CH 5440 Multivariate Data Analysis in Process Monitoring and Diagnosis

Assignment 3

1. Multivariate calibration model using PCA

Multivariate calibration of spectral measurements is a technique that is used in chemometrics to develop a model relating spectral measurements (obtained using instruments such as UV, FIR or NIR or MS spectrophotometers) to properties such as concentration or other properties of species (usually liquid or gases). The application we consider is to obtain a model relating UV absorbance spectra to compositions (concentrations) of mixtures. Such a model is useful in online monitoring of chemical and biochemical reactions.

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). (Water and ethanol are generally used as solvents since these do not absorb in the UV range. Also the nitrate ions do not absorb in the UV range. So an aqueous solution of nitric acid is used to dissolve the metals in this experiment). Five replicates for each mixture were obtained. The measurements were made at 2 nm 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 are also given in the data file. 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.

Principal Component Regression is a method that can be used to develop a multivariate calibration model. In this method PCA is first applied to the absorbance matrix to obtain the scores corresponding to different mixtures. In the second step, a regression model is used to relate the concentrations to the scores using OLS (assuming concentrations are the dependent variables). In order to use this model for predicting the concentrations of a mixture whose absorbance spectra is given, we first obtain the scores and then use the OLS regression model to predict the concentrations. Note that the true rank of the absorbance matrix is equal to the number of 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. Pick the first sample 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 calibration model by PCR (Note that there is no offset term and evaluate the RMSE for different choice of number of PCs selected in step 1 of PCR. Report the RMSE results in the form of a table for different number of PCS chosen. Are you able to estimate the number of species correctly using LOOCV?

(b) Apply PCR on the averaged (over the five replicates) absorbance measurements. Do your results improve? If so, give reasons.

(c) The absorbances are very noisy near the ends of the instrument. Estimate the standard deviation of errors in absorbance measurements using the five replicates for each wavelength and for each mixture. Assume that the error standard deviations vary significantly with respect to wavelength but are almost same for all mixtures (verify this by plotting the estimated standard deviations wrt wavelength and mixtures). Therefore, obtain the average standard deviation or errors with respect to each wavelength. Use these standard deviations to scale the absorbance measurements for each wavelength before applying PCA in the first step. Do your results improve? If so give reasons (this method is also known as Maximum Likelihood PCR).

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.

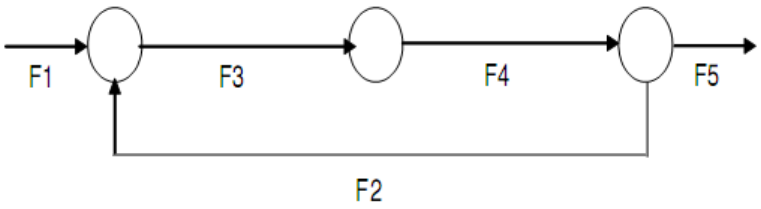

Fig. 1. Schematic of a flow process

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).