Prof. Dr. Petra Imhof
Prof. Dr. Felix Höfling
Irtaza Hassan (irtaza06@zedat.fu-berlin.de)

Freie Universität Berlin
FB Mathematik und Informatik

Computational Molecular Physics **Winter 2017/18**

Problem set 7 **14 December 2017**

Problem 7.1 *Data clustering*

The file *"brownian_data.txt"* contains trajectory from Brownian simulation. The first column is the time step, the second and third are x and y components.

a) Plot the data and cluster the xy plane into small boxes of size 3.0×3.0.

b) Compute the transition matrix by counting the transitions between the clusters.

c*) Choose lag-time (τ) such that Markov property maintains.

d*) Repeat (a)-(c) by clustering the data into the boxes of 1.5×1.5 and 2.0×2.0 and discuss the results.

Problem 7.2 *Principal compenent analysis (PCA)*

The file *"traj.xyz"* contains 1000 frames of a time series of a collection of seven atoms (Alanine-Leucine backbone atoms). Each frame is written in the format

number_of_atoms
title_line
atom_name x-coordinate y-coordinate z-coordinate

a) Calculate the mean value μ for each of the 3*7 coordinates.

b) Calculate the mean position for each of the seven atoms, expressed as x,y,z coordinates.

c) Set up the covariance matrix $\mathbf{C}$ of all the 21 coordinates (r=x,z, or y) with

$$C_{ij} = \langle (r_i - \mu_i)(r_j - \mu_j) \rangle$$

d) Calculate eigenvalues and eigenvectors of the covariance matrix.

e) Choose two eigenvectors to reduce the dimensionality of the system to. Project the trajectory, i.e. each frame of the time series, onto the chosen pincipal components with

$$\mathbf{PA} = \mathbf{B}$$

where the rows of $\mathbf{P}$ are the pincipal components, PC, $\mathbf{A}$ is a matix (21x1000) where each column corresponds to the full coordinates of a frame, and $\mathbf{B}$ is the matrix containing the projected trajectory.

f) Plot the data points of the projected two-dimensional trajectory as PC1 vs PC2

g) Plot the time series of the data points projected only onto PC1 and PC2, respectively (PC1 vs. time, and PC2 vs. time)

You can use e.g. the `PCA()` class from the matplotlib.mlab library (in python) or `princomp` from the statistics toolbox in matlab.

Due date: ***21 December, 12 p.m.***