Partitioning data cube

- First get the datacube (26×502×1535), where the first dimension is the 24 excitation wavelengths plus 2, the second dimension is 501 emisison wavelengths plus 1, and the third dimension is the number of observations.
- Partition the datacube into (1) train, (2) valid, and (3) test sets. Then we obtain train set (26×502×515), valid set (26×502×397), and test set (26×502×623).
- The resulting 3 sets (cubes) are in partition_data.mat (for MATLAB).

Converting train, valid, and test sets (cubes) into matrices

- For each of the three cubes, extract matrices of biographical information (bio), emission/excitation pairs (exem), and the measured spec values corresponding to exem pair (spec).
- For example, for train set, the size of bio is 515×25, the size of exem is 2×12024 (where 12024=24*501), and the size of spec is 515×12024.

Removing NaN's in spec matrix

- Some columns of spec matrix contain all NaN's (which means nobody has a measurement at certain exitation/emission pairs) and some columns contain some NaN's (the measurement at a given ex/em wavelength was obtained for somebody but not for everybody).
- We decided to throw away columns that contain all NaN's (6624 out of 12024) and to further investigate the columns that contain some NaN's (1823 out of 12024).
- Upon further investigation, we concluded that we can throw away any columns that contain any NaN's.
- The final, raw spec matrix (for train set) is 515×3577 (where 3577=12024-6624-1823). Also created the corresponding exem matrix (2×3577) where now there are 16 excitation wavelengths and diffeeing number of emission wavelengths within each excitation wavelength.

Smoothing spec matrix

- When first inspecting the countour plots of spec matrix, it was determined that smoothing may be desired.
- For each of the 515 observations, we first obtain average values for each of 16 excitation wavelengths, so that we get 515 by 16 matrix of average values. Then we smooth over 16 values, and obtain a matrix of smoothed values of size 515 by 16. We then multiply each component of spec matrix by (smoothed value/averaged value) corresponding to correct observation and excitation wavelength.

- When smoothing, we have decided to use fixed smoothing parameter chosen by inspection, because the automated method for parameter selection yielded non-satisfactory result.
- The final smoooth matrix as well as other items of interest (bio, exem, spec) are in train_final.mat.
- NOTE: When plotting, log10 transform is used.

Principal component analysis

- First, get the centered spec matrix by subtracting off the column means of spec.
- Perform SVD on centered spec matrix, and obtain eigen vectors (V) and square root of eigenvalues (diag(D)).
- Determine the number of eigenvalues that are 99.9% of total. It's 380.
- Save 380 eigenvectors and the first 380 diagonal values of D.
- Multiply entered spec matrix by 3577×380 reduced V matrix, to get a principal component matrix, pc_matrix (515×380). (or just get U*D).

SVM

- Classify 515 obs according to if they have HG (+1) or (normal or LG) (-1). We get this from the 9th column of bio matrix.
- When we feed the data (pc_matrix) into the SVM, the sensitivity (classifying HG as normals) is 0% and specificity is 100%.
- Even when we tried to see classification performance of HG versus normals (leaving out LG), we still get the same result.
- Will try kernel logistic regression and PSVM (ridge regression).

Contents of .mat files

- train_final.mat
  - bio (515 X 25) The 25 biogrphical information for each of the 515 obs.
  - exem (2 X 3577) The valid excitation/emission pairs. THe first row is the excitation, and the second row is the emission.
  - spec (515 X 3577) The matrix of 515 obs by 3577 values corresponding to exitation/emission pairs. This is smoothed (see Smoothing spec matrix section above).
- train_pc.mat
  - D_pc (381 X 1) The square roots of the eigenvalues of cov(X), X=spec matrix.
  - V_pc (3577 X 380) The first 380 eigenvectors of cov(X).
  - spec_mean (1 X 3577) The column means of spec matrix.
- train_centered.mat
  - spec_centered (515 X 3577) The spec matrix with each column subtracted by column mean.
- train_pc_matrix.mat
  - pc_matrix (515 X 380) The principal component matrix.