Final Task 12
Principal Component Analysis on a Torus

Return: Hand in a PDF containing the discussion and analysis of **all** tasks (including the Gromacs simulation procedure) and a separate well-documented source code file (e.g. Jupyter notebook). If feasible, provide the `.xtc` as well.

Information: The tasks are only to guide you and therefore do not contain a complete list of work orders. Due to the randomness of the simulation, the findings will vary and not all suggested parameters will be optimal.

Introduction

Task I: Basic Data Consideration

In the following we consider the backbone angles of the alanine dipeptide (Ac-Ala-NHCH$_3$) to evaluate the differences between the different principal component analysis methods.

- The data given to you contains 2.5×10^6 points of a trajectory of alanine dipeptide, a peptide consisting of two amino acids.
- The first column contains the dihedral angle ϕ, the second column the dihedral angle ψ (see lecture for information on dihedral angles). The points of the trajectory are separated by 200 fs.
- Load the trajectory in your notebook and create a two-dimensional histogram/free energy, i.e., a Ramachandran plot.
- What do you see? How many states, i.e, point accumulations in the coordinate space, do you observe? How are the barriers between the states compared to each other? Why os a periodic treatment of the dihedrals necessary?
- For the last task we need the 1d dimensional free energy projection along ϕ and ψ. Generate them and discuss their distributions as well. What do you see?

Task II: Principal Component Analysis (PCA)

Now, we want to perform the PCA, just as you know it from sheet X. To this end, we calculate the averages $\langle\phi\rangle$ and $\langle\psi\rangle$. The covariance between ϕ and ψ is defined by

$$\mathrm{Cov}(\phi, \psi) = \frac{1}{N}\sum_{i=1}^{N} (\phi - \langle\phi\rangle)(\psi - \langle\psi\rangle) \tag{1}$$

where N is the number of steps of the simulated trajectory. Therewith, the variance can be defined as

$$\mathrm{Var}(\xi) = \mathrm{Cov}(\xi, \xi) = \frac{1}{N}\sum_{i=1}^{N} (\xi - \langle\xi\rangle)^2 \quad \text{with} \quad \xi = \phi, \psi \tag{2}$$

Calculate the covariance matrix

$$C = \begin{pmatrix} \mathrm{Var}(\phi) & \mathrm{Cov}(\phi, \psi) \\ \mathrm{Cov}(\phi, \psi) & \mathrm{Var}(\psi) \end{pmatrix} \tag{3}$$

Estimate the corresponding eigenvalues λ_i and eigenvectors $\boldsymbol{e}_i$ with $\lambda_1 > \lambda_2$. Afterwards, we project the trajectory onto the principal components x_1, x_2 (PCs) by

$$(\boldsymbol{e}_1, \boldsymbol{e}_2)^T \cdot (\phi(t), \psi(t))^T = (x_0(t), x_1(t))^T \tag{4}$$

where $(\boldsymbol{e}_1, \boldsymbol{e}_2)$ represents the matrix of normalised eigenvectors. With the projected data, plot again the two-dimensional free energy and compare it to the previous Ramachandran plot. What has changed?

As motivated in the introduction, the main usage of PCA is to reduce the number of dimensionality. Hence, visualize and discuss the one-dimensional projections on both principal components. Which dynamic would be neglected by projecting onto the first principal component x_1?

Task III: Dihedral Principal Component Analysis (dPCA)

As emphasized above, if we have periodic data it is not straight forward to perform a PCA. Even though it is feasible to implement a periodic metric (like you have implemented in the first few exercises) the mean value is not well-defined, because it normally measures the concentration of angles, and therefore neither the covariance nor the variance. This becomes obvious if one considers equally distributed angles.

To overcome this problem, dPCA was proposed. The idea is to project the periodic data onto a none periodic space. In this case, one projects each input coordinate onto sine and cosine, so

$$\begin{pmatrix} \phi \\ \psi \end{pmatrix} = \begin{pmatrix} \sin(\phi) \\ \cos(\phi) \\ \sin(\psi) \\ \cos(\psi) \end{pmatrix} \tag{5}$$

As it can be seen, this projects the periodic space onto a non-periodic one, $[-\pi, \pi) \times [-\pi, \pi) \mapsto \mathbb{R}^4$. After this projection, one performs a PCA treating the data points as elements of $\mathbb{R}^4$.

- Project both angles, as described above. Using the straight forward generalization of Task II to 4 dimensions, perform a PCA.
- Project as well only onto the first two principal components and visualize the two-dimensional free energy. How did it change compared to PCA and the Ramachandran-plot?
- Visualize the one-dimensional projections on the first and second principal components. Which dynamic would be neglected by projecting onto only the first principal component x_1? Does it differ compared to PCA? Discuss your findings. Does dPCA perform better or worse compared to PCA?

Task IV: Principal Component Analysis on a Torus (dPCA+)

As you have hopefully noticed, dPCA overcomes the problem of periodicity but it introduces new ones. For that reason, an alternative approach was introduced: Principal Component Analysis on a torus (called dPCA+ to emphasize that it is superior to dPCA). Instead of projecting the data, this time we try to minimize the error induced by ignoring the periodicity. Therefore, the input data is shifted such, that the maximal gap lays at the periodic boundary, followed by a non-periodic standard PCA. So, we are left with identifying the maximal gap. There are in general different approaches, either by maximizing the squared distance of the cut to the nearest data points, to minimize the data point density in a corridor (with finite width) or to use the dynamic information of the trajectory by choosing the cut such that we minimize the number of border crossing events.

For this task we will define the maximal gap by minimizing the point density of a corridor. We will try to automatize this step, by calculating the optimal shift. Later check if it agrees with the one-dimensional projections obtained in Task I.

- Generate a density (histogram) of your trajectory along both dihedral angles. Use a reasonable number of bins, to be able to still resolve the density and to have not to much noise at the border regions (between states).
- Once calculated, define the bin which is most far away from the next occupied bins. This bin will define the offset ξ_{offset} (where ξ is the angle projected onto)[1] Compare this finding with the visualized free energy. Does it agree? If not, try to tune your binning.
- Repeat the procedure for the second dihedral angles. Finally, shift periodic boundary condition such that

  $$(\phi, \psi)^T \mapsto (\phi + \phi_{\text{offset}}, \psi + \psi_{\text{offset}}) \tag{6}$$

  so only the periodic boundary should be shifted to

  $$[-\pi, \pi) \times [-\pi, \pi) \mapsto [-\pi + \phi_{\text{offset}}, \pi + \phi_{\text{offset}}) \times [-\pi + \psi_{\text{offset}}, \pi + \psi_{\text{offset}}) \tag{7}$$

- Afterwards perform a PCA as described in Task II.
- Plot again the two-dimensional free energy and compare it to the previous results (Task I-III). What has changed?
- Visualize and discuss as well the one-dimensional projections on both principal components.

Is their any difference compared to dPCA and PCA? If so, explain. Discuss the pros and cons of PCA, dPCA and dPCA+ when handling dihedral angles.

[1] 1t should be remarked that the offset does not correspond to the value of the bin. Instead, it is the difference between the periodic boundary an the identified region.