D5: Databases and Information Systems
Data Mining and Matrices, SS 2015
Programming assignment #1: SVD and pre-processing
Due: **14 June 2015** at 23:29 CEST

max planck institut informatik

You must hand in machine-typed report in PDF format and a script file (e.g. `*.R` file). The report must explain your approach to the problem, the results you obtained, and your interpretation of the results. Naturally, the report must also answer to any direct question presented in the problems. You can, and in many cases should, add plots and other illustrations to your answers. The script file must show every step you have taken to solve these tasks, that is to say, if we run the script file we must get the same results you reported and see the same figures you presented. You can discuss these problems with other students, and you are encouraged to discuss with the tutor, but everybody must hand in their own answers and own code. Return your answers by email to `dmm15@mpi-inf.mpg.de`*. Remember to write your name and matriculation number to every answer sheet!*

Task 1: Normalization

Download the data and utility files from `http://resources.mpi-inf.mpg.de/departments/d5/teaching/ss15/dmm/assignments/assignment_1.zip`. That package contains file `assignment1.R`. You can fill your answers to that file and return it as a part of your solution.

Follow the steps in the file to load the `worldclim` data. This data contains information about the bioclimatic conditions (minimum, maximum, and average temperature in degrees Celcius and average precipitation in millimetres) per months in Europe. Spend some time exploring the data (you do not need to report these explorations, though). Compute the SVD of the data, and plot the first two left singular vectors to a map of Europe. This is also explained in the provided file. Can you interpret the results?

Play around with different color schemes and markers. Can you make the results more interpretable that way? Does the color scheme effect the interpretation?

Normalize the data to z-scores. Given the type of data we have, do you think this normalization is sensible?

Compute the SVD of the normalized data, and plot again the first two left singular vectors. Have they changed? Has your interpretation changed? Why?

Task 2: Selecting the rank

In this task, we use the normalized `worldclim` data that you did in the previous task. Compute the SVD of the data. Try the following rank selection methods to decide what would be a good rank for the truncated SVD:

- Guttman–Kaiser criterion
- 90% of explained Frobenius norm
- Scree test
- Entropy-based method
- Random flipping of signs

Report the rank each method suggests (and when subjective evaluation is needed to decide the rank, explain why you choose that rank). Discuss the results: are they same or different? Why? Do you think some methods work better than others? Why? What rank would you choose? Why?

max planck institut informatik

Task 3: Clustering and PCA

For this task, our goal is to cluster the rows of the data into five clusters and visualize the result. One way to do this is to first cluster the data into five clusters using the k-means algorithm, and then plot the data points into the map in such a way that the color of the marker identifies the cluster (the provided file explains the process).

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

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 the marker is defined by the clustering.

Are the clusters well-separated from each other in the plot or are they mixed? Do some of the clusters look like outliers?

Now, apply the Karhunen–Loève transformation (i.e. PCA) to project the data into a 2-dimensional subspace. Repeat the clustering and visualization steps with this new data. Did the results change? Why do you think the results changed or did not change?