Exercise 9

Principle component analysis

Deadline: Please hand in your protocol in pdf format by Thursday, the 27th of July 2019, 10 am to jan.joswig@fu-berlin.de and marco.manni@fu-berlin.de.

The files needed for this exercise are provided at
`login.bcp.fu-berlin.de:/home/janjoswig/MD19/Ex09/`.

9.1 Protein C_α-trajectory

The given trajectory of 3110 frames holds coordinates of the 2018-atoms of holo-langerin. In Cartesian space the system is fully described by 6054 dimensions (xyz-coordinates for every atom). While you can visualise for the analysis all dimensions at once in a molecule viewer like VMD (trajectory.xtc), this can be highly subjective and hard to quantify. Principle component analysis (PCA) is a way to find new dimensions describing the system in a different space by a linear transformation of the original coordinates. The crucial point is, that these new dimensions can be ranked according to their importance, which allows to simplify the analysis of the system. Importance in terms of PCA means large covariance in the input data, in this case collective molecular motion with large spatial amplitude.

The steps to perform a PCA on the given data would be (the analysis is only done for the 128 C_α-atoms):

1. Load the trajectory (from the trajectory_calpha.xyz textfile) into a 2D array (*time* × *coordinate*).
2. Subtract the mean value from each input coordinate to obtain a mean-free trajectory (mean of 0):

$$x_i(t) = x_i'(t) - \mu\left(x_i'(t)\right) \tag{1}$$

3. Construct the covariance matrix:
(you can check your result with the numpy.cov() function)

$$C_{ij} = \frac{1}{t_n - 1} \sum_{t=0}^{t_n} x_i(t) x_j(t) \tag{2}$$

4. Calculate the eigenvalues and eigenvectors of this matrix:
(you can use numpy.linalg.eig)

$$C = \Upsilon \Lambda \Upsilon^{\mathrm{T}} \tag{3}$$

You will obtain as many eigenvectors/new dimensions/principle components (column vectors of $\varUpsilon$) as you had original input dimensions. The corresponding eigenvalues are a measure of the variance described along the eigenvectors. Plot the eigenvalue spectrum and and their cumulative sum.
5. Project your original trajectory $x(t)$ in Cartesian space onto the new principle component space by left hand side multiplication of the eigenvectors (now as row vectors) with the transposed original data matrix (coordinates as rows, time frames as columns):

$$y(t) = \varUpsilon^{\mathrm{T}} x(t)^{\mathrm{T}} \tag{4}$$

6. Visualise the subset of the projection onto the first two principle components (highest two eigenvalues) as 2D histogram. Can you identify maxima in the distribution? Transform the probability surface into a free energy landscape and plot it again.
7. Which molecular structure(s) (all atoms) correspond(s) to the highest maximum? Include a figure of this conformation in your report (*hint*: Find the trajectory frame indices contributing to the histogram bin with most counts. You can then visualise the corresponding frames directly in VMD or use `gmx trjconv -sub` to generate a subtrajectory).