CS 322 Homework 5 — Solutions

out: Tuesday 27 March 2007
due: Monday 2 April 2007

SVD detective work

Often one is faced by some multidimensional experimental data in which a lot of measurements are generated from some physical system where the underlying process is simple. For instance, in absorption spectroscopy one might measure absorption spectra (sampled at hundreds of wavelengths) for a number of different samples which are actually just mixtures of a few substances, or one might have camera observations of one moving object from many views. In cases like these the data is redundant, and reducing it to the relevant dimensions is the first task in analyzing it. If we are lucky the redundancy will show up as low-rank structure in the data and we can find the right space for looking at it by using the SVD.

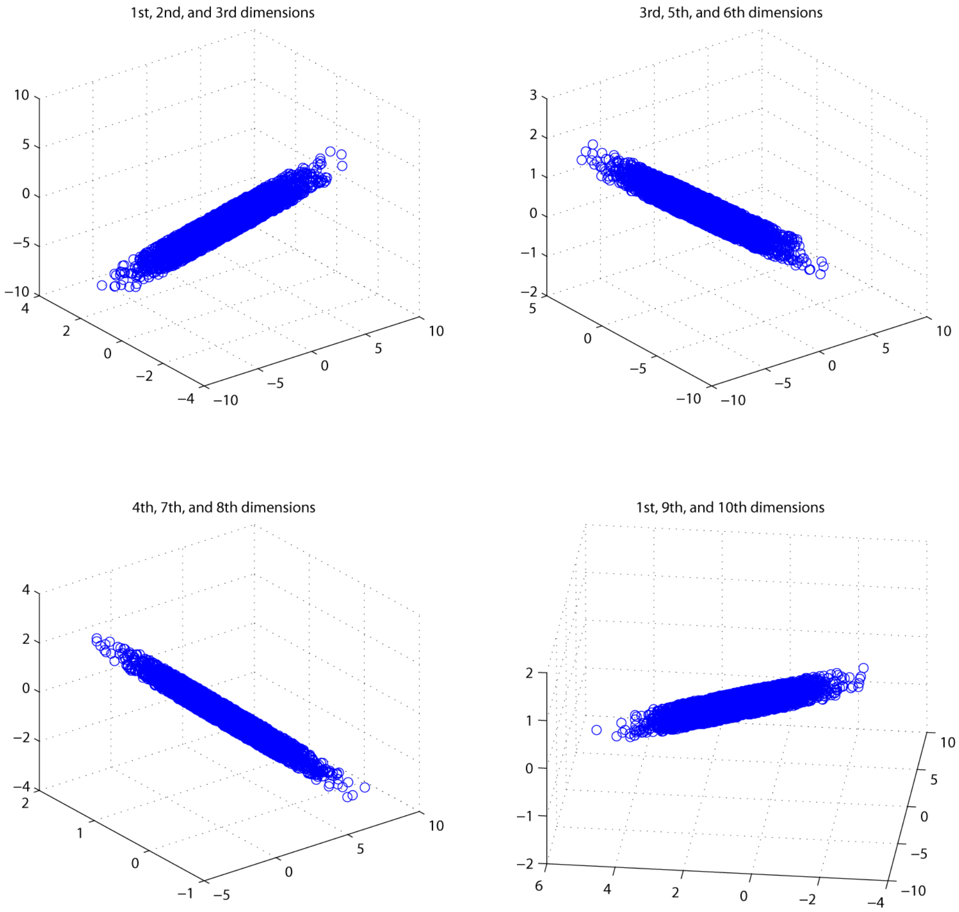
I generated the set of "measurements" in the file `hw5data.txt` by starting with a set of points in $\mathbb{R}^k$ and transforming them into $\mathbb{R}^{10}$ by a linear transformation. Then I contaminated them with a little bit of noise. This closely follows the models often used to analyze experimental data that is expected to be low rank.

Your job is to answer the following questions using the SVD and MATLAB's plotting tools, and to explain how you arrived at your answer.

1. What, if anything, can you tell about the data by looking at 2D and 3D scatterplots of the measurements against one another?

   Answer: Taking the data, picking three of the coordinate axes, and plotting them in a scatterplot shows that there appears to be a strong linear correlation between any three of the axes (four such plots are shown below). This appeared to be true regardless of the axes chosen (although we did not test all 120 such plots). For each plot, the linear correlation appeared strongest in one particular direction, with the magnitude of the spread in the orthonormal directions much less in magnitude. Beyond this, though, there is very little indication of some lower dimensional process occurring.

2. What is the dimension of the space the data started in? What assumptions did you need to decide this?


1st, 2nd, and 3rd dimensions

3rd, 5th, and 6th dimensions

4th, 7th, and 8th dimensions

1st, 9th, and 10th dimensions

Figure 1: Graphs of the 10 dimensional data plotted with respect to 3 canonical axes

Answer: Taking the data as a 6127×10 matrix and performing the SVD, we obtain the following singular values:

$$\begin{bmatrix} 282.2434218521066 \\ 198.6578791758413 \\ 35.1117403019177 \\ 0.1594845289428 \\ 0.1578403396736 \\ 0.1570806031596 \\ 0.1566071477911 \\ 0.1554201199952 \\ 0.1545383094732 \\ 0.1535098026629 \end{bmatrix}$$

We observe that there are two relatively large singular values, a third value that still appears to be significant, and then a bunch of much smaller singular values, all of about the same magnitude. If we assume that the relevant signals are much larger in magnitude than these small singular values, then we can conclude that there are three dimensions of relevant data and seven dimensions of noise, so the data originally started in a three dimensional space.

3. What are the most interesting two axes to project the data on to make a scatterplot? Show the resulting plot.

   Answer: The most interesting axes to plot on are the right singular vectors of the SVD of the data matrix corresponding to the three interesting singular values above. These vectors are:

$$\mathbf{v}_1 = \begin{bmatrix} 0.4437 \\ -0.2015 \\ 0.5729 \\ 0.4349 \\ -0.2733 \\ -0.1157 \\ -0.0672 \\ -0.2388 \\ 0.2865 \\ -0.1166 \end{bmatrix}, \mathbf{v}_2 = \begin{bmatrix} -0.1400 \\ 0.0257 \\ -0.6162 \\ 0.5178 \\ -0.3794 \\ 0.1284 \\ 0.0888 \\ 0.0155 \\ 0.4041 \\ -0.0016 \end{bmatrix}, \mathbf{v}_3 = \begin{bmatrix} 0.1271 \\ 0.8056 \\ -0.0391 \\ 0.0396 \\ 0.1077 \\ -0.4147 \\ -0.0082 \\ -0.0678 \\ 0.1181 \\ -0.3600 \end{bmatrix}$$

   Looking at the attached plots, we can see that the second and third singular vectors are the most interesting axes to plot against, as the true structure of the data is readily apparent — it is a happy face. However, we cannot necessarily discard the first dimension of data, because the magnitude of the signal in that direction is actually stronger than the magnitude in the other two interesting directions.

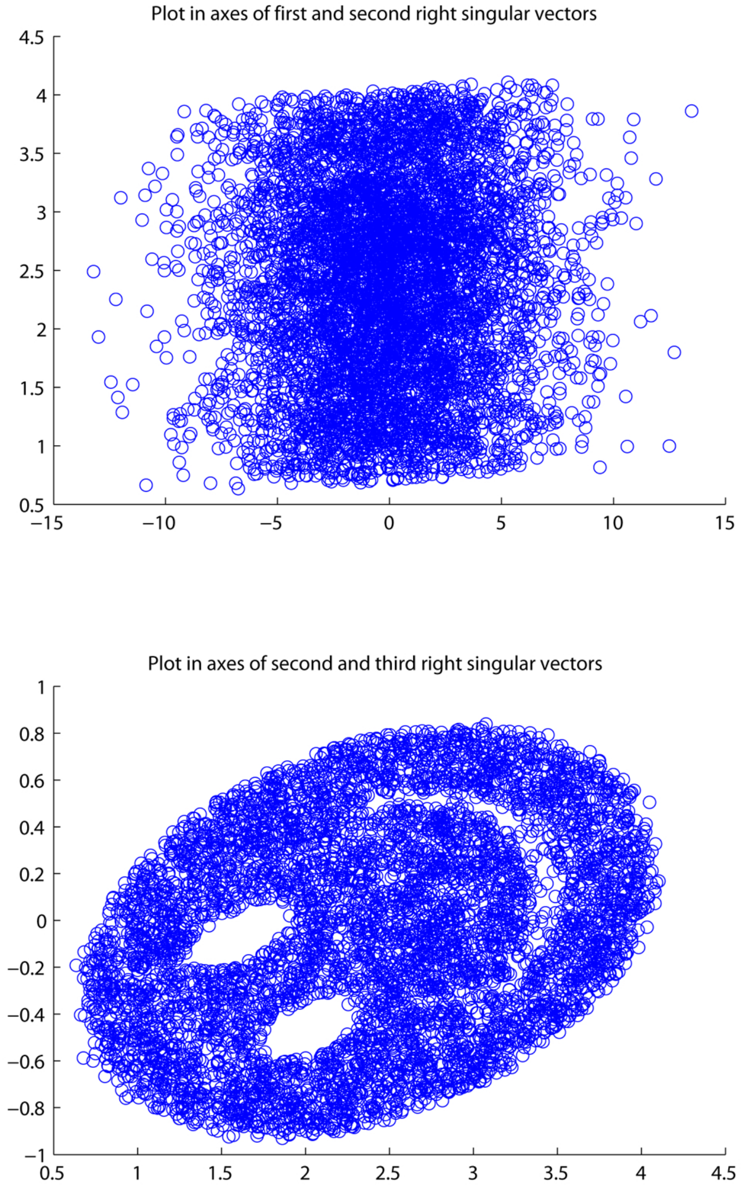


Figure 2: Graphs of the data projected onto the interesting right singular vectors