Final Exam Preparatory Questions

Math 307 Section 202 Jan-Apr 2016

The following questions will be presented at the final exam at 8:30am on Wednesday 27 April 2016. In addition to these questions, there will also be a short-answer section.

The same rules apply to this exam as applied on the midterm: You may work with your classmates and consult any materials you wish in order to prepare. At the exam, you may bring only writing implements and the plots and codes specified below.

Read and answer these questions carefully as a part of your exam preparation.

Question 1

This question is about the outcome of having two different organisms competing in the same environment. We model the populations u and v at generation k according to the difference equation

$$\begin{aligned} u_{k+1} &= a u_k - (1+\epsilon) v_k, \\ v_{k+1} &= -u_k + a v_k. \end{aligned}$$

We constrain $a \geq 1$. What does each term on the right-hand-side represent? What does a represent? What does ϵ represent? Why is *competition system* a good name for this model?

Suppose $\epsilon = 0$. What does this mean (in words) about the inhibition balance between the two populations? Set $u_0 = v_0 = 1$ and use eigenvalue analysis to predict what will happen to the populations after a long time. Provide code and a plot of u and v versus k to demonstrate the behaviour you predict.

Suppose $\epsilon > 0$. What does this mean about the inhibition balance between the two populations? Again set $u_0 = v_0 = 1$ and use eigenvalue analysis to predict what will happen to the populations after a long time. Provide code and a plot of u, v to demonstrate your prediction.

Describe the sensitivity of this competition system to ϵ, and speculate on what this model

can tell us about more elaborate competition systems.

Question 2

This question is about comparing modeling using least-squares fitting to modeling using the SVD. The dataset we'll be using represents qualities of music originating in different parts of the world, and we're attempting to see if these data illustrate whether or not music style is correlated with its place of origin.

Getting the data

Visit the Geographical Origin of Music dataset at the UCI data archive and click on the data folder link. This link directs you to download a zip file. Unpack the zip file and import the file `default_features_1059_tracks.txt` into your favourite programming environment. (There is a second set of data in this zip archive – you won't need it for this question.)

This is a CSV file having 1059 lines. Each line represents a sample of music. The first 68 entries in each line are 68-dimensional vectors representing musical features. The last 2 entries are the latitude and longitude of the location of the music's origin.

For each music sample $i \in 1 \ldots N$, we will denote pairs (p_i, q_i) as the music origin's latitude and longitude, and we will call vector $\mathbf{x}_i \in \mathbb{R}^{68}$ the feature vector.

$$X = \begin{bmatrix} \mathbf{x}_1^T \\ \mathbf{x}_2^T \\ \vdots \\ \mathbf{x}_N^T \end{bmatrix}$$

is the feature matrix. Bring to the exam the code you use to import these data.

Predict music origin using a linear model

Perform least-squares fits of matrix X against latitude, and of X against longitude. That is, suppose that

$$p_i \approx \mathbf{x}_i^T \mathbf{a},$$
$$q_i \approx \mathbf{x}_i^T \mathbf{b},$$

and solve the two resulting least-squares problems for the unknown model parameter vectors $\mathbf{a}, \mathbf{b}$. Note that matrix X is not full-rank. You must comment on the impact this has on least squares, as well as an appropriate fix, for full marks.

Once you have model vectors $\mathbf{a}, \mathbf{b}$, determine the latitudes and longitudes they predict for each music sample, namely,

$$\hat{p}_i = \mathbf{x}_i^T \mathbf{a},$$
$$\hat{q}_i = \mathbf{x}_i^T \mathbf{b}.$$

(Note the hats representing *predicted* p's and q's). Bring to the exam the codes you use for doing the least-squares fit and the prediction.

Plot the actual latitudes and longitudes on a scatterplot, and also plot the predicted latitudes and longitudes. (The actual latitudes and longitudes should roughly remind you of a map of locations on the world map – if it doesn't look right, try swapping the x-y axes.) Bring the two scatterplots (or a single plot, if you used different symbols for actual and predicted values) to the exam. (You do not need to provide plotting code, just the plots.)

Is the proposed linear model a good model for predicting the location of music origin based on musical features? Explain.

Suppose you doubled the size of each element of $\mathbf{x}_i$. If the least-squares model is correct, what would you expect to see happen to the predicted latitude and longitude?

Summarize the music dataset using the SVD

Take the Singular Value Decomposition (SVD) of the feature matrix. Plot the singular values versus their order number as a semi-log plot. (Bring this plot to the exam.) What is the effective rank of this matrix? What does this effective rank tell you about the data? Why was doing least-squares so challenging for this dataset? Can you interpret the four subspaces of matrix X? (Some interpretations are awkward – do your best.)

Group the latitude and longitude data: How many distinct locations are represented in these data? (You could also look at the "world map" data you produced in the least-squares analysis.)

Find two sets of rows of data from X that originate from distinctly different places on

Earth. (This is probably the trickiest bit of programming you will have to do, as you will need to find all row indices corresponding to a single (p, q) location.) For each subset of X, project each row vector onto the subspace spanned by the first two right singular vectors of X. You will have two sets of pair-wise data, one set from each of your two chosen locations. Scatterplot both sets on the same set of axes and bring this plot to the exam.

Find a third set of data from X that originates near to one of the locations you identified above. Repeat the projection and plotting procedure, and bring this plot to the exam. (Just as above, you do not need to provide the plotting code, just the plots.)

(In your preparation, you may wish to try these plots for more than three locations in order to generate good plots for discussion.)

Now, make some comments: Why should you expect this projection procedure to provide a summary of these 68-dimensional data? Do these data support the conjecture that the music feature vectors are sensitive to the music's geographic origin? What does the SVD do that a least-squares model cannot? Provide advice for when a least-squares analysis will provide more or less insight than the SVD.