Solution Set- Assignment 3

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

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| 0.0122 | 0.0088 | 0.0077 | 0.0074 | 0.0072 | 0.0068 | 0.0066 | 0.0063 |

Table : Average RMSE values obtained for different PCs retained

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| Co | 0.0037 | 0.0037 | 0.0030 | 0.0029 | 0.0028 | 0.0026 | 0.0025 | 0.0024 |
| Cr | 0.0212 | 0.0143 | 0.0126 | 0.0118 | 0.0114 | 0.0107 | 0.0103 | 0.0099 |
| Ni | 0.0119 | 0.0086 | 0.0078 | 0.0076 | 0.0077 | 0.0072 | 0.0072 | 0.0067 |

Table : RMSE values of individual species

1. The first replicates of all the samples are assembled together as the absorbance data ( set for the PCR.
2. In LOOCV, the training set of data is chosen by eliminating absorbance data corresponding to the one sample chosen (test sample). 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, , and be the scores, concentrations, number of retained factors, sample size and number of species respectively. PCR follows an OLS regression between and, where() is assumed to be erroneous and () is assumed to be error free. .

Note that dimensions of scores and the corresponding concentrations are reduced to due to LOOCV.

1. The predicted concentration of the training set data is calculated by
2. The concentration RMSEs for each species are calculated between the predicted and the true concentration of the test sample.
3. 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) The average of replicates are chosen for all the samples and LOOCV is applied

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| 0.0123 | 0.0085 | 0.0023 | 0.0022 | 0.0021 | 0.0018 | 0.0018 | 0.0017 |

Table : Average RMSE values obtained for different PCs retained

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| Co | 0.0036 | 0.0036 | 0.0007 | 0.0007 | 0.0006 | 0.0006 | 0.0006 | 0.0005 |
| Cr | 0.0214 | 0.0139 | 0.0033 | 0.0032 | 0.0030 | 0.0026 | 0.0027 | 0.0026 |
| Ni | 0.0118 | 0.0080 | 0.0029 | 0.0027 | 0.0026 | 0.0022 | 0.0021 | 0.0020 |

Table : RMSE values of individual species

Clearly averaging performs better than the previous PCR method to an extent as we can observe a very sharp drop in the RMSE values corresponding to PC = 3. However, the cross validation RMSE continues to reduce as more number of factors are retained which does not give a conclusive picture on the actual number of species present.

C) MLPCR method (wavelengths are scaled with the sample standard deviations calculated from replicates of the sample)

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| 0.0133 | 0.0102 | 0.0003 | 0.0003 | 0.0003 | 0.0003 | 0.0003 | 0.0003 |

Table : Average RMSE values obtained for different PCs retained

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| Co | 0.0031 | 0.0033 | 0.0001 | 0.0001 | 0.0001 | 0.0001 | 0.0001 | 0.0001 |
| Cr | 0.0260 | 0.0212 | 0.0005 | 0.0005 | 0.0005 | 0.0005 | 0.0004 | 0.0004 |
| Ni | 0.0108 | 0.0063 | 0.0004 | 0.0004 | 0.0004 | 0.0003 | 0.0003 | 0.0003 |

Table : 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.
2. The same procedure is followed as previously explained on the scaled data and the RMSE values are calculated correspondingly.

MLPCR also exhibits a similar drop 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 MLPCR is better than the above methods. It can also be concluded as MLPCR performs better in terms of predicting the concentrations of the unknown species from lesser RMSE values obtained. Figure 1 shows RMSE plots against the number of PCs retained for different methods discussed.



Figure 1: Average RMSE values plotted against number of PCs retained for PCR, PCR-averaged and MLPCR

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 and independent flow as, true regression matrix computed as which follows from the model equation

The constraint matrix is estimated from the last 3 eigenvectors

For Data-1 (high SNR)

The maximum absolute difference between regression model coefficients is 0.0108

For Data-2 (low SNR)

The maximum absolute difference between regression model coefficients is 0.4057

1. Plot of log of singular values with the number of PC’s



True model order is not apparent from the above figure for the low SNR case. It can be incorrectly be perceived as 4.

* + 1. Effect of auto-scaling

The regression coefficients () are obtained for Data-1 and Data-2 as in part (i), maximum absolute difference between true and estimated as,

|  |  |
| --- | --- |
| Data-1 | 0.4999 |
| Data-2 | 0.2905 |

This demonstrates that auto-scaling is not a desired a pre-processing technique when the error in variables are small. However, when the SNR of the data is low the auto scaling yields a smaller maximum absolute deviations than when it is not used. This is an indication that auto scaling is suitable for data with low SNR.

* + 1. Choice of dependent and independent variables

The rank of the constraint matrix corresponding to independent variables (indexed by the condition number for different variable pairs) indicates the feasibility of choosing them as the dependent set. The Table 7 shows the condition number as a function of choice of independent variables for dataset 1:

|  |  |
| --- | --- |
| Choice of independent variables | Condition number of R |
| 1,2 | 2.3592 |
| 1,3 | 1.9995 |
| 1,4 | 2.0138 |
| 1,5 | 472.1158 |
| 2,3 | 2.3585 |
| 2,4 | 2.3492 |
| 2,5 | 2.3585 |
| 3,4 | 245.3006 |
| 3,5 | 1.9847 |
| 4,5 | 1.9988 |

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

Choice of {1, 5} or {3, 4} as independent variable set is particularly appears bad as these flows do not form an independent set and this is consistent with the true network.