contains the measured dew point pressure corresponding to temp(i) and liqmf(j). Similarly variable y1meas(i,j) is the measured mole fraction of species 1 in vapour phase corresponding to temp(i) an liqmf(j). It is desired to build nonlinear models to predict dew point pressure and mole fraction of species 1 in vapour phase as a function of temperature and liquid phase composition.

(a) From the given data construct the input data matrix where each row corresponds to a sample and each column corresponds to an input variable. Similarly construct the corresponding output data matrix. Divide the data set into a training set containing about 70% of the samples and the rest for cross validation. (10 points)

(a) Use Kernel PCR to develop a nonlinear model for predicting dew point pressure as a function of temperature and liquid phase composition. Use Gaussian Kernels with widths in range [1, 100] and use cross validation to select the optimum number of PCs and width that gives best model predictions. (10 points)

(b) Use Kernel PCR to develop a nonlinear model for predicting vapour phase composition as a function of temperature and liquid phase composition. Use Gaussian Kernels with widths in range [1, 100] and use cross validation to select the optimum number of PCs and width that gives best model predictions. (10 points)

3. File arx.mat contains 1024 samples of noisy input and output measurements for a linear discrete dynamic system, whose order is less than or equal to 5. The standard deviation of errors in input and output measurements are unequal and are given in variables stdeu and stdey, respectively. It is desired to identify an ARX model of the process from the data.

(a) Apply dynamic PCA (after appropriately scaling the measurements) using two different lags between 10 and 20. Use hypothesis test to estimate the order of the system for each of the lags chosen and report your result. Are you able to obtain a consistent estimate of the model order? (10 points)

(b) Using the estimated order obtained in (a) apply last PCA to estimate the dynamic model coefficients. Obtain the confidence intervals for the model parameter estimates using bootstrap technique (100 bootstrap samples choosing a random sample of 70% of the lagged data matrix in each bootstrap sample). Report the 95% confidence intervals for each model coefficient estimate in the form of a table, clearly indicating the variable (along with its time index) corresponding to each coefficient. (10 points)

(c) The generalized eigenvalues and generalized eigenvectors of a matrix S_z of dimension p with respect to a square matrix Σ_e of same dimension are defined as

$$S_z v = \lambda \Sigma_e v \qquad (1)$$

It can be proved that the above generalized eigenvalues and generalized eigenvectors are the same as the eigenvalues and eigenvectors of the covariance matrix of scaled data S_{z_s} where

$$S_{z_s} = L^{-1} S_z L^{-T} \qquad (2)$$

Note that if error variances are not identical for all variables, we have to determine the eigenvalues and eigenvectors of covariance matrix of scaled data. The above result implies that these can equivalently be obtained by determining the generalized eigenvalues and generalized eigenvectors as defined in Eq. (1). The advantage of using the generalized eigenvalues and eigenvectors is that they can be used even if Σ_e is positive semi-definite (that is even if some of the error variances are zero).

We wish to identify a linear dynamic model using the noisy measured outputs (variable ymeas) and noiseless inputs (given by variable utrue). As in LPCA define a lagged vector of noisy outputs and noiseless inputs, and use the lagged data matrix along with the idea of generalized eigenvalues/eigenvectors to determine the model coefficients. For this purpose assume that you know the order of the system as estimated from (a). In MATLAB the function qz can be used to obtain generalized eigenvalues/eigenvectors, whereas in Python, you can use the function scipy.linalg.eig or scipy.linalg.eigh with appropriate inputs.

(10 points)