CS/ECE/ME 532
Homework 4: The SVD

In this homework set you will work with an analyze the dasaset `jesterdata.mat`, which is available on the moodle site. The dataset contains an $m = 100$ by $n = 7200$ dimensional matrix X. Each row of X corresponds to a joke, and each column corresponds to a user. Each of the users rated the quality of each joke on a scale of $[-10, 10]$.

1. Suppose that you work for a company that makes joke recommendations to customers. You are given a large dataset X of jokes and ratings. It contains n reviews for each of m jokes. The reviews were generated by n users who represent a diverse set of tastes. Each reviewer rated every movie on a scale of $[-10, 10]$. A new customer has rated $k = 25$ of the jokes, and the goal is to predict another joke that the customer will like based on her k ratings. Use the first $n = 20$ columns of X for this prediction problem (so that the problem is overdetermined). Her ratings are contained in the file `newuser.mat`, also on moodle, in a vector b. The jokes she didn't rate are indicated by a (false) score of -99. Compare your predictions to her complete set of ratings, contained in the vector `trueb`. Her actual favorite joke was number 29. Does it seem like your predictor is working well?

2. Repeat the prediction problem above, but this time use the entire X matrix. Note that now the problem is underdetermined. Explain how you will solve this prediction problem and apply it to the data. Does it seem like your predictor is working? How does it compare to the first method based on only 20 users?

3. Propose a method for finding one other user that seems to give the best predictions for the new users. How well does this approach perform? Now try to find the best two users to predict the new user.

4. Use the Matlab function `svd` with the 'economy size' option to compute the SVD of $X = U\Sigma V^T$. Plot the sprectrum of X. What is the rank of X? How many dimensions seem important? What does this tell us about the jokes and users?

5. Visualize the dataset by projecting the columns and rows on to the first three principle component directions. Use the `rotate` tool in the Matlab plot to get different views of the three dimensional projections. Discuss the structure of the projections and what it might tell us about the jokes and users.

6. One easy way to compute the first principle component for large datasets like this is the so-called power method (see `http://en.wikipedia.org/wiki/Power_iteration`). Explain the power method and why it works. Write your own code to implement the power method in Matlab and use it to compute the first column of U and V in the SVD of X. Does it produce the same result as Matlab's built-in `svd` function?

7. The power method is based on an initial starting vector. Give one example of a starting vector for which the power method will fail to find the first left and right singular vectors in this problem.