Dimensionality Reduction Exercises

You will need

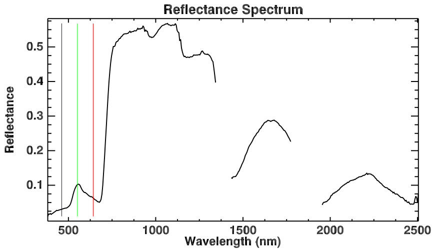
The R Software package from http://www.r-project.org/ or another numerical computing package. Scipy, MATLAB or Octave will work great. If you're using R, you may also want the MASS library – you can download it with R's built-in package installer.

Datasets posted on the course webpage https://class.coursera.org/bigdataschool-001/wiki/Day_9

- **Citrus_Identification_Data.txt** – reflectance spectra for citrus and noncitrus plants
- **Citrus_Identification_Wavelengths.txt** –wavelengths for (a), in nanometers
- **Citrus_Variety_Data.txt** – reflectance spectra for different citrus varieties
- **Citrus_Variety_Wavelengths.txt** - wavelengths for (c), in nanometers

Exploring high-dimensional data

Here we will be interpreting the data in Citrus_Identification_Data.txt. The rows represent spectra of plants observed by an airborne instrument. The first three columns represent: the spectrum number, a binary classification of the plant as either "Citrus" or "NonCitrus", and the number of the associated plant. The remaining columns are the reflectance spectrum, showing the properties of the plant canopy at different wavelengths. It is common to visualize these spectra by graphing reflectance as a function of wavelength. Here's an example:



1. Plot a few example spectra from random rows. Can you see obvious differences in reflectance spectra across plants or species? Are there meaningful correlations between the different wavelengths? How does one know?

2. We might benefit from some more sophisticated visualization to really understand what's going on in this dataset. Perform a Principal Component Analysis (PCA). How do the eigenvalues decay? What does this tell us about the structure of the dataset and the applicability of PCA?

3. Plot some examples of the PCA bases associated with the highest eigenvalues. What does this tell us about our dataset?

4. Plot the datapoints after projection onto top principal components (e.g. component 1 vs component 2, 1 vs. 3, etc.). Find the datapoints located at the extrema, and plot the associated reflectance spectra. What does this tell us about the structure of the data?

Bonus: Citrus/NonCitrus labels are provided. With leave-one-plant-out cross validation, how well can you predict the species of a plant based on its spectral properties alone?

Incorporating classification information

In this exercise we will use linear discriminant analysis to incorporate the class categories in Citrus_Variety_Data.txt. The first three columns represent: the spectrum number, a grouping of the citrus plant into one of five different citrus varieties, and the number of the associated plant. The remaining columns represent the reflectance of the plant canopy at different wavelengths as before.

1. First, start as with a scatterplot of the PCA-projected data. Indicate the class membership with colors or some other obvious marking. How well-separated are the citrus varieties in the low-dimensional space?

2. Perform a multiclass linear discriminant analysis on the dataset. How many different basis directions can you define? Plot the remaining data by projecting it onto the LDA dimensions. Here it will be important to visualize the transformation using a held-out "test" set to ensure that our interpretation is unbiased. Be sure to not mix the same plants in your training and test data. According to this analysis, what varieties are most spectrally similar?

Bonus: How well can you predict the *variety* of a citrus plant based on its spectral properties? What cross-validation strategy is most appropriate to answer this question?

Super Bonus: How well can you predict the specific *plant* based on spectral properties alone? What cross-validation strategy is most appropriate to answer this question?