CH 5440 Multivariate Data Analysis in Process Monitoring and Diagnosis

Assignment 3

1. Multivariate calibration model using PCA

Twenty six samples of different concentrations of a mixture of Co, Cr, and Ni ions in dilute nitric acid were prepared in a laboratory and their spectra recorded over the range 300-650 nm using a HP 8452 UV diode array spectrophotometer (data in Inorfull.mat). Five replicates for each mixture were obtained. The measurements were made at 2nm intervals giving rise to an absorbance matrix of size 130 x 176. The concentrations of the 26 samples, which is a 26 x 3 matrix. In order to predict the concentration of the mixture using absorbance measurements, it is necessary to build a calibration model relating concentration of mixtures to its absorbance spectra. According to Beer-Lambert's law the absorbance spectra of a dilute mixture is a linear (weighted) combination of the pure component spectra with the weights corresponding to the concentrations of the species in the mixture. The quality of the linear calibration model is evaluated using leave-one-sample-out validation and computing the root mean square error (RMSE) in predicting the left out sample concentrations. Randomly pick one out of the five replicates for each mixture to obtain a data matrix of size 26 x 176. Apply each of the following approaches on the selected data set and report the RMSE in the form of a Table for each of the following cases.

(a) Build a linear calibration model using OLS assuming concentration measurements are free of error. However, when using the calibration model for predicting concentrations of a new mixture given its absorbance spectra, we need to assume that the absorbances have error and use OLS to predict concentrations.

(b) Build a calibration model by using PCR. In order to build the calibration model, PCA is first used to reduce (and also de-noise) the absorbance data to a full column rank scores matrix and then OLS is used to relate the concentrations to the scores (assuming scores are free of errors).

(c) Use PCR as in (b) to develop a calibration model except that in step 2, use OLS to relate scores to concentrations, assuming concentrations are free of errors.

(d) For methods in (b) and (c) determine and report the RMSE in the form of a Table for different choices of number of PCs chosen from 1 to 6 in the first step. Is the minimum RMSE obtained for when number of factors chosen is equal to number of species? If not can you give reasons for this anomaly?

(e) Do your results improve if you apply the methods on the averaged values of the five replicates for each mixture instead of randomly selecting one out of the five replicates? If so, give reasons.

(f) The absorbances are very noisy near the ends of the instrument. Wavelength selection can perhaps improve the quality of the multivariate calibration model. Repeat methods (b) and (c) using absorbances in the wavelength region from 350 nm to 600 nm and check whether the predictions improve. Report your conclusions with justifications.

2. Model identification using PCA

Consider the flow process shown in Fig. 1 consisting of five streams, the flow rates of all of which are measured. Two data sets (flowdata1.mat) and (flowdata2.mat) consisting of 1000 samples corresponding to different steady states have been obtained. In data set 1, the signal to noise ratio (SNR) in all flow rates is high, whereas in data set 2 the SNR is low.

(a) Apply PCA to the above data sets in order to identify the linear constraint model relating the variables (assuming that you know that the number of linear relations that exist between variables). In order to verify whether your constraint model is good, choose a set of independent flow variables and obtain the relationship between the dependent and independent variables (regression form of the model) using your estimated constraint model and find the maximum absolute difference between estimated regression model coefficients and true regression model coefficients for your choice of independent variables.

(b) Plot the singular values for both cases and check if you can identify the correct number of linear relations from the singular values.

(c) Apply PCA after scaling the measurements using the standard deviation of the measurements (also known as autoscaling) and identify the linear steady state model relating the flow variables. Does autoscaling lead to a better estimate of the linear model (provide justification).

(d) From the constraint model identified in suggest a procedure (a measure) by which you can determine a set of independent variables for the process. Determine the best and worst possible set of independent variables for this system based on your proposed measure and justify whether these inferences (obtained from data) are consistent with the physical process.

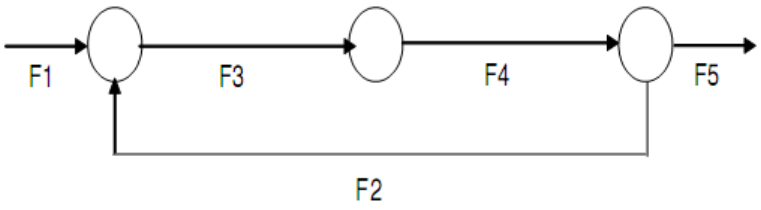

Fig. 1. Schematic of a flow process

Present a summary of your results in the form of a Table for each data set (method applied without scaling/with scaling, choice of independent variables, singular values, regression model, maximum absolute difference between estimated and true coefficient).