DEPARTMENT OF CHEMICAL ENGINEERING
CH5440 - MULTIVARIATE DATA ANALYSIS
END SEM 02/05/17 9:00AM-1:00 PM

**Answer all questions. State any additional assumptions you make. For each question describe the steps of the procedure you are implementing as a MATLAB code. Submit the procedure and answers in the answer sheets provided and upload the MATLAB codes as a zip file in Moodle. Name each file as rollnumberQxx.m and make a zip file of all relevant codes.**

1. A model to predict the glass transition temperatures of polymers using 11 properties of the monomer unit that is used to make the polymer has to be developed. File *tg.mat* contains the input properties of monomer and the measured glass transition temperatures for 57 polymers.

(a) Assuming that all input variables are independent, use OLS to obtain the regression model for predicting the glass transition temperature. For this purpose mean shift and auto-scale the data before applying PCA, and use the model to predict the glass transition temperatures. Report (i) the intercept term in regression model and (ii) the RMSE between measured and predicted values. (5 points)

(b) The input variables may not be linearly independent, therefore use PCR to develop the regression model. For this purpose mean shift and auto-scale the inputs variables before applying PCA and select the number of PCs to capture 95% of the variance in inputs. Use the PCR model to predict the glass transition temperatures. Report (i) the number of PCs retained (ii) the highest three eigenvalues of correlation matrix of inputs (iii) the intercept term in regression model and (iv) RMSE between measured and predicted values of glass transition temperature. (10 points)

(c) In order to account for differences in error variances of all variables, use IPCR to develop the regression model. Use the given function stdest.m to estimate the error variances in input variables. Assume the number of independent variables to be the same as estimated in (b) above. Report (i) the highest three singular values of scaled data matrix, (ii) the intercept term in regression model and (iii) RMSE between measured and predicted values of glass transition temperatures. (15 points)

(d) The relation between glass transition temperatures and input variables may not be linear. Apply KPCR to obtain a nonlinear regression model between glass transition temperature and inputs variables. For this purpose, use a polynomial kernel given by

$$K(\mathbf{x}_i, \mathbf{x}_j) = (1 + \mathbf{x}_i^T \mathbf{x}_j)^p$$

Mean shift and auto-scale all the input measurements to make them dimensionless before applying the kernel transformation. Obtain the optimum value of order $p$ (between 1 and 5) and optimum number of PCs to be selected (between 1 and 10) using leave one sample

out cross validation. Report the optimum value of $p$, optimum number of PCs and corresponding PRESS value obtained. (20 points)

**Note: Matlab function eig for computing eigenvalues and eigenvectors of a square matrix arranges the eigenvalues from smallest to largest (unlike svd).**

(2) We wish to factorize a matrix **Z*: n x N** as a product of two lower rank matrices as follows

**Z* = AP**

where **A** is a n x p matrix and **P** is a p x N matrix both of rank p (p < n, N). The structure of matrix **A** (the location of zero and non-zero elements of **A**) is also specified, and **A** is also NCA compliant. A fast NCA algorithm has been proposed for estimating the matrices **A** and **P** up to a scale factor. The algorithm estimates the non-zero elements of **A** column by column as follows:

Step 1: Perform economical svd($Z^*$) = $U_1 S_1 V_1^T + U_2 S_2 V_2^T$ where $U_1$ and $V_1$ are the singular vectors corresponding to the first p largest singular values. Let $W^* = U_1$ : n x p matrix. Note $W^* = Z^* V_1 S_1^{-1} = APV_1 S_1^{-1} = AQ$. So to estimate A, we can work with $W^*$ instead of the data matrix.

Step 2. Rearrange rows of **W*** such that

$$\begin{bmatrix} W_c^* \\ W_r^* \end{bmatrix} = \begin{bmatrix} a_1 & A_c \\ 0 & A_r \end{bmatrix} \begin{bmatrix} q_1^T \\ Q_r \end{bmatrix}$$

where $\mathbf{a}_1$ is the k x 1 vector of non-zero elements in first column of **A** and $\mathbf{A}_c$ : k x (p-1) and $\mathbf{A}_r$ : (n-k) x (p-1) are the remaining columns of **A** after appropriate rearrangement. $\mathbf{q}_1$ : 1 x p is the first row of **Q** and $\mathbf{Q}_r$ : (p-1) x p are the remaining rows of **Q**.

Step 3. Consider the last n-k rows of rearranged data matrix $W_r^* = A_r Q_r$.

Step 4. Find the projection matrix $S$ such that $Q_r S = 0$. Since $Q_r$ has rank p-1, $S$ can be obtained by performing an economical svd of $W_r^*$ and choosing the last column of $V$ corresponding to zero singular value.

Step 5. Compute $W_c^* S$. Note that $W_c^* S = a_1 q_1^T S + A_c Q_r S = a_1 \tilde{q}_1^T$ where $\tilde{q}_1^T = q_1^T S$ is a 1 x n row vector. This implies that $W_c^* S$ has rank 1.

Step 6. Perform svd of $W_c^* S$ and estimate $a_1$ up to a scale factor using the first column of $U$ corresponding to the only (largest) nonzero singular value.

Step 7. Repeat above procedure to find non-zero elements of each column of **A** one column at a time.

Step 8. Rearrange non-zero elements of each column to obtain A corresponding to the given structure and use OLS to get $P = \left(A^T A\right)^{-1} A^T Z^*$

**Note that if the measurements of $Z^*$ contains noise, in Step 3 other than the largest singular value the remaining singular values will not be zero but hopefully will be small.**

(a) Implement the fast NCA code as a function which returns A and P given the data matrix Z, structural matrix Astruct and rank p. The function should have the following format. (20 points)

function [A P] = fastNCA(Z, Astruct, p)

The structural matrix Astruct contains 1 in the location of non-zero elements and zero otherwise.

Use the following helper functions that have been provided for ease of implementation.

function [$Z_c$ $Z_r$] = rearrange(Z, Astruct, k) returns the row rearranged data matrices corresponding to non-zero elements and zero elements of column *k* of Astruct.

function [A] = reconstruct(Amix, Astruct) constructs the matrix A : N x p given the rearranged matrix Amix : N x p and Astruct. Each column of Amix should first contain the estimated non-zero elements of corresponding column of A followed by zeros. The number of non-zeros in each column of Amix should be equal to the number of 1s in corresponding column of Astruct; otherwise an error will be returned by the function.

(b) UV absorbances for seven 3-component mixtures have been obtained by preparing mixtures consisting of Co, Cr, and Ni salts in nitric acid according to an experimental design. Co is not present in mixtures 3 and 7, Cr is not present in mixtures 2, 4 and 6 while Ni is not present in mixtures 1 and 5. The file *ncadata.mat* contains both the mixture spectra (measabs) and pure species spectra (pureabs). Apply the fastNCA function to obtain the pure component spectra from UV mixture spectra. Report (i) the structural matrix for the mixture absorbances (ii) the estimated matrix A and (iii) the correlation coefficients between estimated and pure component spectra for all three species. (5 points)

(3) The generalized eigenvalue $\lambda$ and eigenvector $v$ of a square matrix **A** with respect to another square matrix **B** is defined as

$$Av = \lambda Bv$$

(a) Assuming that the matrix B is invertible, prove that the generalized eigenvalue and eigenvector of matrix A can be obtained by finding the eigenvalues and eigenvectors of the transformed matrix $C = L^{-1}AL^{-T}$, where $B = LL^T$. Derive the relation between the eigenvalues and eigenvectors of the transformed matrix C and the generalized eigenvalue and eigenvectors of A with respect to B. (5 points)

(b) Let Z be the n x N data matrix consisting N sample measurements of n variables. Let $\Sigma_e$ be the non-singular error covariance matrix. Use the result in (a) to prove that the singular values and left singular vectors of the scaled data matrix $L^{-1}Z$ are the square root of the generalized eigenvalues and eigenvectors, respectively of the covariance matrix $S_z = ZZ^T$ with respect to $\Sigma_e$, where $\Sigma_e = LL^T$ (5 points)

(c) If measurements of some variables are perfect, while measurements of other variables have errors with identical or different variances, then $\Sigma_e$ will be singular and the scaling strategy cannot be used to obtain optimal solution as in MLPCA. An alternative to scaling strategy is to use the generalized eigenvalue and eigenvector of $S_z$ with respect to $\Sigma_e$.

Consider a first order autoregressive dynamic model of a single input single output process with exogenous inputs given by

$$y_k = ay_{k-1} + bu_{k-1}$$

where $y_k$ is the measured value of output at time instant $k$ and $u_k$ is the measured value of manipulated input at time $k$ and $a$, $b$ are the unknown coefficients of the ARX model. 1000 samples of input and output measurements for a first order process is given in file *arx.mat*. It is given that the input measurements are perfect (noise free) while the output measurements contain noise with error variance $\sigma^2 = 0.0292$ (constant for all time instants). Use the generalized eigenvalue and eigenvectors of the data covariance matrix to estimate the parameters a and b. For this purpose, construct a data matrix from the time series data consisting of samples $[y_k \quad y_{k-1} \quad u_{k-1}]$. Thus, a data matrix of size 999x3 is obtained. Report the parameters of the ARX model obtained. (5 points)

(d) In part (c) it is implicitly assumed that we know the true order of the dynamic process and we have therefore appropriately constructed the data matrix to ensure that there will be only one constraint among the variables of the data matrix. In general the true order is unknown and therefore we stack several past instances of the output and input variable in the data matrix. That is the data matrix may contain for every instant $k$ a stacked vector corresponding to $[y_k \quad y_{k-1} \quad \ldots \quad y_{k-m} \quad u_{k-1} \quad u_{k-2} \quad \ldots \quad u_{k-m}]$. If the true process order is 1, how many constraints are present among the stacked vector of variables corresponding to a stacking order *m?* (2 points)

(e) Apply MLPCA to a data matrix obtained using stacking order 10 and obtain an estimate of the constraint matrix. From the estimated constraint matrix, obtain the equations relating the output variables to the inputs

$$\hat{\mathbf{A}}\mathbf{y} = \hat{\mathbf{B}}\mathbf{u}$$

where **y** is a vector of outputs from time *k* to time instant *k-m* and **u** is a vector of manipulated inputs from time *k-1* to time *k-m*. Determine a rotation matrix to transform the estimated $\hat{\mathbf{A}}$ so that it has the same structure as the true **A** with respect to location of non-zero elements. From the transformed matrix obtain an average estimate of the first order ARX model parameters and report these values. (8 points)

**Note: Use the MATLAB function eig(A,B) to find the generalized eigenvalue and eigenvectors of a square matrix A wrt square matrix B.**

*ALL THE BEST*