# CH5540 Multivariate Data Analysis - Assignment 3

AE14B050

March 27, 2018

## Solution 1:

**(a) Procedure:** Given to us is dataset of 1000 samples with both true and measured values. Also given are the error standard deviations and true constraint matrix.

To apply PCA and get the relational matrix between the dependent and independent variables we use the following procedure:

- Create a Cholesky transform inverse from the true error standard deviations $L^T = [diag(1/std)]$
- Scale the measured $Fmeas$ data using $L^T$and create $Fs = Fmeas * L^T/\sqrt{samples}$
- We can now apply PCA as all the error standard deviations are equal
- Perform SVD to obtain the eigenvectors of both the true data $Ftrue$ and scaled data $Fs$
- Transpose and reindex the eigenvectors to obtain the constraint matrices for both the true data and measured data
- Scale back the constraint matrix obtained from $Fs$ back into the original space as $\hat{A} = \hat{A} * L^T$
- Select the possible dependent and independent variables and select only the top '$d$' constraints where $d$ in the number of dependent variables
- Seperate the constraint matrices into dependent and independent parts so that we have the equation $A_D X_D + A_I X_I = 0$ which is a sliced version of $AX = 0$ (Here $X$ is our data matrix, whether true or measured and $A$ is the corresponding constraint matrix)
- Calculate the regression model to get dependent variables as $X_D = R * X_I$ ( where $R = (-A_D^{-1} * A_I)$ ) for both measured and true data
- We can now compare our regression coefficient matrices $R$ and $\hat{R}$ by calculating the differences between them and finding the maximum of each coefficient

**F3, F5 Independent:**

$$\hat{R} = \begin{bmatrix} 0.0058 & 0.9890 \\ 0.9747 & -0.9502 \\ 0.9470 & 0.1059 \end{bmatrix}; R = \begin{bmatrix} 0 & 1.000 \\ 1.0000 & -1.000 \\ 1.000 & 0 \end{bmatrix}$$

$$maxDiff = 0.10595$$

The denpendency of F4 on F5 shows the highest difference compared to the true regression model.

**(b) F1 independent :**

$$\hat{R} = \begin{bmatrix} 1.0016 \\ 2.0018 \\ 2.0017 \\ 0.9993 \end{bmatrix}; R = \begin{bmatrix} 1.0015 \\ 2.0015 \\ 2.0015 \\ 1.000 \end{bmatrix}$$

$$maxDiff = 6.5110e-04$$

Even though the diffenerence is very small both in order and value, the denpendency of F1 on F5 shows the highest difference compared to the true regression model.

**F1,F2,F3 independent :**

$$\hat{R} = \begin{bmatrix} 0.5773 & 0.4894 & 0.4667 \\ 0.5487 & -0.4812 & 0.4658 \end{bmatrix}; R = \begin{bmatrix} 0 & 0 & 1.000 \\ 0.0119 & -0.9881 & 0.9881 \end{bmatrix}$$

$$maxDiff = 0.5773$$

As we observe the diffenerence is very large for almost all the values, the denpendency of F4 on F1 shows the highest difference compared to the true regression model. This is generally a case of wrong assumptions of our independent and dependent variables or more precisely choosing the number dependent variables less then the number of true constraint equations.

**(c)IPCA procedure:** We follow a very similar approach in IPCA when we are comparing our regression models but the selection of variables in PCA application is revamped so that we do need the true standard deviations of the errors. To apply IPCA we use the following procedure:

- Initialize a Cholesky transform inverse using $I$ matrix scaled by $\sqrt{samples}$, $L^T = I/\sqrt{samples}$
- Select a set of indenpendent variables and a convergence criterion to stop our iterative PCA loop
- We then start an iterative loop until we predict error standard deviations correctly
  - Scale the measured $Fmeas$ data using $L^T$and create $Fs = Fmeas * L^T$
  - Perform SVD obtain the eigen vectors of the scaled data $Fs$
  - Calculate sum of the least '$d$' eigenvalues where d is the number of dependent variables
  - Check if the difference betweent the sum from the previous iteration and this new sum is satifies our convergence criterion and if it does, we stop with this iteration
  - Store the eigenvalue sum to be used in the next iteration, if any
  - Transpose and reindex the eigenvectors to obtain the constraint matrices for the measured data
  - Scale back the constraint matrix obtained from $Fs$ back into the original space as $\hat{A} = \hat{A} * L^T$and select only the top $d$ constraints
  - Estimate the new standard deviations using the $stdest()$ function
  - Calculate the new $L^T = [diag(1/std)]/\sqrt{samples}]$
- We output the final constraint matrix $\hat{A}$ to be used to calculate our regression model and comparision with the true constraint matrix just as in the PCA format.

**F1,F2 independent:** Upon following the basic procedure to obrain the regression matrix as given by the PCA procedure we obtain the regression model for our IPCA variant

$$\hat{R} = \begin{bmatrix} 0.9666 & 1.0335 \\ 1.0231 & 0.9769 \\ 1.0029 & -0.0036 \end{bmatrix}; R = \begin{bmatrix} 1 & 1 \\ 1 & 1 \\ 1 & 0 \end{bmatrix}$$

$$maxDiff = 0.0335$$

**Determining independent variables:** There are theoretical aspects to be noted in IPCA. If we select the number of dependent variables to be too small, the number of constraint equations we get might have not be

enough to perform the estimation step and it requires at least $m$ constraint equation such that $\frac{m(m+1)}{2} \geq n$ where $n$ is the number of varibles for which we need to obtain the error standard deviations.

For our current case this limits us to atleast 3 dependent variables to obtain error standard deviation estimates for all the 5 variables.

To find the number of independent variables we just loop over the two cases of 1 and 2 independent variables and find which of the two cases has a lower $Svar = \frac{1}{d}\sum_{i=1}^{i=d}(\lambda_i - 1)^2$. If we have chosen correctly, this value will be very low as the $d$ least eigen values will be distrubuted around 1. It has to be noted that this procedure only applies when the scaling is done using the $\sqrt{samples}$as in this case.

In the general sense we just iterate and find the case with the lowest $Svar$ that also satisifies the basic IPCA solution constraint.

**(d) Determining set of independent variables:** We shall use the information obtained from the before step that the number of independent variables is 2.

We now iterate over all the possible pairs of independent variables and calculate the value of maxDiff for each case. We tabulate below the results

| | F1 | F2 | F3 | F4 | F5 |
|---|---|---|---|---|---|
| F1 | - | 0.0335 | 0.1094 | 0.1157 | inf(4.4683e15) |
| F2 | - | - | 0.1170 | 0.1105 | 0.0369 |
| F3 | - | - | - | inf(2.014e15) | 0.1087 |
| F4 | - | - | - | - | 0.1150 |
| F5 | - | - | - | - | - |

duplicate and same variable pairs calculations are left blank
inf - denotes that the dependent variable constraint matrix is singular and hence is a worst case

Table 1: maxDiff for all two variable pairs

The cases of $(F1, F5)$ and $(F3, F4)$ both result in singular dependent variable constraint matirces and are hence worst choices, the former case being the worse of the two pairs.

The case of $(F1, F2)$ gives the lowest maxDiff and is hence the best possibles choice of independent variables.

## Solution 2:

Given to us are 2 dataset, DATA comprising of the absorbance spectrum for each sample and CONC, detailing the concentration of each non-reacting elements in the sample. We also have 3 pure absorbance spectrum vectors corresponding to each of the elements and their corresponding concentrations.

As directed in the question we seperate the 1st sample from the 5 samples for each mixture to be used as our LOOCV or the testing data. We use the rest to train our OLS models.

**(a)Pure component spectra:** From the pure component spectra we obtain the maximum wavelength for each of the species.

| Co | Cr | Ni |
|---|---|---|
| 510nm | 406nm | 394nm |

From these wavelengths we build an OLS model (no scaling involved). Given below is the coefficient matrix that relates the abosorbance spectrum of the maximum wavelengths to the concentrations $C = XA$, $X$ is the data matrix while $A, C$ are the coefficient and concentration matrices respectively.

$$A = \begin{bmatrix} -0.0047 & -0.0031 & 0.2150 \\ 0.2671 & -7.216 & -0.2803 \\ -0.2251 & 0.8106 & 0.2321 \end{bmatrix}$$

Here each column corresponds the to coefficients for each element in order.

The RMSE error is also reported for each of the variables as $Co = 7.564e-4 \quad Ni = 0.0025 \quad Cr = 7.513e-4$ .

**(b)PCR :** We take all the 176 variables for create our PC's. For applying PCA, we autoscale the data across the samples dimension. We then perform SVD over the autoscaled data and iteratively calculate RMSE error over 1 to 104 PC's, 104 is the maximum number of PC's obtainable instead of 176 as the number samples of 104 limits this. Following is the graph showing the RMSE values over all the PC's chosen. Minimum RMSE we obtain is at 80 PC's with a value of 0.0196 in the scaled domain.

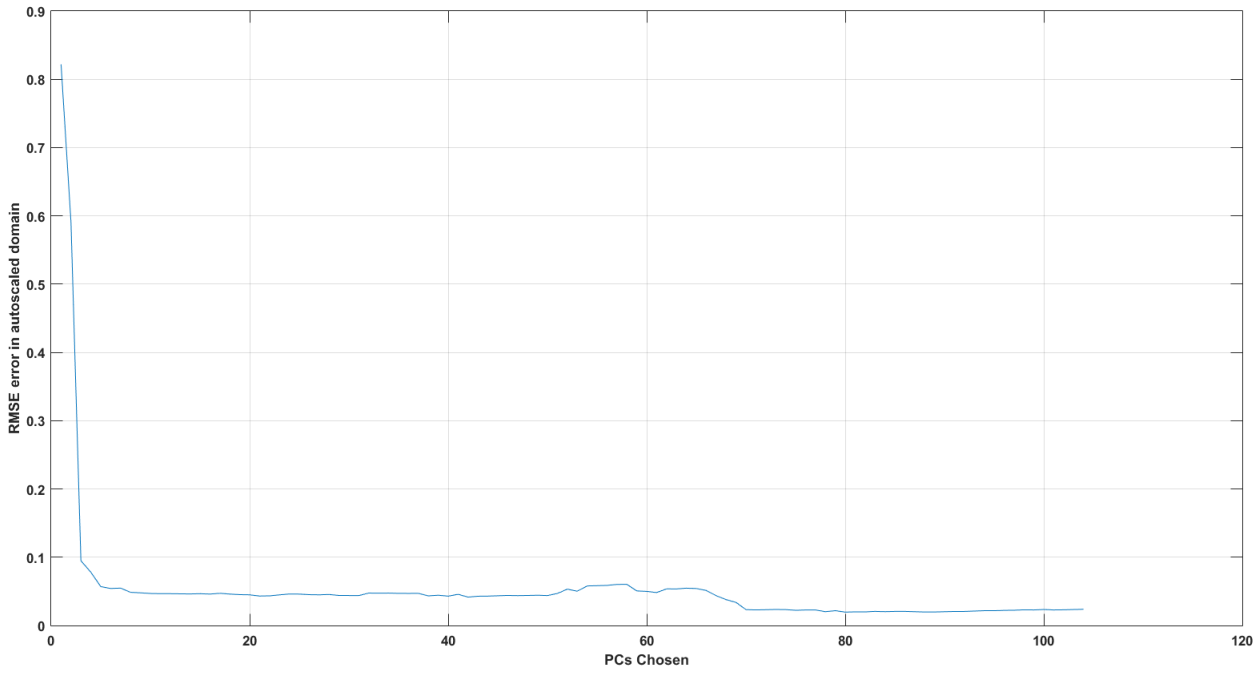


Figure 1: RMSE plot for PCR

Using the condition of capturing 99% of variaance over the data, the number of PC's we obtain are 40. Hence using PCR and LOOCV as it is we identify 40 different underlying species in our measurement.

**(b2)PCR-averaged :** Averaging the data we reduce our number of samples to 26 thereby limiting our maximum number of PC's to 26 too. We follow just as in the previous case and plot our RMSE values. Minimum RMSE we obtain is at 24 PC's with a value of 0.0926 in the scaled domain. Indicating no improvement of our model over the non averaged case.

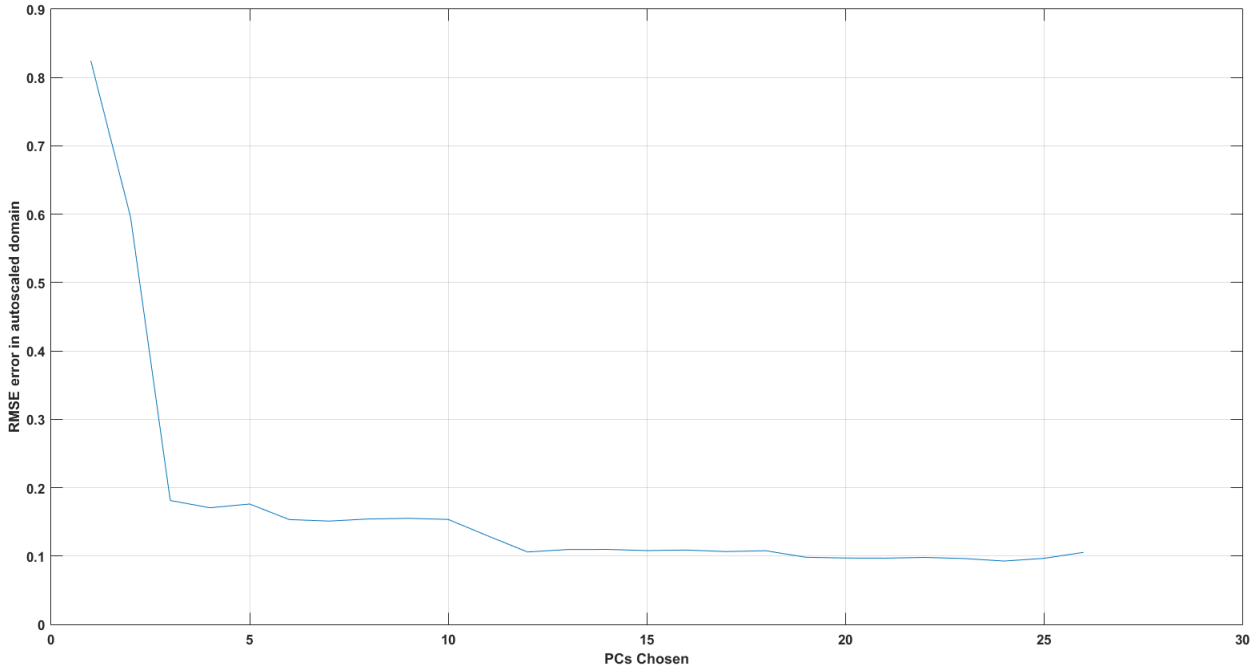


Figure 2: RMSE plot for PCR-averaged

Using the condition of capturing 99% of variaance over the data, the number of PC's we obtain are 18. Hence using PCR and LOOCV as identify 18 different underlying species in our measurement.

**(c)MLPCR :** From the given data we can estimate there exists some error function for each measurement.

For the first case we suppose this error function stays constant over each mixture or atleast doesn't vary much from mixture to mixture. Plotting the error standard deviation across the wavelengths for each of the 26 different mixture we observe this assumption is true and also we verfiy that the measurements are very noise towards the ends of the absorbance spectrum.

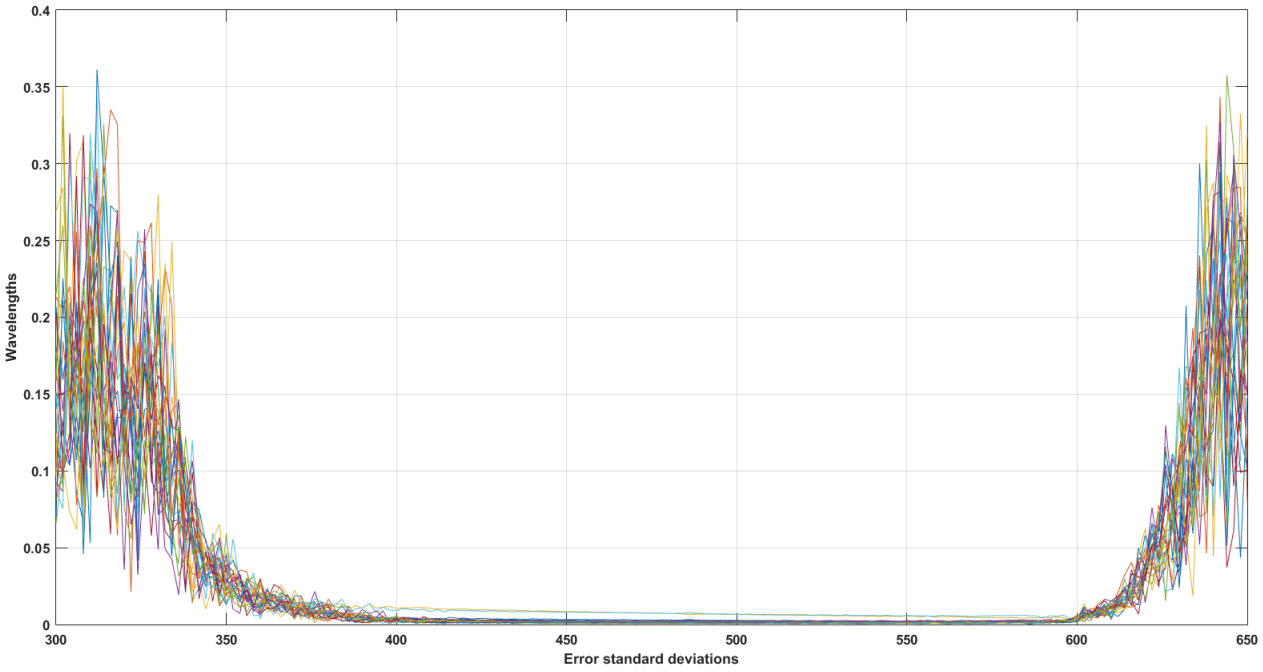


Figure 3: Error standard deviations across wavelengths for all mixtures

Let us average this over all the samples for better observations.

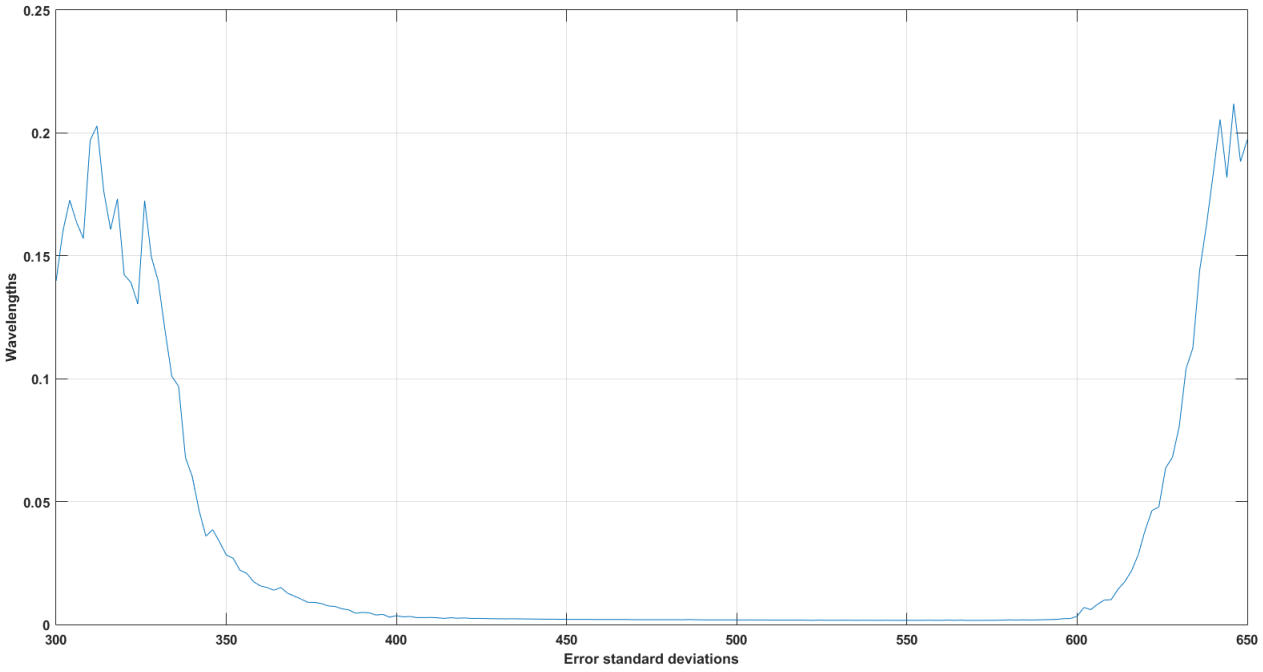


Figure 4: Error standard deviations across wavelengths averaged for all mixtures

We shall also plot the error standard deviations across mixtures to check if they are significant. From the plot it is plain to observe there isn't much change or atleast changes in the error standard deviations are not as as

significant as across the wavelengths.

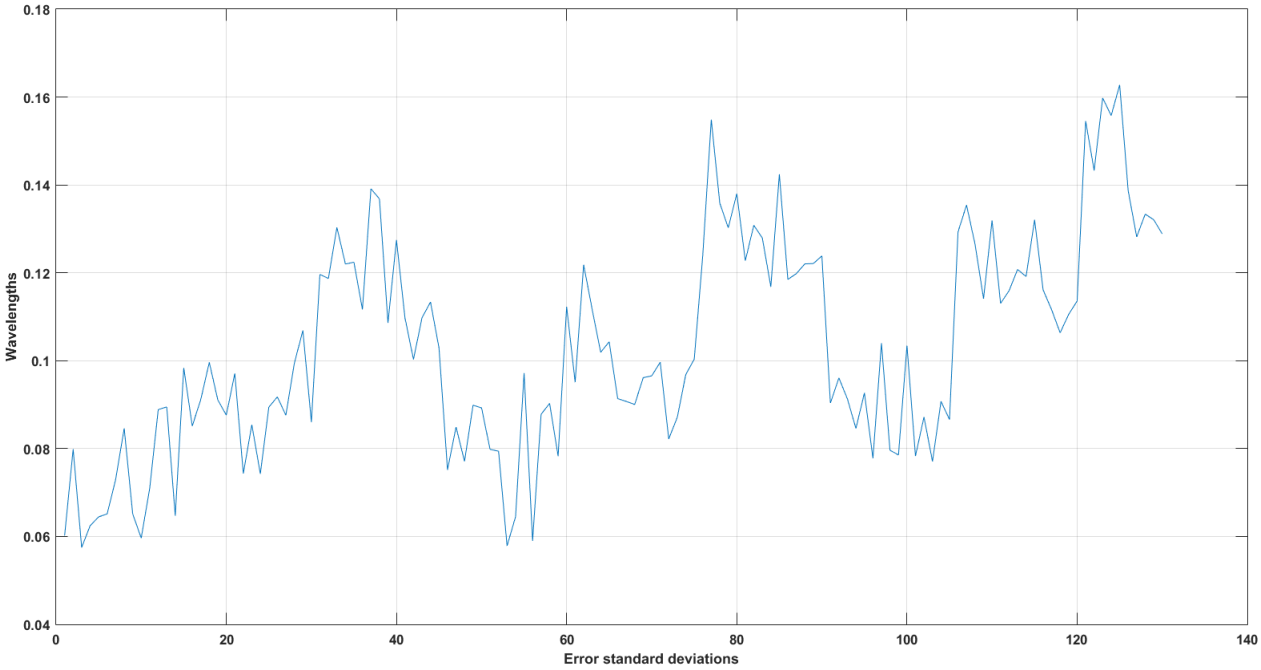


Figure 5: Error standard deviations across mixtures

We use the error standard deviation obtained across wavelengths to scale our data and obtain equal error standard deviations for PCA application.

We as usual perform SVD and repeat the steps as in the previous cases. This time though we see a huge improvements in the number of PC's required to capture 99% of the data variance as only 3 are sufficient. This is also verified by the RMSE plot we get which sees no further improvement at all after rapidly reducing over the first three variables. The minimum RMSE error value is also very small at $1.9742e-4$. Though this in a different scaled domain it is still comparitively much less than the previous cases.

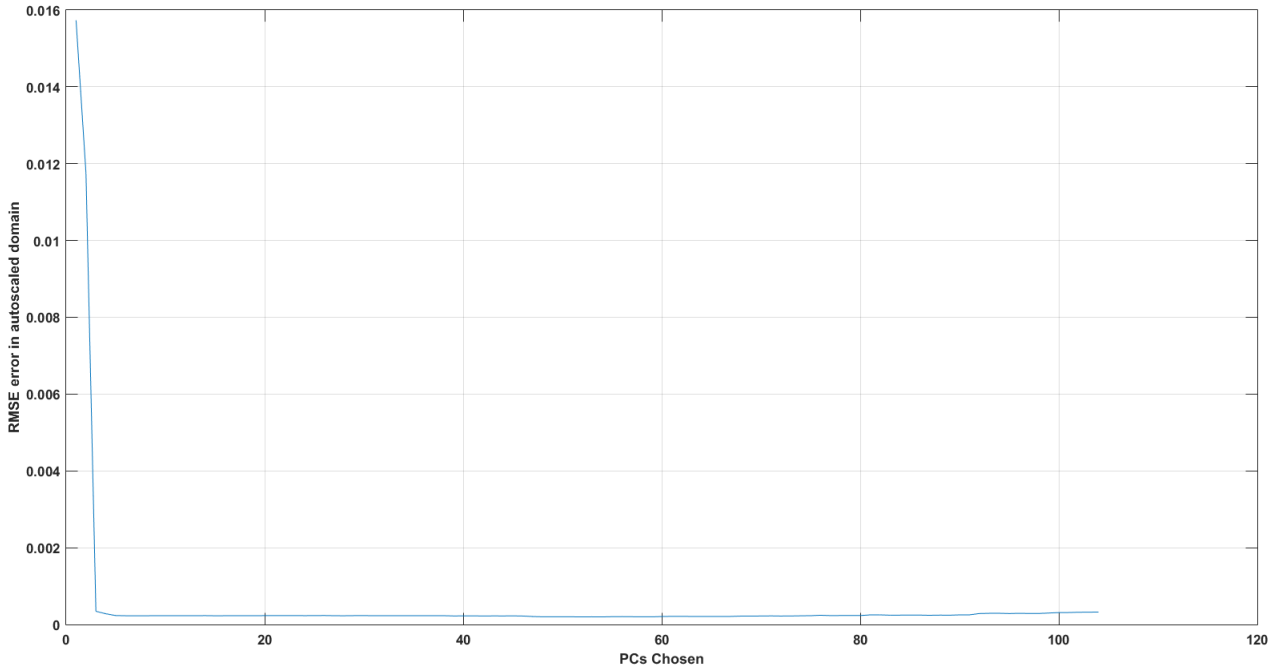


Figure 6: RMSE plot for MLPCR

Hence using MLPCR and LOOCV we identify 3 different underlying species in our measurement.

**(d)IPCR :** We shall now couple the methods used in question $1c$ and $2c$ to apply IPCA. First we apply IPCA to our data and use the helper function *stdest*() to calculate the estimated standard deviation errors and number of independent variables which come out to be 3. We then scale our data using the estimated standard deviations and apply PCA. We reapeat the steps followed in previous cases to find the number of PC's to capture 99% variance is 2. Verified by the RMSE plot we get which sees no further improvement at all after rapidly reducing over the first two variables. The minimum RMSE error value is also very small at $1.7468e-4$ which is further improved over the previous MLPCR case.

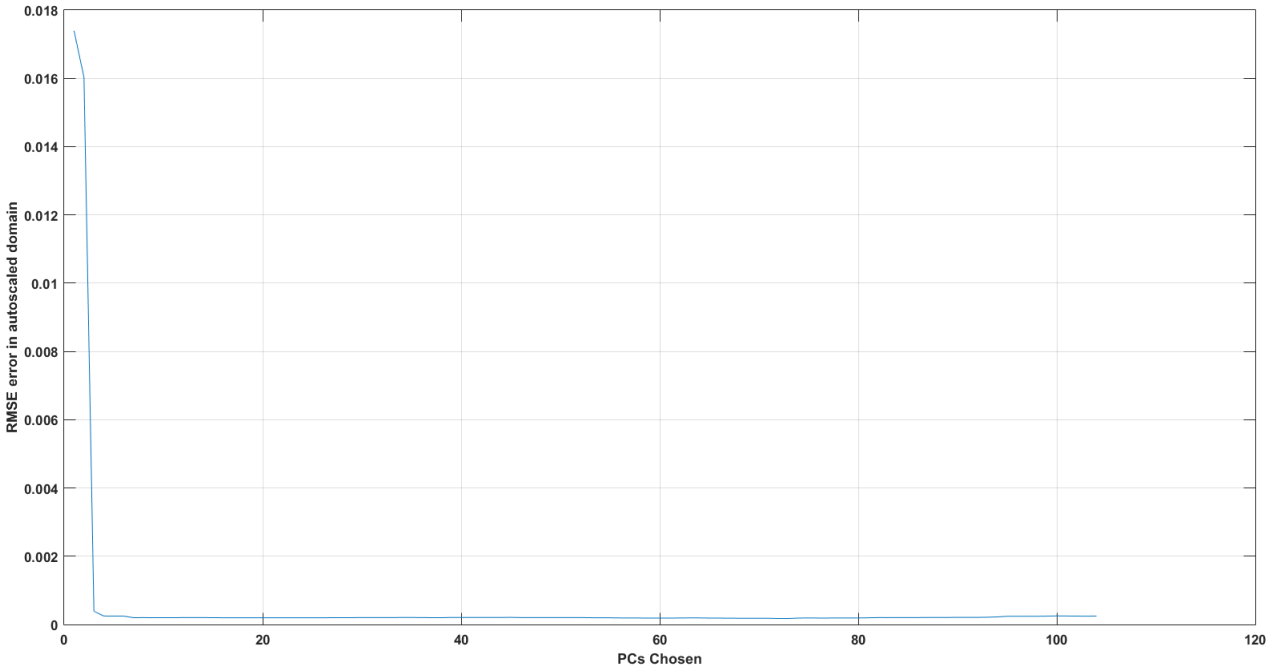


Figure 7: RMSE plot for MLPCR

Hence using IPCR and LOOCV we identify 2 different underlying species in our measurement.

We finally compare our estimated standard deviations to the true standard deviations.

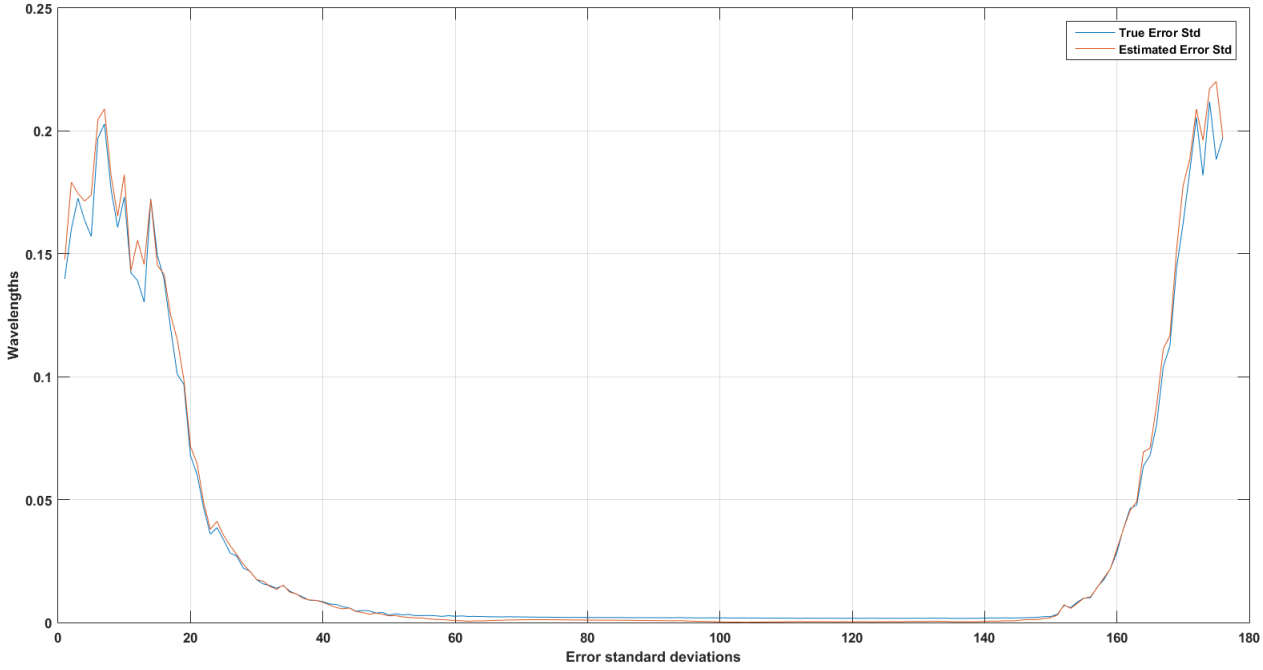


Figure 8: Estimated vs true Standard deviations plot