Matrices and Statistics with Applications
Computer Exercise: SVD

SeSE, LU, SLU, UU

September, 2020

The tasks within this Lab are related to SVD. You are welcome to extend some of the exercises with additional tasks, relevant to the theme of this lab.
At the end, the results of the exercise should be sketched and shown to the lab-responsible.
The tasks require programs and data files, which can be downloaded from
`http://user.it.uu.se/~maya/Courses/NGSSC/index_Stat_2020.html`.
Make a copy of these files in one of your directories.

Exercise 1 (Easy, optional)
Download the files `yeastarray_t2.txt` and `Gene_test.m`.
You have at your disposal the test for the Gene analysis data. You could reproduce the tests and try to analyse the results from the SVD decomposition.

Exercise 2 (Data compression)
Consider the example how to compress an image via SVD. The image is in the file `blue_bird.jpg` and the Matlab code to use is `SVDC.m`. An example how to do analogous operations in R can be found at `https://aaronschlegel.me/image-compression-singular-value-decomposition.html`
You can also see `http://timbaumann.info/svd-image-compression-demo/`
How many dimensions do we need to compress the file with 25%? How can we estimate the quality of the compressed image?
Note: The code is downloaded from Matlab Fileexchange.

Exercise 3 (Eigenvalues vs singular values)
Consider the matrix

$$A = \begin{bmatrix} 0100 \\ 0020 \\ 0003 \\ 0000 \end{bmatrix}.$$

Compute the eigenvalues and the eigenvectors of A. How many linearly independent vectors you see?
Compute the SVD decomposition of A.
Then permute A but changing the entry $A(4,1)$ to $1/60000$ and repeat the experiments.
Comment on the effect of the perturbation of A on its eigenvalues and singular values.

Exercise 4 (Large data sets and covariance matrices)
Load the file `formaya.mat`. It contains four files

```
Name             Size    Bytes      Class     Attributes

 covmatrix     2500x2500    50000000  double
 covmatrix2    2500x2500    50000000  double
 cutout          50x50x50766   253830000  int16
 cutout2         50x50x50766   253830000  int16
```

We will work only with `covmatrix2` and `cutout2`.

Tasks:

1. Consider `covmatrix2`. Find all its eigenvalues and study them.
   Explain the very small eigenvalues, some of which are even negative.
   How many dominant components are observed? Can you illustrate them graphically?

2. The number of nonzero components in the matrix is 5621641 and the number of entries is $2500^2 = 6250000$, thus, the matrix is full. (Note that there are negative entries in the covariance matrix.)

   A study of the size of the entries shows that some of them are very small by absolute value.

   (a) What would be the impact on the PCA if we delete some of the entries, say, that are by absolute value less than 0.005 or 0.05? (This effect can be achieved during the computation of the covariance matrix - we can neglect some small entries upon computation.)

   (b) What properties of the matrix have to be preserved?

3. Consider the data set `cutout2`. Note that it is provided in single precision and if we want to perform algebraic operations with it, we have to convert it to double.

   (a) Estimate the computer resources needed to compute the SVD on the matrix? Note that each cut has very few nonzero elements and the total number of nonzero elements is not $50 * 50 * 50766 = 126915000$ but 26392058.

   (b) Reshape the data as (50766x2500) and (try to) compute its SVD. In Matlab try to create the matrix as sparse.

   (c) Do a performance study: check the time needed to compute SVD on parts of the reshaped data, say 2500x2500, 5000x2500 etc. What is the increase factor in computing time?