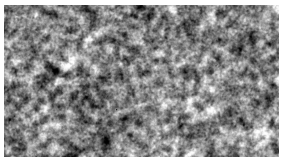
Data and supplementary information can be found online on a heiBox at
https://heibox.uni-heidelberg.de/d/8790b1f9fdc54e3089e5/
The corresponding password reads: *TDAworkshop*

Dataset 1 *Cosmic Microwave Background temperature maps*

Though very homogeneous, the Cosmic Microwave Background (CMB), i.e., the electromagnetic radiation filling all space that is a remnant from an early stage of the universe, shows tiny temperature fluctuations on the μK-level. They have been measured in detail for instance by the Planck satellite and are believed to be generated by quantum fluctuations of matter, which subsequently expanded immensely within an inflationary phase. In inflationary theories the universe itself expanded right after the Big Bang to the size of today's observable universe, starting at a size numerous orders of magnitude smaller.



Excerpt of CMB temperature fluctuations, © ESA and the Planck Collaboration, 2018 Results

A standard cosmological paradigm is provided by the ΛCDM model, which takes into account a cosmological constant (Λ) and cold dark matter (CDM). By today, the CMB is the most important observational probe for the ΛCDM model's validity. Namely, the ΛCDM model together with inflationary theories predict the CMB temperature fluctuations to be realizations of a homogeneous and isotropic Gaussian random field. While it has largely been agreed upon that the CMB exhibits characteristics of a homogeneous and isotropic Gaussian field, there are lingering doubts.

While full-sky CMB maps from Planck are available online at http://pla.esac.esa.int/pla/#home, we here focus on 10 excerpts of so-called mollview projections. The actual data is given in the healpix-format, tailored to images on a sphere and hence slightly more complicated to analyse. The data provided can be easily read into Python via

```
import pandas as pd
import numpy as np

# Note: the .dat data file needs to be in the folder from which Python is started
data = np.array(pd.read_csv("COM_CMB_IQU-commander_2048_R3.00_hm1_excerpt1.dat",
    sep="\t", engine='python'))
data -= np.mean(data)
```

The mean value of the image is to be substracted from the entire data set, for instance using the command provided in the last line of code above.

The goal of this CMB data analysis session is to compare the CMB map excerpts with different real-valued random fields via TDA methods. Random field models on the rectangle $[0, L_1] \times [0, L_2]$ include for instance the following,

(a) superposition of Gaussian random fields. Let $\{p_1, \dots, p_n\}$ be a chosen number of uniformly distributed points in the rectangle $[0, L_1] \times [0, L_2]$. Choose a variance $\sigma^2 > 0$ and sample amplitudes $\{a_1, \dots, a_n\} \subset \mathbb{R}^n$ according to a Gaussian distribution of arbitrary mean and variance. Define a random field realization $f : [0, L_1] \times [0, L_2] \to \mathbb{R}$ as

$$f(x, y) := \sum_{i=1}^{n} a_i e^{-|(x,y)-p_i|^2/2\sigma^2}.$$

(b) Fisher random field. Let $m, n > 0$ be integers and $\eta_i : [0, L_1] \times [0, L_2] \to \mathbb{R}$, $i = 1, \dots, m+n$, be a zero-mean and unit-variance Gaussian random field. The Fisher random field $F_{m,n}$:

$[0, L_1] \times [0, L_2] \to \mathbb{R}$ is defined as

$$F_{m,n}(x, y) := \frac{(\eta_1^2(x, y) + \cdots + \eta_m^2(x, y))/m}{(\eta_{m+1}^2(x, y) + \cdots + \eta_{m+n}^2(x, y))/n}.$$

(c) Student random field. Let $n > 0$ be an integer, $\eta_i : [0, L_1] \times [0, L_2] \to \mathbb{R}$, $i = 1, \ldots, n$, be a zero-mean and unit-variance Gaussian random field. The Student random field $T_n : [0, L_1] \times [0, L_2] \to \mathbb{R}$ is defined as

$$T_n(x, y) := \frac{\eta_1(x, y)}{\sqrt{(\eta_2^2(x, y) + \cdots + \eta_n^2(x, y))/n}}.$$

A script called "Gaussian_random_field.py" is provided which generates realizations of Gaussian random fields. The script can be easily adjusted to generate Fisher and Student random fields as well as others which are potentially of interest.

In general, a statistical analysis of topological structures might be favorable. As a first starting point for a persistent homology investigation, you could construct sublevel sets of the CMB maps and corresponding random field model realizations and study the homology of the corresponding grid points — using the maps themselves or possibly other scalars constructed from them. In this approach, one needs to bear in mind that the 'real' fields are continuous. How could one take care of it in persistent homology?

This explorative data example is inspired by Pranav *et al.*, A&A 627, 2019, https://doi.org/10.1051/0004-6361/201834916.

Dataset 2 *Mapper on scRNAseq data*
Provided by Sebastian Damrich, IWR Heidelberg.

Single-cell RNA sequencing data (scRNAseq) consists of measurements for the expression levels of multiple genes for individual cells. The resulting data is a matrix, where each row corresponds to a single cell and each column to the expression levels of a particular gene. There are typically measurements for a few thousand cells and genes. Biologists hope to identify structure based on these gene expressions. One such structure are developmental trajectories, i.e., the pathways along which generic cells such as stem cells differentiate to more specific cells such as liver cells, neurons or skin tissue. To do so, a sparse, often tree-shaped structure has to be extracted from the high-dimensional gene expression measurements.

In this project we would like you to try to find a graph representation using the Mapper algorithm. This is similar in spirit to Exercise 4 (Mapper on breast cancer data) from Monday. We provide preprocessed gene expression measurements for two datasets ("dentate_gyrus.h5" and "endocrine_pancreas.h5") which we expect to contain developmental trajectories. Your task is to try to explore them. The datasets contain the preprocessed data ("data"), label information based on some clustering result ("labels"), a 50-dimensional PCA dimension reduction ("pca") and a 2-dimensional UMAP dimension reduction of the PCA ("umap") to plot and look at the data.

To get started execute in your terminal the following:

(a) `pip install h5py matplotlib`

(b) Download the folder "2 - scRNAseq" from the above heiBox

(c) Move to the respective folder via `cd`

(d) `ls` [should show the two .h5 files among other files]

(e) `python read_scRNA_data.py` [shows one dataset and plots dimensions of the contained arrays]

(f) Change file_name in read_scRNA_data.py

(g) `python read_scRNA_data.py` [shows the other dataset and plots dimensions of the contained arrays]

This should give you some idea of what the provided data looks like.

To apply Mapper, you need to select a filter function. This is very much an open problem, but the choices for the breast cancer data from Exercise 4 on Monday might give you some ideas. You can apply the filter either directly on all the genes. Alternatively, you can also try on the PCA or even the UMAP dimension reduction. Play around with different filter functions and parameters of Mapper and try to find a graph representation, that looks "tree-like" (although some loops might be also plausible if there are different pathways to the same cell type). You can sanity-check the graphs you found by extracting from Mapper the points that form a node and place a centroid of the corresponding 2D UMAP embeddings in the 2D plot. Link up these centroids as given by the Mapper graph structure. This way you overlay the 2D visualisation of the dataset with the Mapper graph. Ideally, the graph has little self-intersection and provides a visuably plausible skeletonisation of the data.

Dataset 3 *EEG recordings of a human adult performing a visuospatial working memory task*

Provided by the group of Prof. Daniel Durstewitz, ZI Mannheim, and based on Johnson *et al.*, Front. Syst. Neurosci. 12(65), 2019, https://doi.org/10.3389/fnsys.2018.00065.

Note that the data to handle in this explorative data example is quite large with many additional parameters included.

How does the human brain rapidly process incoming information in working memory? In growing divergence from a single-region focus on the prefrontal cortex (PFC), recent work argues for emphasis on how distributed neural networks are rapidly coordinated in support of this central neurocognitive function. Previously, the authors of the above mentioned study showed that working memory for everyday "what," "where," and "when" associations depends on multiplexed oscillatory systems, in which signals of different frequencies simultaneously link the PFC to parieto-occipital and medial temporal regions, pointing to a complex web of subsecond, bidirectional interactions. Here, the authors used direct brain recordings to delineate the frontoparietal oscillatory correlates of working memory with high spatiotemporal precision.

This data set contains EEG recordings from the lateral frontal and parietal regions from 7 human adults while they were performing a visuospatial working memory task. The recordings were intracranial (iEEG) recording subdurally or stereotactically with 4, 5 or 10 mm spacing between channels, sampled at 1 kHz or 512 Hz. Subjects were 7 epileptic patients (mean ± SD [range]: 25.4 ± 3.3 years of age, 4 males). Primary (filtered) and derived (fully preprocessed) iEEG data, and analysis scripts included.

The recordings from each participant were made during a visuospatial working memory task. Each trial consisted of five phases: pretrial, encoding, pre-cue delay, post-cue delay, and response. Following a 1-s pretrial fixation interval, participants were cued to focus on either IDENTITY or RELATION information. Then, two common-shape stimuli were presented for 200 ms each on the in the center of the screen in a specific spatiotemporal configuration (i.e., top/bottom spatial and first/second temporal positions). After a 900- or 1150-ms jittered pre-cue delay fixation interval, the test prompt appeared, followed by a post-cue delay fixation interval of the same length. Working memory was tested in a two-alternative forced choice test, resulting in a 0.5 chance rate. In the identity test, subjects indicated whether the pair was the SAME pair

they just studied (50% yes/no). In the spatiotemporal relation test, subjects indicated which shape fit the TOP/BOTTOM spatial or FIRST/SECOND temporal relation prompt (50% per stimulus). Participants completed 120 trials.

The file "crcns_fcx-3_data_description.pdf" contains further information on the different data files contained in the folder for this exercise. To load the EEG time-series data into Python, execute the following commands

```
import numpy as np

data = np.load("eeg_data.npy", allow_pickle=True)
```

In particular, the spatial coordinates of the EEG sensors around the participant's head are given in "s3_MNI_grid.csv". More information regarding the MNI coordinate system can be found here, if of relevance: https://imaging.mrc-cbu.cam.ac.uk/imaging/MniTalairach An example is given by https://neurosynth.org/locations/.

Scientifically, a single trial should already suffice to obtain meaningful deductions. This strongly reduces the size of the data to analyze.

A specific exercise task for this data set is not existing. The goal is the very meaning of *exploring* the data set. Do you find topological structure somewhere? Might Mapper be a promising method to apply to this data?

Inspiration for persistent homology explorations may be given by Wang, Ombao, Chung, Ann. Appl. Stat. 12(3), 2018, https://doi.org/10.1214/17-AOAS1119

Inspiration for explorations using the Mapper algorithm might be given by Saggar *et al.*, Nature comm. 9, 2018, https://doi.org/10.1038/s41467-018-03664-4

Dataset 4 *Action and outcome activity state patterns of the anterior cingulate cortex of rats*

Provided by the group of Prof. Daniel Durstewitz, ZI Mannheim, and based on Hyman *et al.*, Cerebral Cortex 23(6), 2013, https://doi.org/10.1093/cercor/bhs104.

Although there are numerous theories regarding anterior cingulate cortex (ACC) function, most suggest that it is involved in some form of action or outcome processing.

The data in the study characterized the dominant patterns of ACC activity on a task in which actions and outcomes could vary independently. Patterns of activity were detected using a modified form of principal component analysis (PCA), termed constrained PCA in which a regression procedure was applied prior to PCA to eliminate the contribution of nontaskrelated activity.

Histology and tetrode-lowering records confirmed that all neurons were recorded from dorsal regions of the mPFC (either ACC or dorsal prelimbic). Since all recordings were made using moveable tetrodes that were advanced over multiple sessions, it was not possible to precisely identify each neuron's recording location.

To load the time-series data into Python, execute the following commands

```
import numpy as np

data = np.load("james_data.npy", allow_pickle=True)
```

The data contains multiple single unit recordings from rat brain. In total, 22 trials, each trial with a different length (between 75 and 341 time steps), 10 neurons recorded, are available. Everything is saved as a numpy array of trials of shape (22,) with each entry being a numpy array of e.g. (341, 10) (time steps, dimensions).

A specific task for this data set is not existing. The goal is the very meaning of *exploring* the data set. Do you find topological structure somewhere? Might Mapper be a promising method to apply to this data?

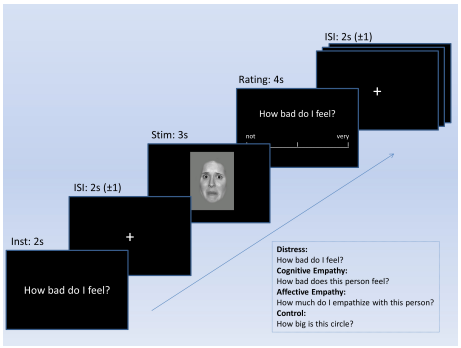
A source of inspiration for the use of persistent homology in this example might be Petri *et al.*, J. R. Soc. Interface 11, 2014, https://doi.org/10.1098/rsif.2014.0873.

Dataset 5 *EEG measurement data of different empathy tasks*
Provided by Dr. Stephanie Schmidt, Clinical Psychology, U Konstanz.

The data consists of task-related recordings of 20 participants. The group recorded 128 electrodes of EEG time series.

The task is an empathy task, in which pictures of emotional faces were presented, and participants had to think about different aspects while looking at the faces.

The data was recorded at 5000 Hz and downsampled to 256 Hz, subsequently; the task lasts for about 15 minutes. In the preprocessing process of the data a notch filter at 50 Hz, a highpass filter at 0.01 Hz and a low pass filter at 80 Hz have been applied. The picture on the right displays the rough task design of a single (extensive) trial. There are 4 trials per condition. The conditions are repeated 5 times in a fixed order, summing up to 20 trials in total per condition.



Task design, © Dr. Stephanie Schmidt

The data can be loaded into Python e.g. via the package `pandas`. You can install it by executing the command `pip install pandas` in your terminal. Then, you can load the two data files "WIN1_102_empathie_Segmentation affective emp.dat" and "WIN1_102_empathie_Segmentation distress.dat" using the following piece of code, for instance,

```
import numpy as np
import pandas as pd

data = np.array(pd.read_csv("WIN1_102_empathie_Segmentation_affective_emp.dat", sep=" ",
    header=None, engine='python'))[:,3:15363]
```

In the data file loaded this way, each row corresponds to an individual EEG channel, i.e., an individual electrode located on the head of the participant. The columns then specify a number of concatenated time series, each consisting of 3 seconds of recording at a sampling frequency of 256 Hz. Each of the concatenated 20 time series correspond to a question that the participant had to think about, i.e., correspond to a single trial.

As well as with the other data sets, the goal of the analysis of this EEG data is the very meaning of being *explorative*. Collect and exchange ideas, try to capture the scientific essence behind the given time series. Do you expect particular topological structures somewhere? Might Mapper be a promising method to apply to this data?

A source of inspiration for the use of persistent homology in this example might be Petri *et al.*, J. R. Soc. Interface 11, 2014, https://doi.org/10.1098/rsif.2014.0873.

Inspiration for explorations using the Mapper algorithm might be given by Saggar *et al.*, Nature comm. 9, 2018, https://doi.org/10.1038/s41467-018-03664-4