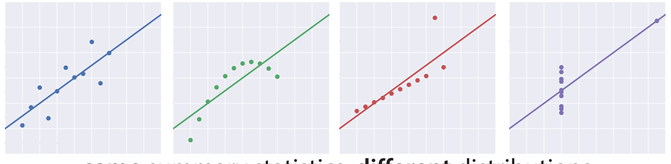
HyperTools: A python toolbox for gaining geometric insights into high-dimensional data

Andrew C. Heusser, Kirsten Ziman, Lucy L.W. Owen and Jeremy R. Manning
Dartmouth College

Overview

Visualizing data can reveal trends and patterns that are not otherwise obvious from the raw data or summary statistics.

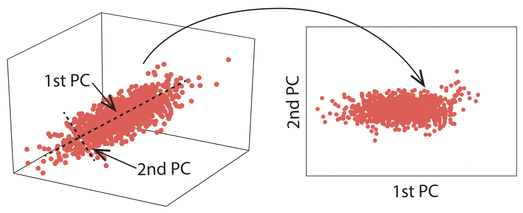
Anscombe's quartet



same summary statistics, **different** distributions

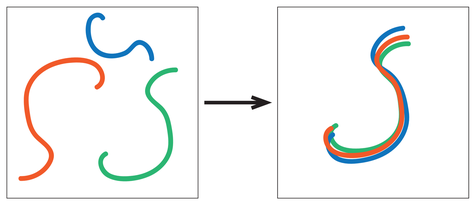
HyperTools is designed to facilitate dimensionality reduction-based visual explorations of high-dimensional data, allowing one to perform complex analyses often in a *single line of code*.

Dimensionality reduction



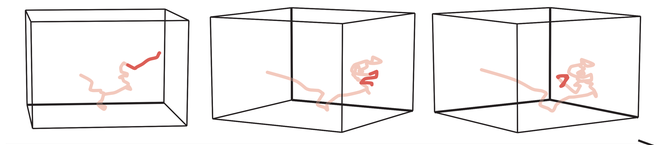
`hyp.plot(data, 'o')` `hyp.plot(data, 'o', ndims=2)`

Hyperalignment



`hyp.tools.align([array1, array2])`

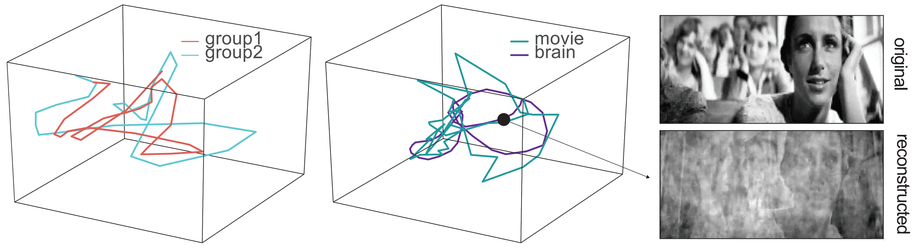
High-dimensional timeseries data can be visualized by reducing it to its first 3 principle components and plotting the resulting coordinates as a trajectory.



`hyp.plot(data, animate=True)`

Visualizing high dimensional neural data

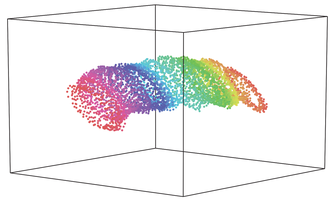
Aligning brains and movies to a common space



`hyp.plot([group1, group2])` `hyp.plot([movie, brain])`

(Left) Group-averaged hyperaligned activity (BOLD) from ventral visual cortex for two groups of subjects (each n=6) watching the same movie. (Middle) Ventral visual activity hyperaligned to the movie trajectory (changes in pixel intensity over time). (Right) Reconstructed movie frame from "movie-aligned" brain data.

Cyclic EEG patterns over time

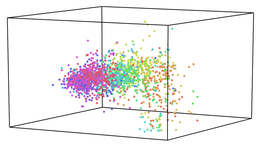


`hyp.plot(data, '.', group=time, azim=220)`

EEG (8 channels, n=1) recorded during rest reduced with principal components analysis. Each timepoint is plotted as a dot and colored by time. A cyclic pattern is present in 2 of the axes and the other is likely explained by signal drift.

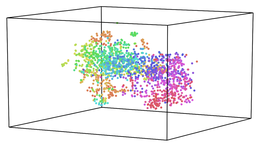
Spatial coding in hippocampal population responses

Principal Components Analysis



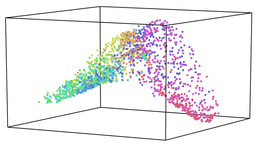
`hyp.plot(data, reduce='PCA')`

t-SNE



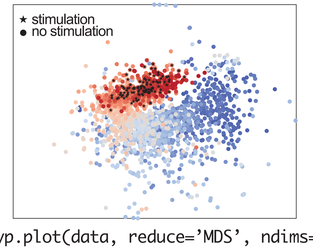
`hyp.plot(data, reduce='TSNE')`

Spectral Embedding



`hyp.plot(data, reduce='SpectralEmbedding')`

Multidimensional Scaling

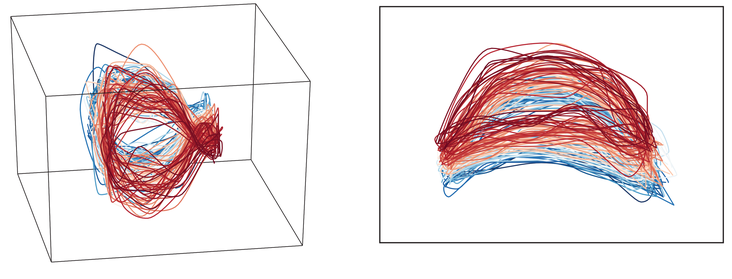


`hyp.plot(data, reduce='MDS', ndims=2)`

Single unit spikes (42 unique cells) were recorded from mouse hippocampus (dentate gyrus) during during locomotion on a linear track. The 3D plots above represent average population response patterns colored by location using different dimensionality reduction methods. Optogenetic stimulation was triggered for a subset of the cells at a particular spatial location. In the 2D plot to the left, stimulated locations are highlighted with a start and all locations are colored by track position. (Thank you Sam McKenzie for sharing this data!)

Visualizing other types of data

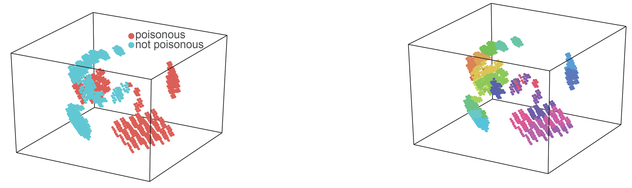
Cyclic global weather patterns over time



`hyp.plot(temps, group=year)` `hyp.plot(temps, group=year, ndims=2)`

Global temperatures are cyclic due to seasonal changes (left), but there is also a gradual increase in temperature over the past 130 years (right).

Feature vectors of various mushroom types



`hyp.plot(mushrooms, '.', group=labels)` `hyp.plot(mushrooms, '.', n_clusters=23)`

Descriptive features (i.e. cap size, color, odor) from a dataset of mushrooms. Mushrooms' poisonousness appears stable within each cluster (e.g. mushrooms that have similar features), but varies across clusters (left). Using the built-in k-means clustering algorithm, subclusters can be labeled and easily visualized (right).

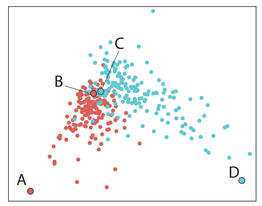
Topic models fit to political tweets

A
Why doesn't the failing @nytimes write the real story on the Clintons and women? The media is TOTALLY dishonest!

B
"@justiningly: @realDonaldTrump great speech today!! It's all about America and that's why you will become president!!!!" Thank you.

C
The labor movement pioneered the basic work that made America great: If you work hard and do your part you should be able to get ahead.

D
Obama on whether Trump could be trusted with US nuclear weapons: "Make your own judgement" https://t.co/6OZtrflwim https://t.co/Nj20PaXF2o



A
trump https donald vote today women president says millions believe https thank amp makeamericagreatagain crowd rights fight votetrump making realdonaldtrump clinton hillary https trump tax did woman rncincle failing

B
realdonaldtrump clinton hillary https trump tax did woman rncincle failing people americans https said speech enjoy american years history hillary trump https donald vote today women president says millions believe

C
trump https donald vote today women president says millions believe great america trump2016 https media plan carolina isis ratings tickets job work jobs better look wow man world life hard

D
trump https donald vote today women president says millions believe new https going say hillaryclinton york hillary timkaine people run president obama bad hillary states totally united person senator crooked

`hyp.plot(tweets, '.', group=labels, ndims=2)`

We fit topic models to tweets from Donald Trump and Hilary Clinton during the 2016 election (middle). On the right are the top 10 words for the 3 topics that best fit some exemplary tweets. Example tweets are shown on the left.

For questions, comments, or reprints, please visit our website: www.context-lab.com. This work was supported in part by NSF EPSCoR Award Number 1632738.