Chem 623 HW#3 (20pts total scaled from 100 pts)

1. (30 pts) (Modified From Brereton) quantitative structure-activity relations (QSAR) and related areas rely on PLS regression. The goal here is to build a predictive model for the property A1 for a benzodiazepine compound using the retention of this compound measured on several chromatographic stationary phases. The data are in HW#3-1.csv .

(a) Load and standardize the data, then do cross-validated PLS-1 on these data. Use a leave-one-out cross validation. Plot the RMSECV vs number of LV in the model. Where is the optimum number of LVs? What is the approximate expected error in the property estimate? Why is standardization the proper choice here?

(b) Select the optimal number of LVs, rebuild a PLS-1 model and plot the leverage vs. Studentized residuals (yes, you'll need to calculate these). Are there any suspect samples in your model? If so, remove any suspect samples and repeat the analysis in (a) and (b).

(c) Plot the measured property versus the predicted property for the 13 samples.

Compound	**Property**	**Retention parameter for bonded phases**					
	A1	**C18**	**Ph**	**CN-R**	**NH2**	**CN-N**	**Si**
1	-0.39	2.90	2.19	1.49	0.58	-0.76	-0.41
2	-1.58	3.17	2.67	1.62	0.11	-0.82	-0.52
3	-1.13	3.20	2.69	1.55	-0.31	-0.96	-0.33
4	-1.18	3.25	2.78	1.78	-0.56	-0.99	-0.55
5	-0.71	3.26	2.77	1.83	-0.53	-0.91	-0.45
6	-1.58	3.16	2.71	1.66	0.10	-0.80	-0.51
7	-0.43	3.26	2.74	1.68	0.62	-0.71	-0.39
8	-2.79	3.29	2.96	1.67	-0.35	-1.19	-0.71
9	-1.15	3.59	3.12	1.97	-0.62	-0.93	-0.56
10	-0.39	3.68	3.16	1.93	-0.54	-0.82	-0.50
11	-0.64	4.17	3.46	2.12	-0.56	-0.97	-0.55
12	-2.14	4.77	3.72	2.29	-0.82	-1.37	-0.80
13	-3.57	5.04	4.04	2.44	-1.14	-1.40	-0.86

2. (35 pts) This problem requires PCA (Please do the labels for the data in the plots!) and both PLS-1 and PLS-2. To get PLS-2, you will supply more than one column for the y-block, so be sure to use PLS with the NIPALS algorithm (not PLS based on SIMPLS, which won't do PLS-2).

The problem is one that occurs all the time in foods analysis and quality assurance. Suppose that we have 8 blends of cocoa. We have obtained a set of assessments from a taste panel – a group of experts who taste the sample and give it "scores" (no, not PC scores) on several sensory variables that we pre- selected. We also ran analyses of the cocoa, sugar and milk in each of the samples. Thus, our data (HW#3-2.csv) looks like the following:

Sample	*Ingredients*			*Assessments*					
	%COCOA	%SUGAR	%MILK	Lightness	Color	Cocoa-odor	Smooth-text	Milk-taste	Sweetness
1	20.00	30.00	50.00	44.89	1.67	6.06	8.59	6.89	8.48
2	20.00	43.30	36.70	42.77	3.22	6.30	9.09	5.17	9.76
3	20.00	50.00	30.00	41.64	4.82	7.09	8.61	4.62	10.50
4	26.70	30.00	43.30	42.37	4.90	7.57	5.96	3.26	6.69
5	26.70	36.70	36.70	41.04	7.20	8.25	6.09	2.94	7.05
6	26.70	36.70	36.70	41.04	6.86	7.66	6.74	2.58	7.04
7	33.30	36.70	30.00	39.14	10.60	10.24	4.55	1.51	5.48
8	40.00	30.00	30.00	38.31	11.11	11.31	3.42	0.86	3.91

Our aim is to predict the sensory assessments from the chemical analysis. After all, instruments are cheaper and more precise than expert testers.

(a) Begin by standardizing the data over all of the variables. Do PCA separately on the 8 x 3 ingredients **(X)** matrix and the 8 x 6 assessments **(Y)** matrix. In each analysis, keep 2 PCs and plot the scores and loadings for the data.

(b) Now do six separate PLS-1 for the 6 y-block variables against the 3 x-block variables. In each of the PLS-1 runs, keep 2 LVs, and plot measured vs. predicted values for the y-block variable. Make a table of RMSECV (2 components) vs. assessment

(c) Obtain the set of predicted y-block variables (you will need to convert these back to the original measurement space) and calculate the correlation coefficients between the observed variables and the predicted variables. Plot the percent RMSECV vs the correlation coefficient. Comment on this plot.

(d) Repeat the above, using PLS-2 with NIPALS to calibrate all 6 y variables at once, and regenerate %RMSEP values as above. Which method produces better estimates, PLS-1 or PLS-2? Explain your results.

3. (35 pts) The Shootout 2 mixture (HW#3-3.csv) includes water, isopropanol (IPA), tert-butyl alcohol, (TBA) and acetone.

(a) Note that this dataset has a design and is closed, in that the components add to 100%. What should the rank of the data matrix be? What is the observed rank? How many LVs are expected purely from chemical considerations?

(b) Using the data in HW#3-3.csv, load the "cal" samples (1-24), and choosing preprocessing and outlier removal as needed, develop a PLS-1 model to predict acetone. Select the optimal number of components by repeated, random subset cross-validation (Explain why leave one out CV is ***NOT*** ok here).

(c) Plot the loading weights for the LVs you keep, and also plot the regression vector. Explain these plots, given that acetone is present at intermediate levels and is polar but not an -OH containing compound.

(d) Calculate the RMSECV for acetone from your model. Then apply this model to the test set, "val" samples 1-12 and calculate the RMSEP for this set. Plot the measured versus predicted relationship for the validation set. Are the cross-validation and external validation errors commensurate? If not, why is there a difference? Discuss the results briefly.