# Solution Set- Assignment 3

1. The following is an example to demonstrate an important validation test for the constraint matrix through a regression matrix (other being the cross validation). The solution for the problem is followed first by the choice of dependent and independent sets of variables.

   (a) Assuming dependent flows to be $[1,2,4]$ and independent flows as $[3,5]$, true regression matrix computed as $\boldsymbol{R}_{true} = \boldsymbol{A}_D^{-1}\boldsymbol{A}_I$ which follows from the model equation $\boldsymbol{A}z = 0$

$$\boldsymbol{R}_{true} = \begin{bmatrix} 0 & 1 \\ 1 & -1 \\ 1 & 0 \end{bmatrix}$$

   The constraint matrix is estimated from the last 3 eigenvectors of the given data matrix flowdata3.mat and the regression matrix obtained for flows [3 5] chosen as independent variables

$$\boldsymbol{R}_1 = \begin{bmatrix} 0.0066 & 0.9815 \\ 0.9846 & -0.9540 \\ 1.0053 & -0.0156 \end{bmatrix}$$

   The maximum absolute difference between regression model coefficients is 0.0460.
   The eigenvalues are [2406.4 0.758 0.033 0.019 0.003].

   (b) Estimated error variances = [0.1020 0.2111 0.0638 0.1064 0.1886]
   Eigenvalues = [138642 66 1.87 1.08 0.10]. Maxdiff = 0.0146

   (c) Eigenvalues = [138642 66 1.87 1.08 0.10]. From the eigenvalues we can claim that number of constraints is not equal to 4 since last 4 eigenvalues are not equal to unity

   (c) Choice of dependent and independent variables

The rank of the constraint matrix corresponding to dependent variables (indicated by the condition number of the matrix for different choice of the dependent variables) indicates the feasibility of choosing them as the dependent set. The following table shows the condition number as a function of choice of independent variables:

| Choice of independent variables | Choice of independent variables | Condition number of R |
|---|---|---|
| 1,2 | 3,4,5 | 1.6075 |
| 1,3 | 2,4,5 | 1.8192 |
| 1,4 | 2,3,5 | 2.4064 |

| | | |
|---|---|---|
| 1,5 | 2,3,4 | 625.412 |
| 2,3 | 1,4,5 | 2.4036 |
| 2,4 | 1,3,5 | 3.3033 |
| 2,5 | 1,3,4 | 2.719 |
| 3,4 | 1,2,5 | 283.6 |
| 3,5 | 1,2,4 | 2.638 |
| 4,5 | 1,2,3 | 3.12 |

*Table 1: Variation of condition number with the choice of independent variables*

Choice of {2,3,4} or {1,2,5} as dependent variable sets appear a bad choice. The corresponding choice of {1,5} or {3,4} as independent flows is incorrect as can be observed from the flow network.

2. A) The first replicate of each sample is chosen and the model is built using PCR.

| 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|
| 0.0122 | 0.0088 | 0.0077 | 0.0074 | 0.0072 |

*Table 2: Average RMSE values obtained for different PCs retained*

| | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| Co | 0.0037 | 0.0037 | 0.0030 | 0.0029 | 0.0028 |
| Cr | 0.0212 | 0.0143 | 0.0126 | 0.0118 | 0.0114 |
| Ni | 0.0119 | 0.0086 | 0.0078 | 0.0076 | 0.0077 |

*Table 3: RMSE values of individual species*

1) The first replicates of all the samples are assembled together as the absorbance data ($\boldsymbol{Z}$) set for the PCR.
2) In LOOCV, the training set of data ($\boldsymbol{Z}_{train}$) is chosen by eliminating absorbance data corresponding to the one sample chosen (test sample $-\ \boldsymbol{Z}_{test}$). The left out sample concentration is predicted using the calibration model built using the other samples. This is repeated for all the samples of the dataset.
3) Apply PCA to the training set and the number of PCs is chosen, then the scores corresponding to chosen number of PCs is calculated.
4) Let $\boldsymbol{T}$, $\boldsymbol{C}$, $p, m$ and $s$ be the scores, concentrations, number of retained factors, sample size and number of species respectively. PCR follows an OLS regression between $\boldsymbol{C}_{train}$ and $\boldsymbol{T}$, where $\boldsymbol{T}$ $(Nm - 1 \times Np)$ is assumed to be erroneous and $\boldsymbol{C}$ $(Nm - 1 \times Ns)$ is assumed to be error free. $\boldsymbol{B} = (\boldsymbol{T}'\boldsymbol{T})^{-1}\boldsymbol{T}'\boldsymbol{C}_{train}$.
5) The predicted concentration of the training set data is calculated by $\boldsymbol{c}_{test} = \boldsymbol{z}_{test}\boldsymbol{V}_1\boldsymbol{B}$
6) The concentration RMSEs for each species are calculated between the predicted and the true concentration of the test sample.
7) Repeat the same for different number of PCs and the average RMSEs are calculated and the results are given in Table 1 and 2.

Using this technique, it can observed that the RMSE values gradually reduces and makes it difficult to ascertain the number of species.

b) Scaled PCR method (wavelengths are scaled with the sample standard deviations calculated from replicates of the sample)

| 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|
| 0.0133 | 0.0102 | 0.0003 | 0.0003 | 0.0003 |

*Table 4: Average RMSE values obtained for different PCs retained*

| | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| Co | 0.0031 | 0.0032 | 0.00011 | 0.00011 | 0.00012 |
| Cr | 0.0262 | 0.0217 | 0.00052 | 0.00050 | 0.00050 |
| Ni | 0.0109 | 0.0064 | 0.00038 | 0.00038 | 0.00040 |

*Table 5: RMSE values of individual species*

1) The standard deviations are calculated as stated in the problem and the PCR is applied for the scaled data $\boldsymbol{Z}_s = \boldsymbol{Z}\boldsymbol{L}^{-1}$.
2) The same procedure is followed as previously explained on the scaled data and the RMSE values are calculated correspondingly.

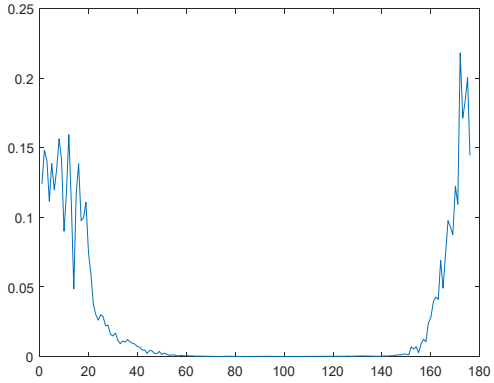
Scaled PCR exhibits a significant drop in RMSE values at PC = 3 and the RMSE values remain constant for different number of PC > 3 which is indicative of the number of species present in the mixture. Thus performance of scaled PCR is better than PCR in terms of predicting the concentrations of the unknown species since RMSE values are lower for PC >= 3.

c) In order to estimate the standard deviation of errors with respect each wavelength IPCA is applied in first step instead of PCA. However, it is difficult to estimate all 176 error variances simultaneously, due to the large number of decision variables. We can use a divide and conquer method where we divide the absorbance matrix into smaller sub-blocks of wavelengths say of 25 wavelengths each and estimate the error variances for each block separately. After estimating the error variances we can combine them into a single vector and apply scaled PCR to develop the calibration model. Using this procedure, the RMSE values obtained for different number of PCs are as given below

| 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|
| 0.0133 | 0.0102 | 0.0003 | 0.0003 | 0.0003 |

| | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| Co | 0.0019 | 0.00086 | 0.00087 | 0.000073 | 0.000074 |
| Cr | 0.0299 | 0.0315 | 0.0011 | 0.0005 | 0.0005 |
| Ni | 0.0126 | 0.0004 | 0.0004 | 0.00015 | 0.00016 |

It is observed that again RMSE drops for PC = 3 and remains more or less same for higher number of PCs. The RMSE values obtained using IPCR are comparable to Scaled PCR indicating that error variances are estimated well. A plot of estimated standard deviations with respect to wavelength for PC = 3 is shown below.




(d)