Mathematical Tools for Neuroscience (NEU 314)
Princeton University, Spring 2016
Jonathan Pillow

Homework 4: PCA and Regression

Due: Friday, Mar 25, 11:59pm

1 Principal Components Analysis (PCA)

Loadthefile PCA.mat into your MATLAB environment. You'll find a matrix M , which contains the responses of four neurons to a variety of different stimuli. Each row of M gives you the spike count of the four neurons to one of the 100 stimuli. We cannot visualize the data in this form, but would like to know how the neurons as a population are encode the stimulus space.

1. Plot the raw data. Make a plot showing the response of each of the four neurons to each of the 100 stimuli, so the x axis is "stimulus" and y axis is "neural response". Use a different line color for each neuron and add a legend using the `legend` command, with labels 'neuron1', ... 'neuron4'.

2. Start by computing the covariance of the responses. You can do this using "cov", or, by subtracting off the mean response from each neuron (each column of M) and computing $M^T M/(n-1)$. Compute the eigenvalues, λ_k, of the covariance matrix, and plot them as a function of k, for $k = 1, 2, 3, 4$. Do the data points live close to a subspace of dimensionality less than four?

3. Look at the axes of the subspace where the neural responses are varying most (i.e., the eigenvectors corresponding to the largest eigenvalues). How would you describe these? Which neurons appear to have the most similar stimulus tuning?

4. Project the data in M onto the first principal component (i.e., compute the inner product of the data vectors with the eigenvector corresponding to the maximal eigenvalue). Plot a histogram (using `hist`) of these values. Show that the sum of squares of these values (divided by $n-1$) is equal λ_1. What proportion of the total variability of the data (sum of squared data vector lengths) does this component account for?

5. Show a scatter plot of the data projected onto the first two principal components (that is, plot the dot product of the data with the first component versus the dot product with the second component). Use `plot`, requesting circular plot symbols and no connecting lines. Use '`axis equal;`' to set the two axes to use equal scales.

6. Show that the sum of the squared lengths of these projected vectors (divided by $n-1$) is equal to $\lambda_1 + \lambda_2$. What proportion of the total variability of the data do these two components account for?

7. Finally, for comparison, make a scatter plot of the data projected onto the first and third principal components. (Use '`axis equal`' in matlab to make x and y axes have the same scaling.)

Least-Squares Regression

8. **Polynomial regression**. Download the file `r`egress1.mat into your MATLAB environment. Scatter plot variable Y as a function of X. Find a least-squares fit of the data with polynomials of order 0 (a constant), 1 (a line), 2, 3, 4, and 5.

   For polynomial regression of order k, generate the design matrix M as

   $$M = [\vec{1}, \vec{X}, \vec{X}^2, \ldots, \vec{X}^k],$$

   where $\vec{1}$ is the vector the same length as $\vec{X}$ consisting of of all ones, and $\vec{X}^p$ is the vector $\vec{X}$ with all elements raised to the power p. Then find the regression coefficients $\vec{w}$ by minimizing $||\vec{Y} - M\vec{w}||^2$.

   Plot the polynomial fit of each order for a grid of x values ranging from $x = -1$ to $x = 5$ in increments of 0.1. (To do this, make a column vector `xx` out of the grid points, and place them in a matrix with `x.^p` in each column, just like you did to form M above, and multiply it by the fitted weights $\vec{w}$.) (5 points)

9. On a separate graph, plot the squared error as a function of the order of the polynomial. Which fit do you think is "best"?

10. Verify that the "residual" error vector $Y - M\vec{w}$ is orthogonal to "linear prediction" vector $M\vec{w}$ for the polynomial you deemed to be best-fitting.