CH 5440 Multivariate Data Analysis in Process Monitoring and Diagnosis

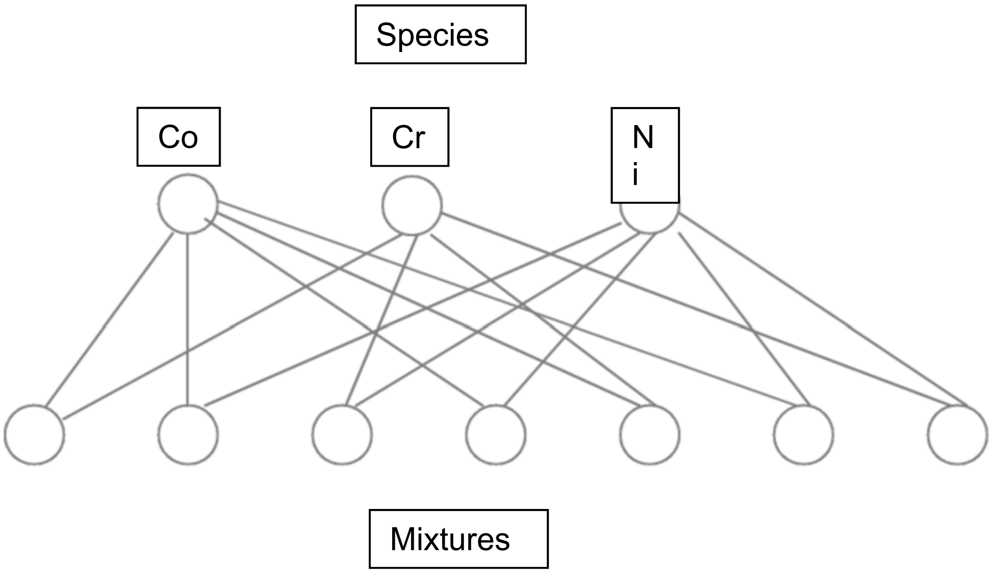
Assignment 4

1. UV absorbances for seven 3-component mixtures obtained by preparing mixtures consisting of Co, Cr, and Ni salts in nitric acid according to the experimental design shown in figure below is given in ncadata.mat.

(a) Show that the given network is NCA complaint.

(b) Apply PCA to the data set and obtain denoised spectra of mixtures assuming that the number of pure species is known. Determine the rotation matrix using the experimental design information and apply it to the denoised spectra and also estimate the pure component spectra using this rotation matrix. Report the rotation matrix as well as correlations between the estimated and true pure component spectra for all three species.

(c) Apply NCA to estimate the pure component spectra and compare with the true spectra given as part of the data set (use correlation coefficients). For applying NCA use the NCA toolbox (which has been downloaded from website of Prof. Liao's from UCLA).



2. Micro-array data for yeast cell cycle were obtained under different conditions by Spellman et al. (1998). A sub-matrix of the microarray data consisting of 441 genes over 56 time points is provided in data yeast.mat (variable microarraydata). An NCA compliant connectivity matrix corresponding to 33 transcription factors (TFs) which influence the expression levels of these genes is given in variable Astruct. Out of these 33 TFs, eleven are known to regulate expression levels of different genes during the cell cycle. Apply NCA to estimate the connectivity strengths as well as the temporal expression levels of the 33 TFs. Identify the eleven TFs involved in cell cycle regulation by examining the variance of all 33 TFs and choosing those that have the maximum variance. How many TFs are you

able to identify correctly using this procedure? (Note since there is a scale ambiguity, normalize each column of A and correspondingly scale the expression profile of TFs before comparing their relative variances).

The TFs implicated in cell cycle regulation have been identified as Ace2, Fkh1, Fkh2, Mbp1, Mcm1, Ndd1, Skn7, Stb1, Swi4, Swi5, Swi6. The names of all genes and TFs along with data re also given in excel file NCA_Yeast_dataset.xls.

3. Non-negative matrix factorization (NMF) can be used to extract the pure component spectra from the mixture UV absorbance data set used in assignment 3. NMF requires the number of pure species to be specified by the user, which can be assumed to be known to be three for this example. The NMF code downloaded from Prof. Haesun Park of Georgia Tech can be used for this purpose. Before applying NMF it is better to set all negative absorbance values in the data matrix to zero. Also an initial estimate of the non-negative matrices can be obtained by first using PCA to reduce the rank of the absorbance matrix and use absolute values of the loadings and scores matrix as initial guesses for the non-negative matrices.

(a) Use the first sample from the five replicates for each of the 26 mixtures and apply NMF to the data. Compare the extracted pure component spectra to the measured pure component spectra using correlation and identify which of the pure component spectra are being extracted well (Note that the order in which NMF extracts the pure components cannot be ascertained, that is, there is a permutation ambiguity). Therefore you have to compare each extracted spectra with every pure spectra to determine the permutation order.

(b) Determine the average of the five replicates for each mixture. Apply NMF to the averaged data and determine whether the pure components spectra are extracted more accurately.

Report the correlations for the two cases in the form of a table and report your conclusions.

(4) Optional Problem (previous years' end sem problem): Implement the following FASTNCA algorithm in MATLAB and apply it to the extract pure component spectra from the dataset given in problem 1

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

$$\mathbf{Z^* = AP}$$

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 by 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 6 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.

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.