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

Assignment 4

Multivariate calibration model using Scald PCR & MLPCR

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 aqueous 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 this spectral 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 (concentration units are expressed in moles/lit or millimoles/lit etc.)

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 $\lambda_{max}$ values and report them. Build a calibration model using OLS by selecting the mixture absorbances at the $\lambda_{max}$ wavelengths. The RMSE can be computed for each species separately and the total RMSE can also be computed. 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 10) selected in step 1 of PCR.

(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 10) chosen. Determine the RMSE using LOOCV.

(d) Assume that the error variances vary with respect to both mixtures and wavelength. Use the replicate measurements for each sample to estimate the error variances for all data elements of the absorbance data matrix. Assume that the errors in different data elements are uncorrelated. Use Maximum Likelihood PCA developed by Wentzell et al. (1997) to determine the calibration model and RMSE of the model using LOOCV for different choices of number of PCs (1 to 10). For this purpose you should code the MLPCA method based on alternate least squares method described by Wentzell et al. in their paper.

For cases (b) to (d) provide the RMSE results as a function of number of PCs chosen for different methods. Which method results in the best calibration model? Are you able to correctly determine the number of species using LOOCV?