CH5460 HW3

Prem Sagar S - AE14B021

March 19, 2017

Problem 1

Part a

Number of PCs	RMSE
1	0.0246282722564009
2	0.0165370341625899
3	0.0138826359662914
4	0.0126437489662498
5	0.0124521793534631
6	0.0103387843439066
7	0.0103207936823860
8	0.0103507971832113
9	0.00663314737349578
10	0.00710262577155226
11	0.00758997113939423
12	0.00821254047680662
13	0.00815626246730665
14	0.00840445762381331
15	0.00800441150939719
16	**0.00439941935104714**
17	0.00453928134158650
18	0.00487021088783960
19	0.00550457080937712
20	0.00618889113069417
21	0.00679818206742099
22	0.00853339402109971
23	0.00809004733741021
24	0.0161789187560918
25	0.0306101233852825
26	0.00815196168016452
27-176	0.00815196168016452

Table 1: RMSE for different choice of PCs

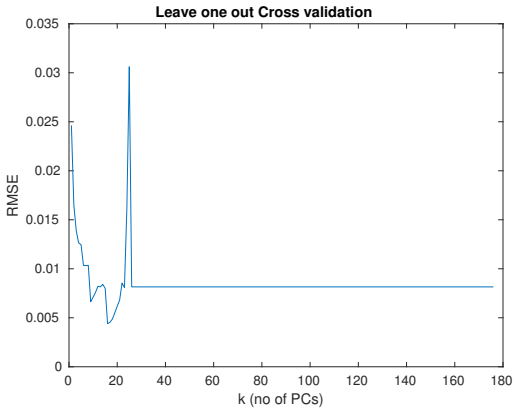

Figure 1: RMSE from LOOCV vs no of PCs

Yes. From the RMSE data for each choice of PCs, the one corresponding to least RMSE is 16 PCs.

Part b

Number of PCs	RMSE
1	0.0246915066597355
2	0.0161568670144919
3	0.00363575123522478
4	0.00351976179006457
5	0.00351581967732227
6	0.00274662475638195
7	0.00287341736655089
8	0.00276583545447820
9	0.00197312477495101
10	0.00203917604194091
11	0.00157523898492577
12	0.00151436762184238
13	**0.00148493009627178**
14	0.00158871561852440
15	0.00183277738570059
16	0.00199166723804930
17	0.00178953151512532
18	0.00221868542524504
19	0.00228489935750614
20	0.00255959708643358
21	0.00271559558784269
22	0.00215569449246215
23	0.00407201471590375
24	0.00544319407121081
25	0.0248850188557172
26	0.00262338226948231
27-176	0.00262338226948231

Table 2: RMSE for different choice of PCs

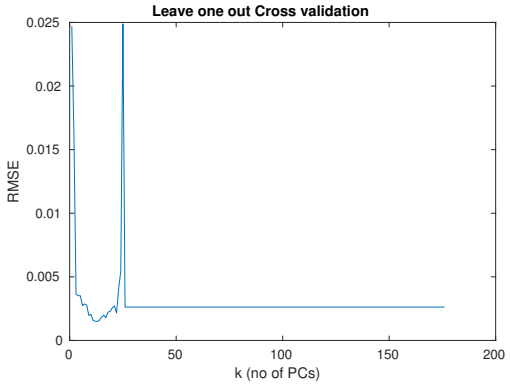

Figure 2: RMSE from LOOCV vs no of PCs

By taking the mean of the data measurements, the number of PCs has reduced from 16 to 13. RMSE has improved. Because repeated experiments averages out the errors that were probably biased by some fluctuations.

Part c

The measurements are noisy near the ends. This is verified from the standard deviations for some mixture for all wavelengths.

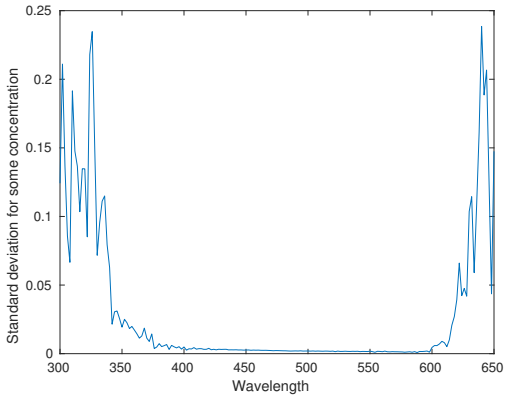

Figure 3: STD for some mixture vs Wavelength

The standard deviation of measurements for a particular wavelength is plot below. They're all not same but taking a mean of the isn't bad as they are seem evenly distributed about the mean.

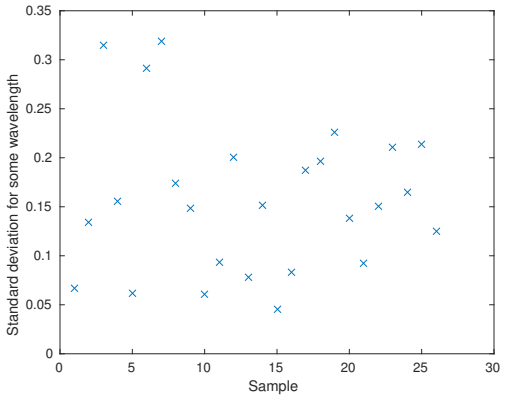

Figure 4: STD for some wavelength vs Mixtures

The table below gives RMSE taking only the 1st data of each block of 5 measurements for each sample and then scaled with mean error standard deviation. The optimal number of PCs has reduced to 7 and significant improvement in RMSE can be seen.

Number of PCs	RMSE
1	0.0283142261811793
2	0.0217153152085491
3	0.000661972874295058
4	0.000611819386737660
5	0.000547984721134788
6	0.00054634625014 4158
7	**0.000546157766336328**
8	0.000567296498945971
9	0.000625163925397771
10	0.000653911323671472
11	0.000693000063698665
12	0.000751684637960156
13	0.000756303882811698
14	0.000879174392785865
15	0.000855816728218118
16	0.000881083812017657
17	0.000794658061486840
18	0.000886057472553013
19	0.00103268524081273
20	0.00123520623343307
21	0.00131648637798961
22	0.00187491034221159
23	0.00250004809326108
24	0.00277053376559627
25	0.00490547263900845
26	0.00143104958590139
27-176	0.00143104958590139

Table 3: RMSE for different choice of PCs

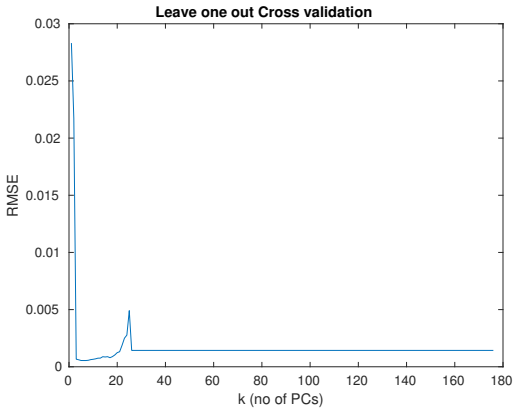

Figure 5: RMSE from LOOCV vs no of PCs

The table below gives RMSE taking the mean of 5 measurements and then scaled with error standard deviation. The number of optimal PCs has further reduced to 6 and RMSE has improved even more.

Number of PCs	RMSE
1	0.0282521006459083
2	0.0217232327559754
3	0.000445477970805625
4	0.000323585586580870
5	0.000306839144052206
6	**0.000289469708228384**
7	0.000307341142634353
8	0.000307540763187457
9	0.000323664397366716
10	0.000345683918786343
11	0.000341514549152057
12	0.000366938522579584
13	0.000370412250288105
14	0.000381969859943723
15	0.000416642871000959
16	0.000461085518332247
17	0.000500297153507263
18	0.000592975085483032
19	0.000631316310517178
20	0.000759590835751427
21	0.000904241303704409
22	0.00125649838086309
23	0.00154743438399033
24	0.00316680440878699
25	0.00485723219512706
26	0.000818202811399507
27-176	0.000818202811399507

Table 4: RMSE for different choice of PCs

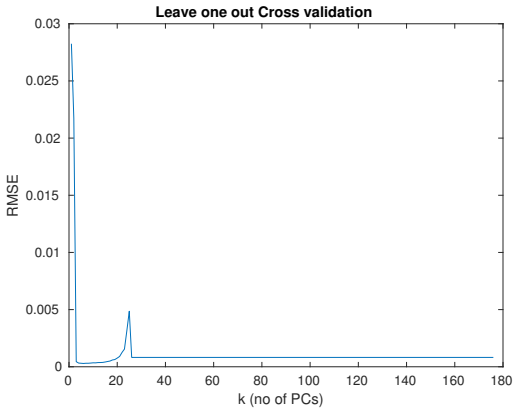

Figure 6: RMSE from LOOCV vs no of PCs

	Scaled by σ_{err}?			
	No		Yes (MLPCA)	
Data	1st reading	Mean of meas.	1st reading	Mean of meas.
Optimal PCs	16	13	7	6
RMSE	0.004399	0.001484	0.000546	0.000289

Table 5: Summary

Problem 2

Part a

For the flow system, conservation equations can be written as below.

$$\begin{aligned} F_1 + F_2 &= F_3 \\ F_3 &= F_4 \\ F_4 &= F_2 + F_5 \end{aligned}$$

Or,

$$\begin{bmatrix} 1 & 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & -1 & 0 \\ 0 & -1 & 0 & 1 & 1 \end{bmatrix} \begin{bmatrix} F_1 \\ F_2 \\ F_3 \\ F_4 \\ F_5 \end{bmatrix} = 0$$

Let's choose the independent variables to be F_1 and F_2. Let Z_D and Z_I be the vectors of dependent and independent vectors such that,

$$Z_D = BZ_I$$

where B is the regression matrix.

$$B = -\begin{bmatrix} -1 & 0 & 0 \\ 1 & -1 & 0 \\ 0 & 1 & 1 \end{bmatrix}^{-1} \begin{bmatrix} 1 & 1 \\ 0 & 0 \\ 0 & -1 \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 1 & 1 \\ 1 & 0 \end{bmatrix}$$

$$Z_D^* = B^* Z_I^*$$

Flow data set 1

From the measured data set, the regression matrix is obtained from PCA as

$$B^* = \begin{bmatrix} 0.9993 & 1.0004 \\ 1.0051 & 0.9961 \\ 1.0046 & -0.0052 \end{bmatrix}$$

The true measurements give the following regression matrix which is consistent from that obtained previously.

$$B = \begin{bmatrix} 1 & 1 \\ 1 & 1 \\ 1 & 0 \end{bmatrix}$$

Maximum absolute difference between estimated regression model coefficients and true regression model coefficients:

$$abs(B - B^*) = \begin{bmatrix} 0.0007 & 0.0004 \\ 0.0051 & 0.0039 \\ 0.0046 & 0.0052 \end{bmatrix}$$

Flow data set 2:

The regression matrix is obtained from PCA as

$$B^* = \begin{bmatrix} 0.9203 & 1.0798 \\ 0.9453 & 1.0546 \\ 1.2343 & -0.2345 \end{bmatrix}$$

Again, the true measurements give the following regression matrix which is consistent from that obtained previously.

$$B = \begin{bmatrix} 1 & 1 \\ 1 & 1 \\ 1 & 0 \end{bmatrix}$$

Maximum absolute difference between estimated regression model coefficients and true regression model coefficients:

$$abs(B - B^*) = \begin{bmatrix} 0.0797 & 0.0798 \\ 0.0547 & 0.0546 \\ 0.2343 & 0.2345 \end{bmatrix}$$

Part b

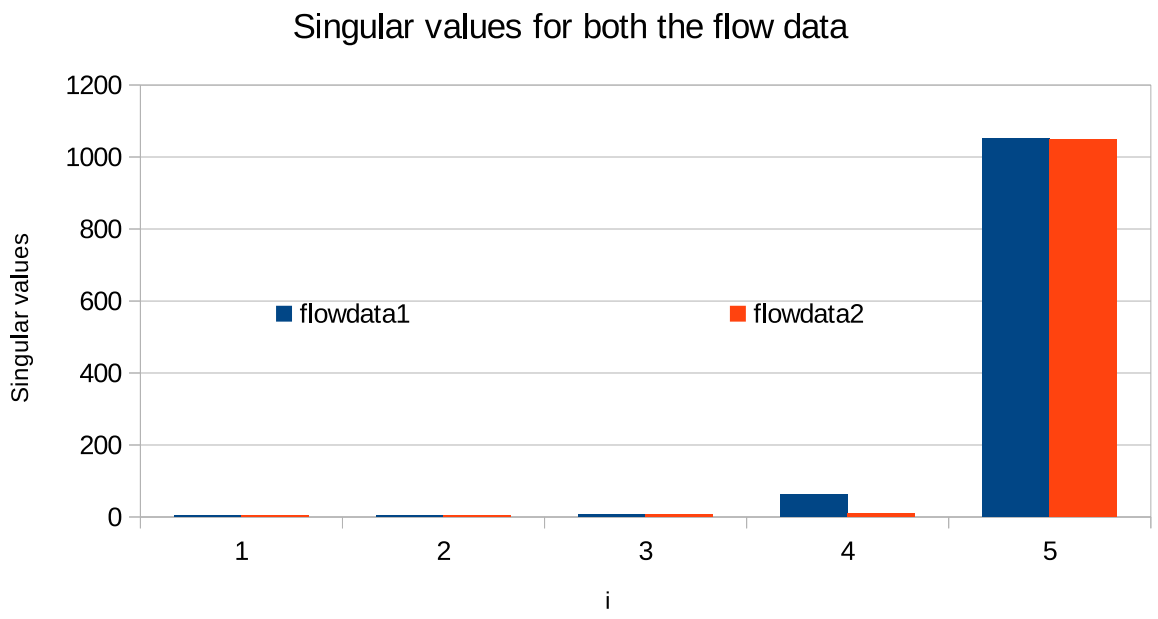

Figure 7: Singular value plot

Part c

Flow data set 1:

On applying autoscaling for the true flow rates, the regression matrix is obtained as below:

$$B = \begin{bmatrix} 0.6667 & 0.5333 \\ 0.5000 & 0.4000 \\ 0.5556 & -0.0000 \end{bmatrix}$$

This is proportional to the true regression matrix obtained before. It is just scaled since our data is scaled. The regression matrix for the autoscaled measurement data of the first set is below,

$$B^* = \begin{bmatrix} 0.6647 & 0.5326 \\ 0.4954 & 0.3963 \\ 0.5573 & -0.0010 \end{bmatrix}$$

Flow data set 2:

The regression matrix from the true data is,

$$B = \begin{bmatrix} 0.6667 & 0.5333 \\ 0.5000 & 0.4000 \\ 0.5556 & 0.0000 \end{bmatrix}$$

The regression matrix for the autoscaled measurement data of the second set is below,

$$B^* = \begin{bmatrix} 0.6549 & 0.5453 \\ 0.5202 & 0.3885 \\ 0.5648 & 0.0026 \end{bmatrix}$$

Autoscaling nearly gives comparable absolute errors in the regression matrices. They are off in some dimensions. Overall they may be used as a better model.

Part d

The number of independent variables can be chosen by knowing the underlying physics of the problem. Obviously, (F_1, F_3),(F_3, F_4) are bad choices of the independent variables since they represent the same value. The best choices for independent variables are (F_1, F_3), (F_1, F_2) ,(F_1, F_3), (F_2, F_5),(F_4, F_5),(F_3, F_5) etc.

Unscaled data, Z_I= (F_1, F_2)	Measured (B^*)	True (B)	abs($B - B^*$)
Flow data 1	$\begin{bmatrix} 0.9993 & 1.0004 \\ 1.0051 & 0.9961 \\ 1.0046 & -0.0052 \end{bmatrix}$	$\begin{bmatrix} 1 & 1 \\ 1 & 1 \\ 1 & 0 \end{bmatrix}$	$\begin{bmatrix} 0.0007 & 0.0004 \\ 0.0051 & 0.0039 \\ 0.0046 & 0.0052 \end{bmatrix}$
Flow data 2	$\begin{bmatrix} 0.9203 & 1.0798 \\ 0.9453 & 1.0546 \\ 1.2343 & -0.2345 \end{bmatrix}$	$\begin{bmatrix} 1 & 1 \\ 1 & 1 \\ 1 & 0 \end{bmatrix}$	$\begin{bmatrix} 0.0797 & 0.0798 \\ 0.0547 & 0.0546 \\ 0.2343 & 0.2345 \end{bmatrix}$

Table 6: Summary of all the regression matrices for unscaled data

Autoscaled data Z_I= (F_1, F_2)	Measured (B^*)	True (B)	abs($B - B^*$)
Flow data 1	$\begin{bmatrix} 0.6647 & 0.5326 \\ 0.4954 & 0.3963 \\ 0.5573 & -0.0010 \end{bmatrix}$	$\begin{bmatrix} 0.6667 & 0.5333 \\ 0.5000 & 0.4000 \\ 0.5556 & -0.0000 \end{bmatrix}$	$\begin{bmatrix} 0.02 & 0.0007 \\ 0.0046 & 0.0037 \\ 0.0017 & 0.0010 \end{bmatrix}$
Flow data 2	$\begin{bmatrix} 0.6549 & 0.5453 \\ 0.5202 & 0.3885 \\ 0.5648 & 0.0026 \end{bmatrix}$	$\begin{bmatrix} 0.6667 & 0.5333 \\ 0.5000 & 0.4000 \\ 0.5556 & 0.0000 \end{bmatrix}$	$\begin{bmatrix} 0.0118 & 0.5333 \\ 0.0202 & 0.4000 \\ 0.0092 & 0.0026 \end{bmatrix}$

Table 7: Summary of all the regression matrices fo autoscaled data