CH 5440 Multivariate Data Analysis in Process Monitoring and Diagnosis

Assignment 3

1. 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. A data set (flowdata3.mat) consisting of 1000 samples corresponding to different steady states have been obtained.

(a) Apply PCA 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, use thf following two measures

(i) Choose F3 and F5 as independent 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 (maxdiff) between estimated regression model coefficients and true regression model coefficients.

(ii) Determine the subspace angle between the row space of true constraint matrix and estimated constraint matrix.

Report the eigenvalues, maxdiff value and subspace angle

(b) For the same data, apply IPCA to estimate diagonal error variances and identify the linear steady state model relating the flow variables (assuming that you know that the number of linear relations that exist between variables). For this purpose use the function *stdest.m*, which uses a least squares procedure for estimating diagonal error variances. This function takes the constraint model and data as inputs and returns the standard deviation of errors as output. Report the final estimated error variances, eigenvalues, maxdiff value and subspace angle.

(c) Apply IPCA assuming incorrectly that there are four constraints. Report the error variances and eigenvalues obtained? Are you able to determine from the eigenvalues that the number of constraints has been incorrectly guessed? Give reasons for your answer.

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

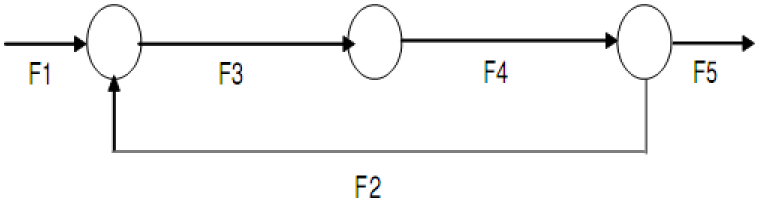

Fig. 1. Schematic of a flow process

2. IPCA method for simultaneously estimating model and non-diagonal error variance matrix

For problem 1, it is known that the errors in measurement of variables F3 and F4 are correlated. In this case, the error covariance matrix is not diagonal, but also contains an off-diagonal element corresponding to the element (3, 4) and (4,3) in the error covariance matrix. Modify the *stdest.m* function to estimate the diagonal and off-diagonal elements of the error covariance matrix and use it to extend the IPCA method to simultaneously estimate the constraint model and non-diagonal error covariance matrix. Report the converged eigenvalues, error variances, maxdiff value and subspace angle.

3. 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.

If absorbances are measured only a minimum number of wavelengths, then OLS can be used to build a calibration model. For example, if a mixture containing n_s non-reacting species, then absorbances at n_s wavelengths need to be measured. Typically, the wavelengths are chosen corresponding to the maximum absorbing wavelengths of individual species. However, if we measure absorbances at $n_w > n_s$ wavelengths, then the absorbance matrix will not be full column rank. In this case, Principal Component Regression 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 cross-validation (LOOCV) 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.

(a) Using the pure component spectra, identify the wavelengths at which the pure species have maximum absorbance. Denote them as λ_{max} values and report them. Build a calibration model using OLS by selecting the mixture absorbances at the λ_{max} wavelengths. Report the RMSE obtained using LOOCV.

(b) Develop a multivariate calibration model using PCR. Evaluate the RMSE for different choices of number of PCs (from 1 to 5) selected in step 1 of PCR. Report the LOOCV 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 RMSE?

(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 for different choices of PCs (1 to 5) chosen. Determine the RMSE using LOOCV. Do your results improve? Are you able to correctly determine the number of species using LOOCV? (This method is also known as Maximum Likelihood PCR).

(d) Use IPCA to estimate the error variances with respect to wavelength in step 1 of PCR and use it to simultaneously develop the calibration model. Again determine RMSE for different values of PCs chosen in first step. Are you able to estimate the number of species by eigenvalue analysis in first step? Are you able to estimate number of species correctly using LOOCV RMSE?