

Problem 1. Analyze the “alcohol” data set with Principal Components Analysis (PCA).

The “Alcohol” data set consists of 65 samples with 52 variables. This data is the clinical evaluation of blood and urine assays for residents of an alcohol treatment program (33 patients) and a suitable group of non-alcoholic individuals (32 patients). Analyze the data by PCA and determine if there are any systematic differences between the alcoholic and non-alcoholic patients.

Load the data into the MATLAB workspace by typing

```
>>load alcohol
```

at the prompt. The data loaded consists of the data matrix `alc`, and matrices containing the sample identifiers `class` and variable identifiers `vars`. You can verify this using the MATLAB `whos` function. The PLS_Toolbox functions `mncn`, `auto`, `pca`, `pltloads` and `pltscrs` will be used. For help on any of these functions, type `help` and the function name at the prompt.

Answer the following questions, supplying score and loadings plots where appropriate.

- a. Before analysis, what would you expect the rank of the data matrix to be? Justify your answer.
- b. Estimate the rank of the raw, mean centered, and autoscaled data. This can be done by observing the relative variance captured by each PC of a PCA model of the data. What are your recommendations for preprocessing of the data? Why?
- c. Variable #20 (U-Glucose) has a very high loading on PC#2 for the raw and PC#1 for the mean centered data, but does not stand out of the autoscaled data. Is variable #20 an outlier? Why does this discrepancy between preprocessing methods occur?
- d. How many PCs are needed to distinguish between the two classes of patients? Is this different from the rank of the system? Explain.
- e. Is there much correlation among the variables? Support this with bi-variate plots of the loadings.
- f. Often, for clinical analyses, each individual measurement is very expensive. Are all 52 measurements (variables) needed? How would you choose a subset of variables (say 6-12) that still provide good discrimination between classes? Choose a good subset. What effect does using this subset have on the discrimination between classes? Support your conclusions with scores plots.

Problem 2. Analyze the archeological data set “arch” with Principal Components Analysis (PCA).

The “arch” data set consists of 75 obsidian samples from several quarry sites. The variables are the concentrations of 10 elements as measured by x-ray fluorescence. The purpose of analyzing this data is to develop a classification model from the quarry data based upon the first 63 samples in the data set which come from 4 quarries. Once the classification model is developed, you must then predict the most likely source for the last 12 samples, which are artifacts. Note that you will probably not be able to predict the source of all the remaining artifacts based on a single PCA model.

Load the data into the MATLAB workspace by typing

```
>>load arch
```

at the prompt. The data loaded consists of the data matrix `arch`, and matrices containing the sample identifiers `samps` and variable identifiers `vars`.

Answer the following questions, supplying score and loadings plots as warranted.

- a. What type of preprocessing is appropriate for this data? Explain.
- b. Create a model using all of the training samples, *i.e.* the first 63 samples. Which PCs are useful in discriminating between the 4 quarries?
- c. Create a model without using the data from the ANA quarry. Which PCs are useful in discriminating between these 3 quarries?
- d. What are the advantages and disadvantages of using each of the two models for discrimination and interpretation? Is it possible to combine the two models?
- e. Which quarry did each of the artifacts originally come from? (Show no more than 3 plots for support.)
- f. Is there much correlation between the variables? Support this with bivariate plots of the loadings.
- g. Using bivariate plots of the scores and loadings, which elements are conspicuous (by their presence or absence) in each of the three quarries? Confirm this with plots of the loadings.

- h. Can a good model for prediction be constructed using just the elements listed in question g? What effect could “extra” variables have on the model? (Hint: Consider question b.)
- i. Extra credit: Develop separate PCA models on each class of samples. Project the unknown samples into each of the 4 PCA models. Confirm your previous assignment of the samples to each class, if possible.

Problem 3. Develop appropriate PCR and PLS calibration models for the dairy data set.

The dairy data set consists of 140 samples of brick cheese. There are 14 variables measured on each sample and 2 analytes of interest. The spectra were collected at 12 discrete wavelengths between 900 and 1100 nm. Also included as variables are two temperature terms (variables 13 and 14). The goal of this exercise is to develop models that are predictive for the fat and moisture content of the cheese bricks based on the spectra.

The data can be loaded into MATLAB by typing

```
>>load dairy
```

at the prompt. The data loaded consists of the data matrix predictor variables `xdat`, and property variables `ydat`, a matrix containing the sample identifiers `samps`, and matrices with the predictor variable identifiers `xvars` and properties `yvars`. You will probably want to use the `modlmker` or `modlgui` function to develop the models. The `delsamps` function may also be useful for editing the data set.

While constructing the calibration model, address the following concerns and support your comments:

- a. Use the plots of leverage vs. studentized residuals to check the data set for outliers. Which samples did you suspect to be outliers? Which samples would you treat as outliers? Why?
- b. A number of samples have a studentized residual error of prediction greater than 2 standard deviations. How many samples would you expect to have a studentized residual greater than this?
- c. What is the optimal number of factors in the PLS and PCR models? Does the number of factors in the models differ for each analyte? Why?
- d. Observe the difference in the variance captured by the PCR and PLS models in both the x- and y-blocks. The x-block variance in the PCR model is monotonically decreasing. This is not the case for PLS. Why not? How do the y-block variance captured numbers compare for the models? What about the sum of the x- and y-block variance captured for each of the models?
- e. Do the temperature terms appear to add to the predictive ability of the model? Substantiate your response.

- f. Assertion: “If two models, *e.g.* one by PCR and one by PLS, have comparable prediction errors for a data set, then the regression vectors from the two models should be comparable. That is, the regression vectors from the two models will give comparable weights to each variable.” Defend or contradict this statement.

Helpful hints:

Choosing the correct number of factors in PLS and PCR models is not always easy. A good way of choosing is to use look for a “knee” in the PRESS plot. Another good rule of thumb is to not add any factor that reduces the PRESS by less than 2%.

Problem 4. Compare an inverse least squares procedure (like PLS or PCR) with classical least squares on the `nir_data` set. Before attempting the assignment, review classical least squares (CLS) regression (see for instance the PLS_Toolbox manual pages 16 to 17).

The `nir_data` set contains the measured near infrared absorbance spectra on 30 pseudo gasoline samples which are a mixture of 5 components. The data can be loaded into MATLAB by typing

```
>>load nir_data
```

at the prompt. The data loaded consists of two matrices of spectra, `spec1` and `spec2`, of the absorbances of 30 samples measured on two different spectrometers, a matrix containing the concentrations of the 5 components of the mixtures `conc`, and a vector containing the wavelengths corresponding to the spectra `lambda`. In this exercise you will use only `spec1`.

- a. Develop separate PLS models on each of the 5 analytes using only the first 20 samples in the data set. What was your choice of scaling, and why? How many factors (PLS latent variables) did you use in the models? Justify your choices.
- b. Calculate the RMSEP (root-mean square error of prediction) for the remaining samples for each of the analytes. Report your results.
- c. Develop a CLS model on the first 20 samples from `nir_data`. Plot your estimates of the pure component spectra. Did you use any scaling of the data? Why or why not?
- d. Predict the concentration of the five analytes in the final 10 samples. Calculate the RMSEP. Compare to the RMSEP for the PLS models developed above.
- e. Develop a CLS model on the first 20 samples using only the first 4 analytes, *i.e.* leave the fifth analyte out. Compare the estimates of the spectra obtained with the estimates from the CLS model based on all 5 analytes. How do the spectra compare?
- f. Use the CLS model from part e to predict the concentrations of the first 4 analytes in the remaining 10 samples. Report the RMSEPs for the samples and compare your results to those obtained with the PLS models for the first four analytes. What is the cause of the large difference?

Problem 5. Use the plsdata set and do separate prediction on the test set.

Problem 6. The XRF data set.

Problem 7. Use the cluster function on ???

Problem 8. Use the cr function on the PLS_data set

Problem