**CH 5440 Multivariate Data Analysis in Process Monitoring and Diagnosis**

Assignment 4

1. Model identification using PCA

Consider the steam distribution network for a methanol synthesis plant shown in Fig. 1. The process consists of 28 steam flows the flow rates of all of which are measured. A data set (steamdata.mat) consisting of 1000 samples corresponding to different steady states have been obtained.

(a) Apply IPCA to the above data set in order to simultaneously identify the linear constraint model relating the variables as well as the error covariance matrix, assuming that the error covariance matrix is diagonal. For this purpose choose different values for number of constraints, and report the eigenvalues corresponding to each choice. Are you able to identify the correct number of constraints for this process based on the eigenvalues? Justify your answer.

(b) The first seventeen variables for this process forms an independent set of variables. Obtain the regression model between the dependent and independent variable. Report the maximum absolute difference between estimated regression model coefficients and true regression model coefficients.

2. Missing data imputation

For the process given in problem 2, a data set of 1000 samples corresponding to different steady states have been obtained and given in data file steamdatamiss.mat. Unfortunately, due to sensor problem some of the measurements are missing (indicate by Nan for the corresponding value).

(a) Eliminate samples corresponding to missing values and apply PCA on auto-scaled data to determine the constraint model for the above process. Assume the actual number of constraints is known for this process.

(b) Use mean imputation for missing values and apply PCA on auto-scaled data to determine the constraint model

(c) Use the PCA model to impute missing valves and iterate until convergence to estimate the constraint matrix

(d) Repeat (a), (b) and (c) using IPCA instead of PCA to estimate the model and error variances simultaneously.

Report your results in the form of a table indicating method and maximum absolute difference between estimated and true regression matrix coefficients for the independent set of variables given in problem 2. Are you able to identify the number of constraints precisely using IPCA in all cases? Justify.