Data Mining and Matrices (FSS17)
Assignment 1: Singular Value Decomposition

Preliminaries

Familiarize yourself with R. A good point to start is the R homepage (`http://www.r-project.org/`), the introductory documentation (`http://cran.r-project.org/doc/manuals/R-intro.pdf`), and the examples we provided in file `00-intro.R`.

To compute the thin SVD of a matrix `M` in R, write

```
Msvd <- svd(M)
U <- Msvd$u
V <- Msvd$v
S <- diag(Msvd$d, ncol(U), ncol(V))
```

To reconstruct the rank-k approximation `Mk`, write

```
Mk <- U[,1:k,drop=F] %*% S[1:k,1:k] %*% t(V[,1:k,drop=F])
```

The archive provided to you contains various R scripts, a dataset and its description as well as more detailed instruction about each part of this assignment. For example, we provide a function called `svdcomp` that performs the two steps mentioned above.

Hand in your solutions via ILIAS until the deadline mentioned there. Provide a single zip archive containing:

- A single document that answers all questions (all figures, all analysis of the results, the main commands you used for the analysis if asked, ...) **in pdf format**, and
- Supplementary material containing the transcript of all commands you issued/all source code **in pdf format**.

Name your file `dmm17-a1-<your-uma-login>.zip`.

1 Intuition on SVD

a) Try to manually obtain the rank of each of the following matrices, as well as its singular values, and the left and right singular vectors corresponding to the non-zero singular values. Do this by "looking" at the data and try to infer how the (compact) SVD needs to look like.

Do not use computational methods such as solving the characteristic equations. If you fail at this task for some matrices, just write that you failed and don't worry.

b) Compute the SVD (e.g., using R's `svd` function) and compare. Have you been correct?

c) How does the best rank-1 approximation look like? Is it "intuitive"?

d) How many non-zero singular values does $\boldsymbol{M}_6$ have, i.e., what is the rank of $\boldsymbol{M}_6$? How many non-zero singular values are reported by R? Discuss!

For convenience, each matrix is also defined in the provided R script.

$$\boldsymbol{M}_1 = \begin{pmatrix} 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix} \qquad \boldsymbol{M}_2 = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 2 & 1 & 2 & 0 \\ 0 & 2 & 1 & 2 & 0 \\ 0 & 2 & 1 & 2 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

$$\boldsymbol{M}_3 = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 \end{pmatrix} \qquad \boldsymbol{M}_4 = \begin{pmatrix} 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 \end{pmatrix}$$

$$\boldsymbol{M}_5 = \begin{pmatrix} 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 1 \end{pmatrix} \qquad \boldsymbol{M}_6 = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \end{pmatrix}$$

2 The SVD on Weather Data

Load the `worldclim` data as described in the provided R file. You can find a description of the dataset under `data/worldclim.txt`. When we refer to "data" below, we mean the `climate` matrix.

a) Normalize the data to z-scores (name the normalized data `climate.normal`). Considering the data we are using, are the assumptions for normalizing the data reasonable?

b) Compute the SVD $\boldsymbol{U\Sigma V}^T$ and the rank of the normalized data.

c) Plot each of the first 5 columns of $\boldsymbol{U}$. Use the longitude and latitude of each data point as the x and y coordinates, respectively, and the corresponding entry in the left singular vector to color each point (see provided R file). Can you interpret the result?

d) Plot some scatterplots between the columns of $\boldsymbol{U}$ using colors to distinguish either their North–South or East–West location (see provided R file). Can you interpret the results?

e) Try the different rank selection methods listed below to decide what would be a good rank for a truncated SVD. Report the rank each method suggests (and when subjective evaluation is needed, say why you picked your choice).

   (i) Guttman–Kaiser criterion
   (ii) 90% of squared Frobenius norm
   (iii) Scree test
   (iv) Entropy-based method
   (v) Random flipping of signs

   What, if any, would be your ultimate choice?

f) Define the root-mean-square error (RMSE) between an $m \times n$ matrix $\boldsymbol{A}$ and an $m \times n$ approximation $\hat{\boldsymbol{A}}$ as

$$\mathrm{rmse}(\boldsymbol{A}, \hat{\boldsymbol{A}}) = \frac{1}{\sqrt{mn}} \|\boldsymbol{A} - \hat{\boldsymbol{A}}\|_F .$$

   Create a noisy version of your normalized climate data by adding i.i.d. Gaussian noise from $\mathcal{N}(0, \epsilon^2)$, where ϵ is a parameter that corresponds to the standard deviation of the noise. You can do this in R via

```
climate.noise <- climate.normal + rnorm(prod(dim(climate.normal)))*eps
```

   Do this for various choices of $\epsilon \in [0, 2]$. Now create a plot with ϵ on the x-axis and the RMSE on the y-axis. Add a line for the RMSE between the original data ($\boldsymbol{A}$ = `climate.normal`) and the noisy data ($\hat{\boldsymbol{A}}$ = `climate.noise`). For $k \in \{1, 2, 5, 10\}$, add a line for the RMSE between the rank-k truncated SVD of `climate.normal` and the rank-k truncated SVD of `climate.noise`. Discuss the results.

3 Clustering and visualizing

For this task, our goal is to cluster the rows of the data into five clusters and visualize the result. The entire process is explained in the provided R file and works as follows: We first load the coordinates, then cluster the data (without the coordinates as before) into five clusters using the k-means algorithm, and finally create a plot where the x-y coordinates coorespond to longitude and latitude, respectively, and the color to the cluster identifier.

a) Look at the resulting clustering and explain what the clusters may represent (remember, the data contains temperature and rainfall information).

b) For another visualization of the results, plot the data so that the x-axis position comes from the first left singular vector, the y-axis position comes from the second left singular vector, and the color of each point is defined by the cluster identifier. Are the clusters well-separated from each other in the plot or are they mixed? Do some of the clusters look like outliers?

c) Compute the PCA scores of the data points for the first k principal components for $k \in \{1, 2, 3\}$, thereby reducing dimensionality to k. Do this using the SVD of the appropriate version of the climate data. Repeat the clustering and visualization steps of a) with this new data. Did the results change? Why do you think the results changed or did not change?