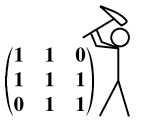

You must hand in a machine-typed report in PDF format and a script file (a `.R` *file). The report must explain your approach to the problem, the results you obtained, and your interpretation of the results. Naturally, the report must also answer to any direct question presented in the problems. You 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* `smetzler@mpi-inf.mpg.de`*. Remember to write your name and matriculation number to every answer sheet!*

Task 1: CX decomposition

Download the data and utility files from `http://resources.mpi-inf.mpg.de/departments/d5/teaching/ss17/dmm/assignments/assignment_3.zip`. That package contains file `assignment3.R`. You can fill your answers to that file and return it as a part of your solution.

In this task you implement the two-phase CX and apply it to the `worldclim` data we used in the first assignment.

First, you have to implement the two-phase CX with exact k algorithm (you can also see Boutsidis et al. 2008 or Boutsidis et al. 2010 for more information). For the RRQR algorithm, you can use the `qr` algorithm from R that computes the (pivoted) QR decomposition (see `assignment3.R`).

We apply the CX decomposition to the climate data we used in the first assignment. Load the data and normalize it to z-scores. The columns of the data are the climate variables, but instead we want to select some locations to $\boldsymbol{C}$. To that end, apply the CX decomposition to the transpose of the climate data. How would you describe the columns of $\boldsymbol{C}$? How about rows of $\boldsymbol{X}$?

The assignment file shows few ideas on how to visualize the decomposition. Based on these, do the results make sense? Can you interpret them? Try different values of k. How do the results change?

Try also to sample more columns (w.r.t. the final number of columns), and sample exactly k columns (in which case you don't have to apply the RRQR algorithm). Do the results change? What if you only select the top-k columns based on their probability (i.e. do not do any sampling or RRQR)? Are these variants worse than the proposed two-phase approach, the same, or better? Why?

Overall, is CX decomposition a good match for this data? Argue!

Task 2: Nonnegative CX decomposition

In this task, we study if nonnegative CX decomposition is a better match to the climate data. To that end, implement the `convex_cone` algorithm from the slides. You can you any method you prefer to find the nonnegative $\boldsymbol{x}$ and $\boldsymbol{X}$ in the algorithm, but you have to explain your choice.

You can use the normalized `worldclim` data from the above task. To remove the negative values, add to each variable it's minimum (so that for every variable, the smallest value is 0).

Try the `convex_cone` algorithm with different values of k and using the visualizations from the previous task. Can you interpret these results better than those from standard CX? Is the reconstruction error better or worse? Which method would you chose to analyse this data? Argue!

Task 3: ICA for housing prices

In this task, we study the applications of independent component analysis to housing price data from the US. The data set `us_housing_prices.csv`[1] contains the monthly house price index for twenty metropolitan areas in the US from January 1987 to June 2014. As is common to time series data, the rows correspond to the locations and the columns to the time stamps.

Get yourself familiar with the data by studying which locations it covers and by plotting the 20 time series. This data contains missing values (denoted `NA` in R), and we have to impute some values before we can proceed with the analysis. For this first round, we simply replace every missing value with 0; we will get back to this later.

Our algorithms expect the rows of the data to be the observations while the columns are the variables, and hence we will transpose the data. The index values vary between different locations, so before we proceed, we will normalize the data to z-scores. Is this normalization sensible? Argue!

We start by computing the full ICA of the scaled data (i.e. we find 20 independent components) using R's `fastICA` package. Compute the ICA and explain what do the different matrices in the solution mean. Show how to reconstruct the housing price index of Los Angeles, CA, from the ICA.

Let us now study the independent components. Plot them as time series. Can you interpret them? Remember that we do not know the sign of the components, so you might want multiply some components with -1 for plotting to have the peaks and well go the right way. For this analysis, it's useful to compare the locations of the peaks and wells with the original time series and/or to your general knowledge of US economy in the studied era.

We can also plot the scatter plot of the first 2 independent components. Can you see any outliers? If yes, identify the dates the outliers correspond to. Can you interpret them (probably using some general knowledge of US economy or by looking at the original data)? Is there any reason we should look for the top-2 independent components instead of, say, 7th and 8th? Try scatter plots of different pairs of independent components. Can you see other outliers and can you interpret them?

The US housing bubble made the housing prices climb heavily from around year 2000 until 2006, when they started to decline. By late 2008, the decline had turned into a crisis with significant drops in the housing prices; the houses would not start to fully recover until 2012. Can you identify these events in the independent components, either as such or from the scatter plots?

We now revise some of our earlier decisions, starting with the imputation of missing values. Earlier we replaced all missing values with 0s. Do you think this was a sensible decision given the data? One alternative would be to replace the missing values with the average of the data. Do you think that would be a sensible decision? Argue! The last option we are going to consider is to notice that all missing values happen at the begin of the time series, so we can replace the missing values with the first observed value for that location. Would this be sensible? Argue! Choose from the aforementioned techniques the one you think is the best and use it to replace the missing values in the original data. Transpose and scale the data again.

We now study the effects of whitening. Compute the SVD of the new transposed and scaled data. The columns of $\boldsymbol{U}$ now correspond to the uncorrelated (whitened) time series. Plot (some of) them. Do they show any patterns? Can you interpret them? Study the singular values. Are all the 20 uncorrelated signals necessary? Use some method to select the rank (e.g. Guttman–Kaiser or scree test) and report the results. Which method and rank you choose? Why?

Re-compute ICA, but compute only k independent components (using as k the rank you chose above). Study the results. Did they change? Can you interpret them? Report you findings and compare them with what you found in the previous step.

[1] Data source `http://data.okfn.org/data/core/house-prices-us`.