DROPP: STRUCTURE-AWARE PCA FOR ORDERED DATA

A GENERAL METHOD AND ITS APPLICATIONS IN CLIMATE RESEARCH AND MOLECULAR DYNAMICS

Anna Beer[1,2], Olivér Palotás[2,3], Andrea Maldonado[3,4], Andrew Draganov[2], Ira Assent[2]

[1]University of Vienna, [2]Aarhus University, [3]LMU Munich, [4]MCML

Talk at ICDE 2024

40th IEEE International Conference on Data Engineering
Utrecht, Netherlands | 13th - 17th May
2024

universität wien

AARHUS UNIVERSITY

LMU LUDWIG MAXIMILIANS-UNIVERSITÄT MÜNCHEN

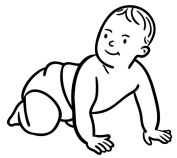
mcml
Munich Center for Machine Learning

MOTIVATION

- There is ordered data
- Dimensionality reduction is important
  - We want to visualize data that comes from similar origin in a comparable way
  - Downstream tasks should get meaningful input

→ We use the order to improve dimensionality reduction

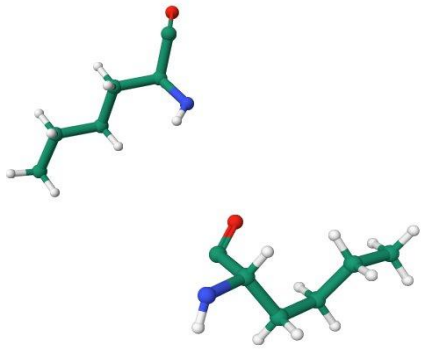
SOME DATA HAS AN INHERENT ORDER

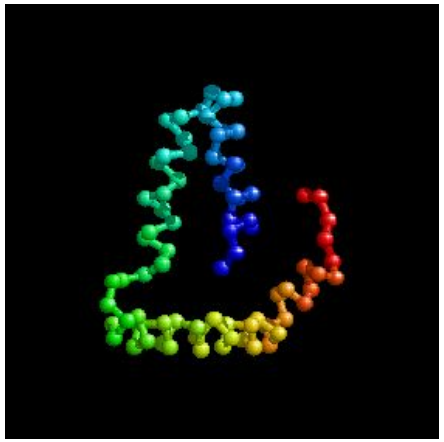


- Order: strict total order $\cdot < \cdot$
- Time series vs. ordered data

PROTEINS ARE ORDERED

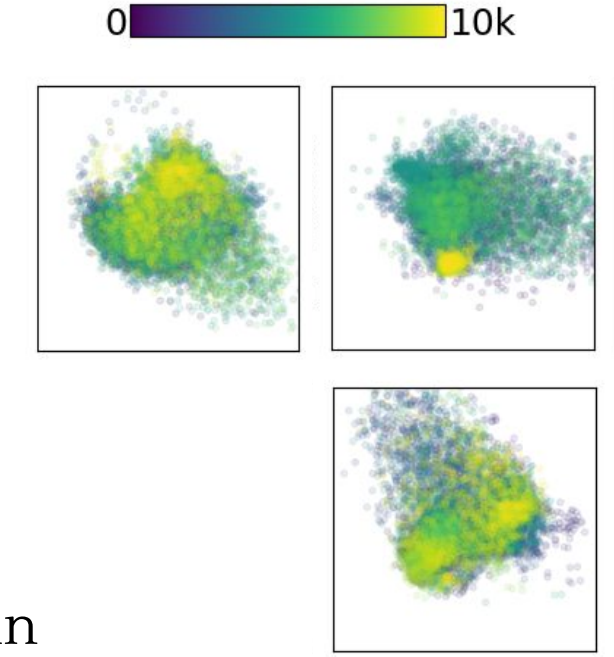
- Proteins are cool and important, e.g., for diseases or drug design
- Proteins are chains of atoms, where the shape is important for the function
- The backbone of a protein consists of C^{α} atoms that are connected by peptide bonds of roughly fixed length





http://www0.cs.ucl.ac.uk/staff/d.jones/t42morph.html 16-05-2024

MOLECULAR DYNAMICS

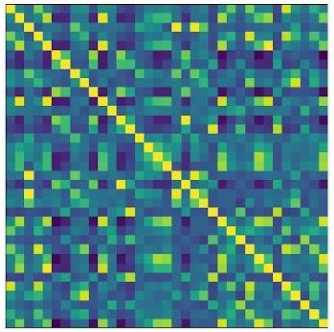


- MD data is **very large** , containing trajectories of all atoms in a protein
- **Tensor** data with several thousand time steps with 3d coordinates for several hundred atoms
- **Relative motion** of atoms within the molecule is important, independently of position and rotation of the complete molecule in space
  - PCA is used for tensor data
- Visualization for analysis, e.g., regarding protein folding
  - Order of atoms is disregarded
  - Visualizations are not comparable
  - Axes are not interpretable

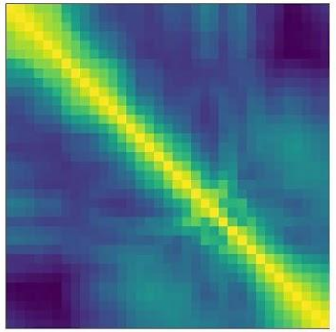
MAIN IDEA

- Incorporating the inherent order of the data exhibits structure in the covariance matrix that we can exploit

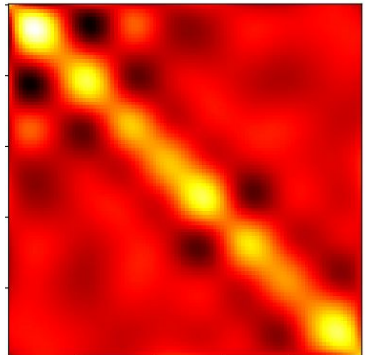
Covariance Matrix C



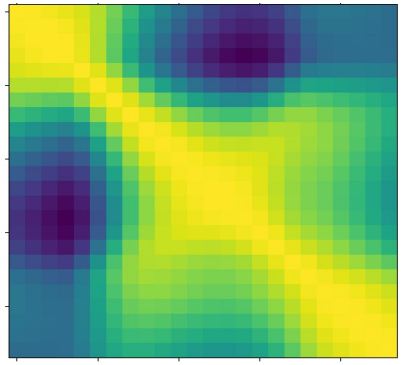
Ordered Covariance Matrices



MD trajectory

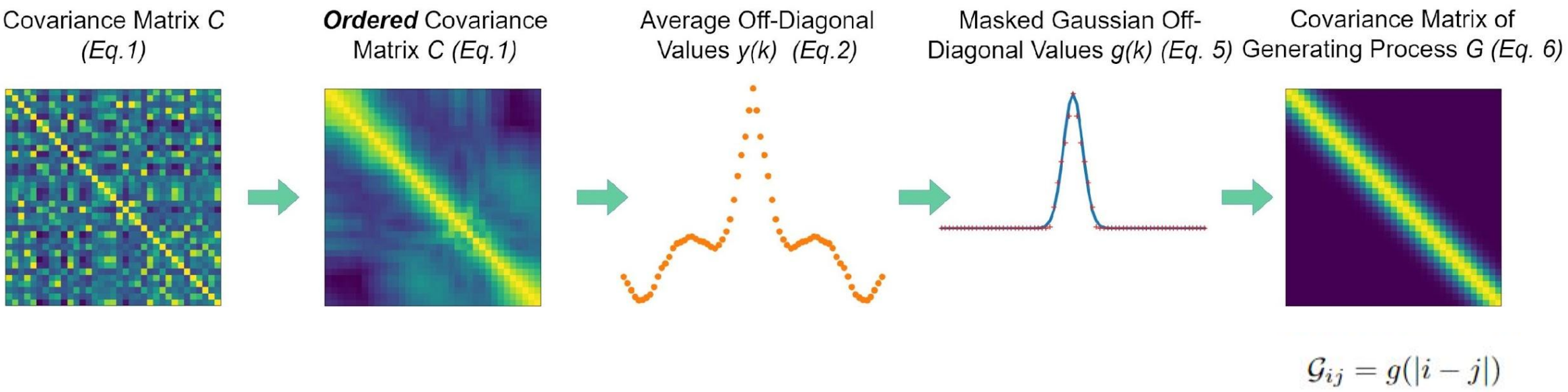


Synthetic random walk data



Climate time series

CONTRIBUTION: DROPP



ADVANTAGES

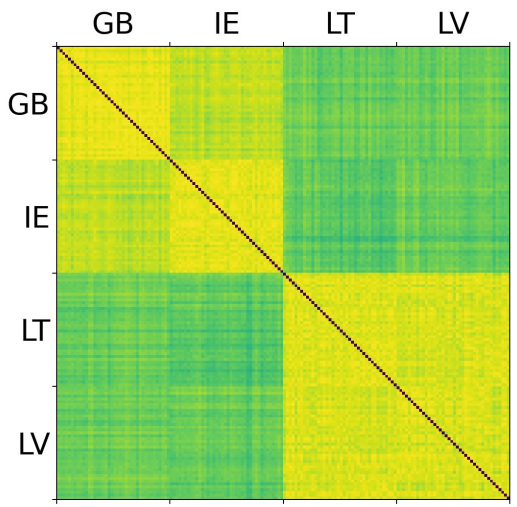
- Compute PCs once, use for all other manifestations coming from the same underlying process
- Compute PCs for small time frame, use for long time series
- Comparable visualization of different manifestations
- Incorporating the order gives similar advantages as knowing the ground truth

EXPERIMENTS

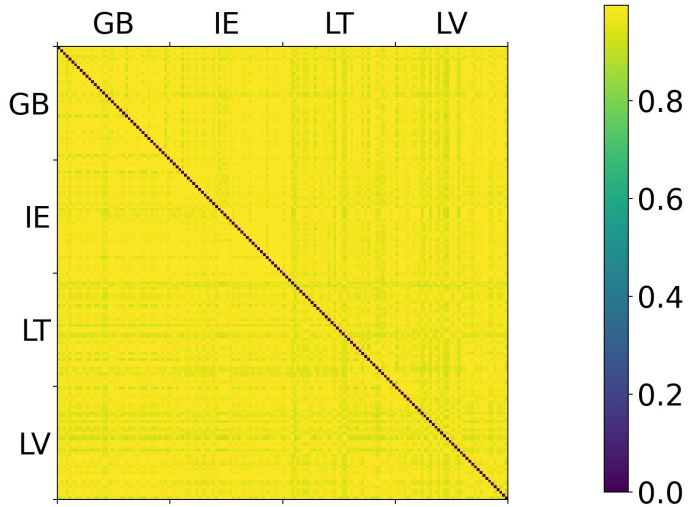
- Similarity of principal components
- Reconstruction error on different manifestations

EXPERIMENTS ON CLIMATE DATA

Similarity of first ten PCs for temperature values over 40 years in Great Britain, Ireland, Lithuania and Latvia



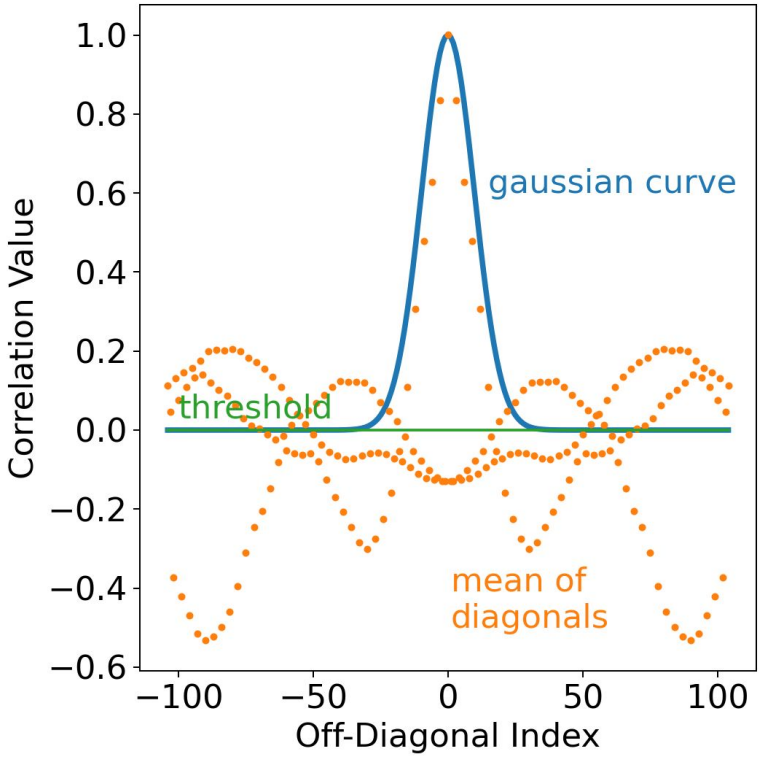
PCA



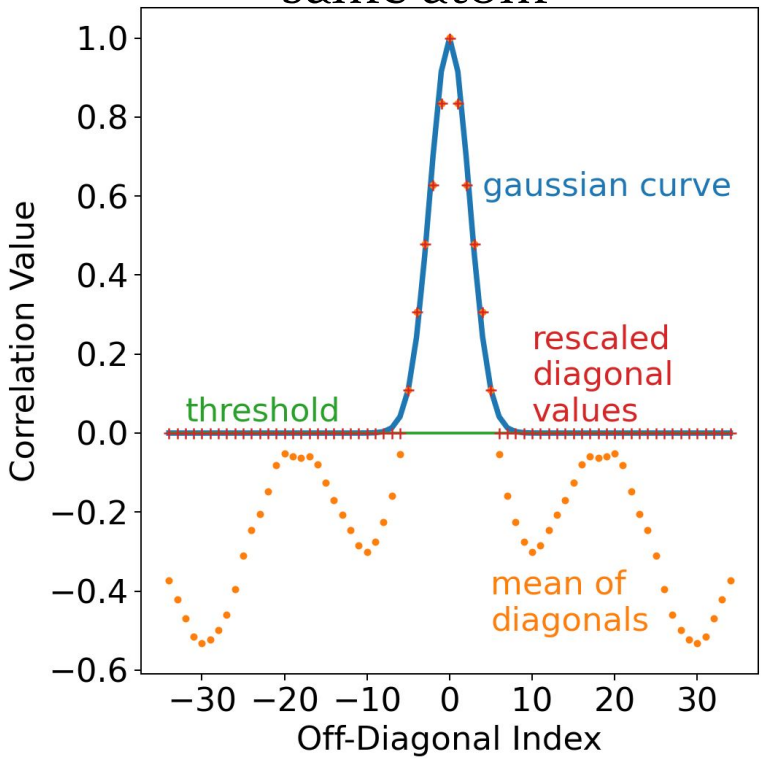
DROPP

EXPERIMENTS ON MD DATA

MD trajectories are tensors

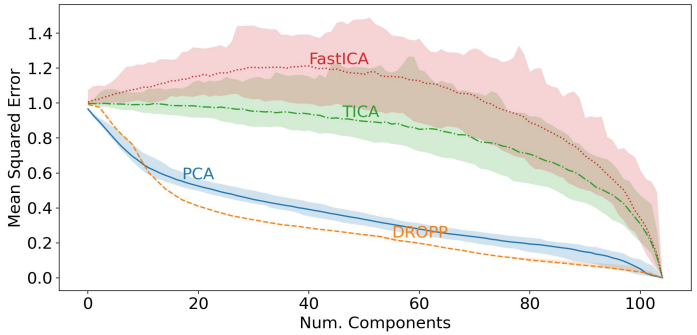


Combine coordinates belonging to the same atom

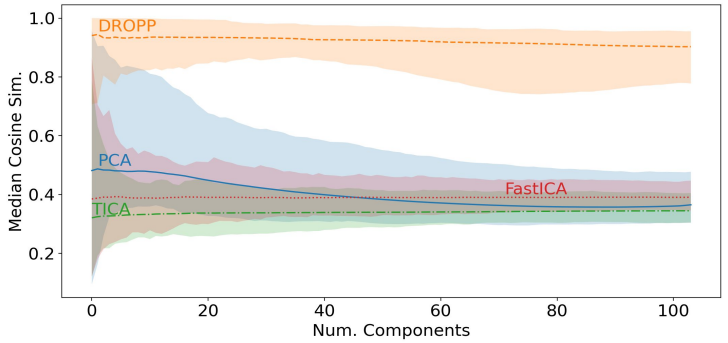


EXPERIMENTS ON MD DATA

Reconstruction error on different instantiations

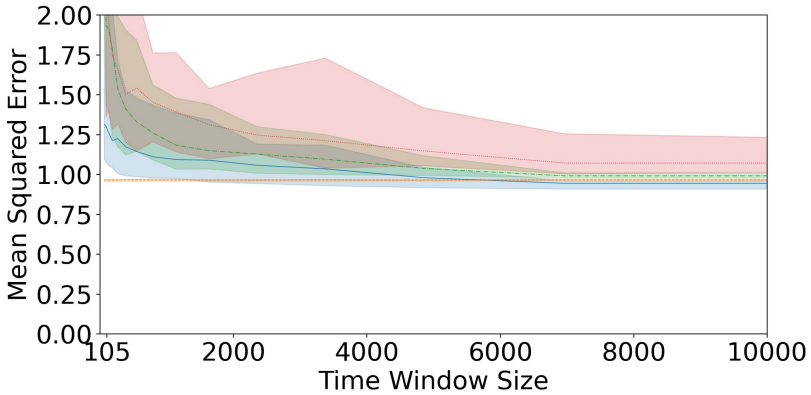


Similarity of components

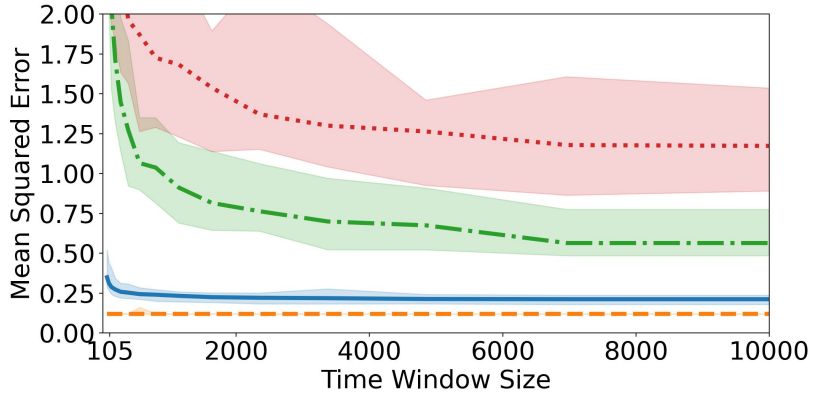


Reconstruction error for varying time window size

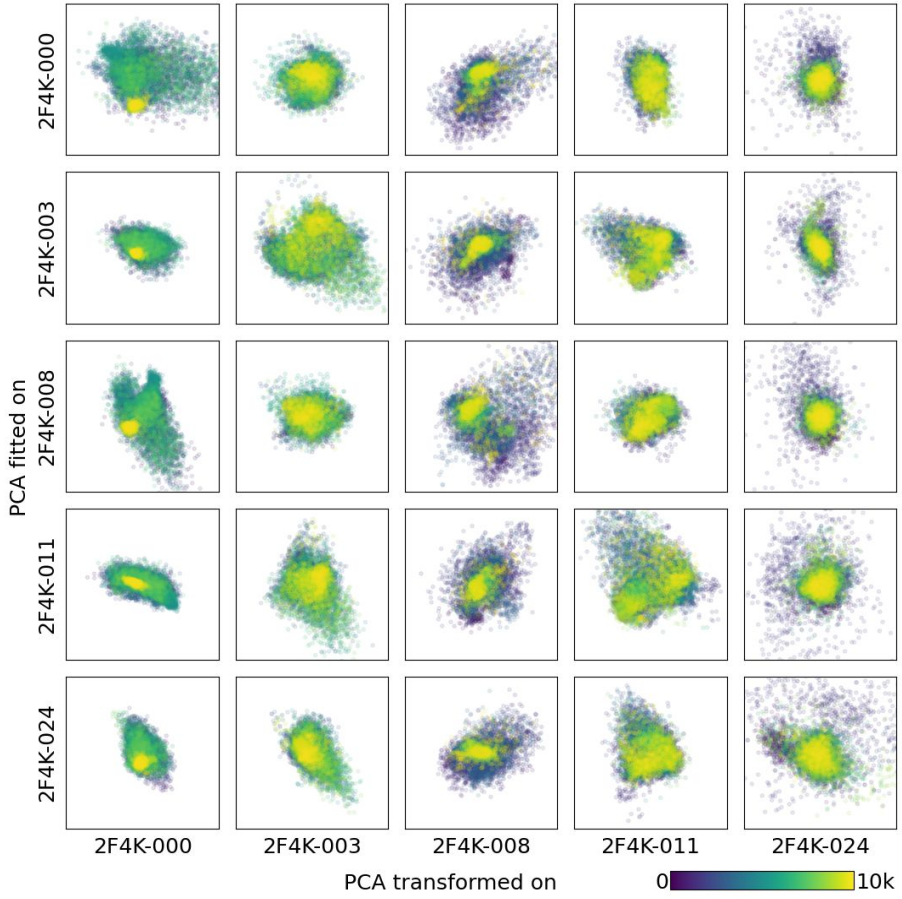
2 PCs

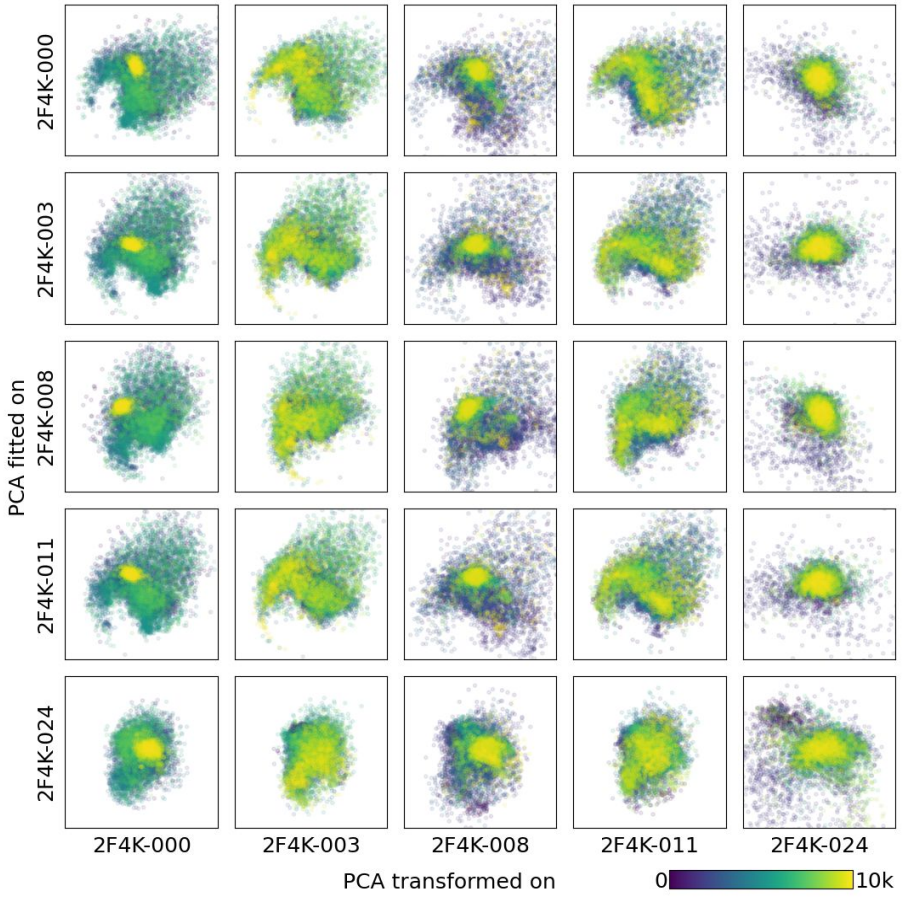


50 PCs



EXPERIMENTS ON MD DATA



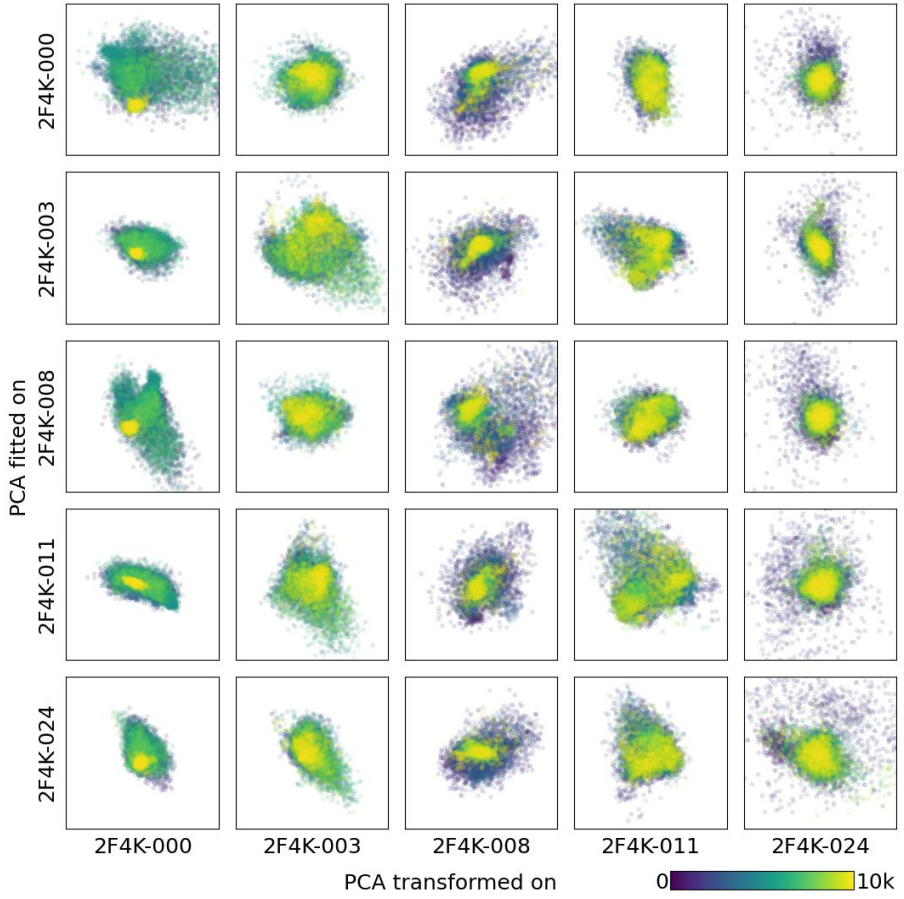


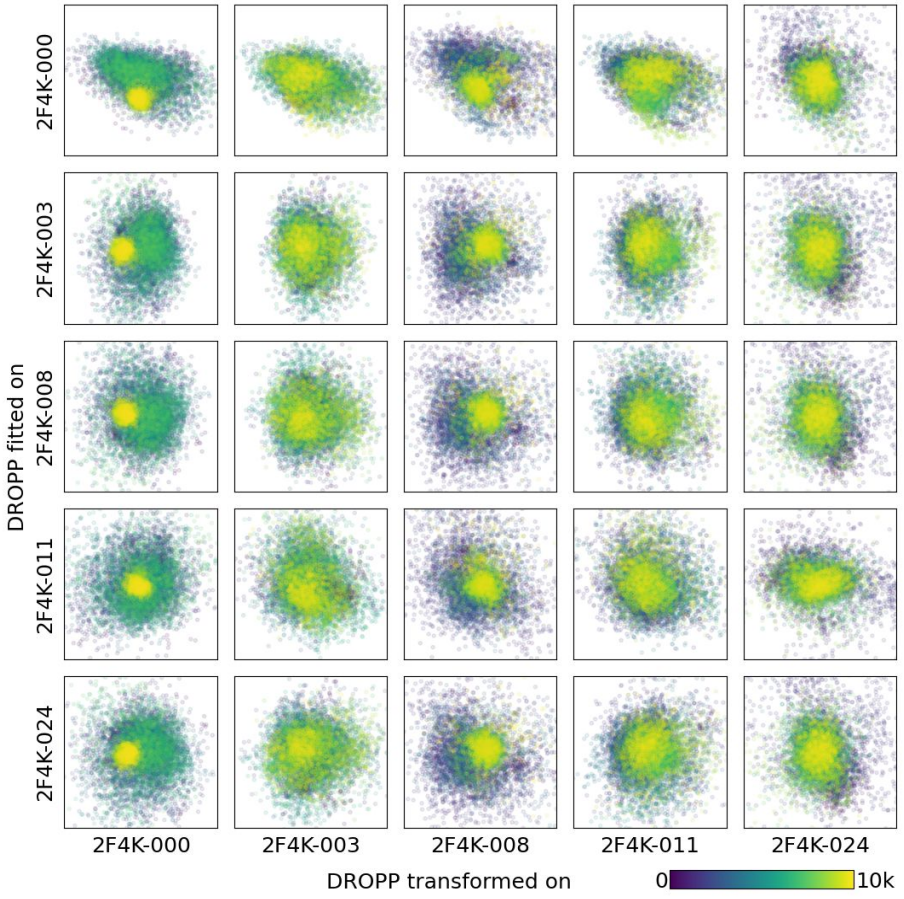
For meaningful comparability between different trajectories, rows should look similar.

2d projections using different trajectories' principal components

Aligned using ground truth

EXPERIMENTS ON MD DATA



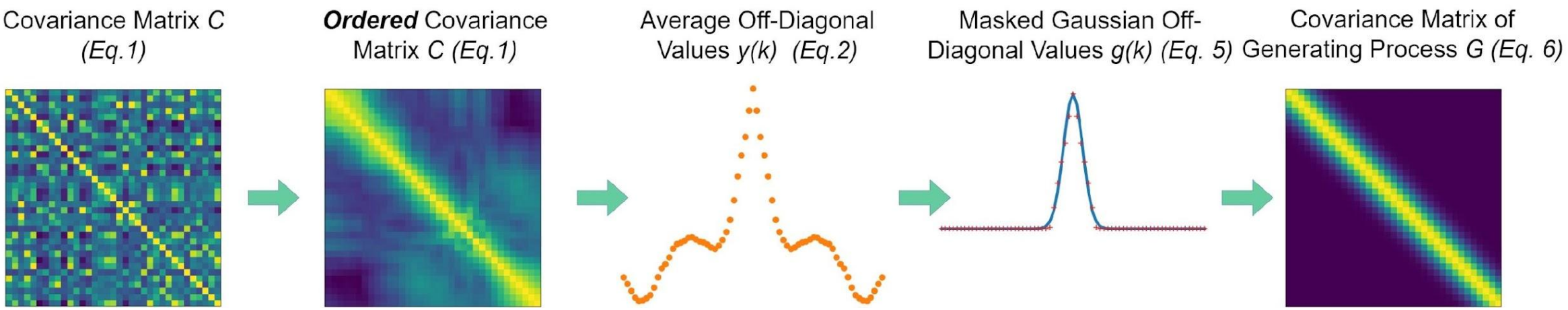


For meaningful comparability between different trajectories, rows should look similar.

2d projections using different trajectories' principal components

Using DROPP

SUMMARY



Questions?

FURTHER STUFF – MATHS

Method description

$$\mathcal{C} = Cov(\mathcal{X}) = \mathcal{X}^\top \cdot \mathcal{X}$$

$$y(k) = mean(c_{ij} \in \mathcal{C}), \text{ where } i - j = k$$

$$k_0 = min\{|k| : y(k) \leq \min(0, min_k(y(k)))\}$$

$$y_G(k) = \begin{cases} y(k) & |k| < k_0 \\ 0 & k_0 \leq |k| < d \end{cases}$$

$$RMSE\left(g(k), y_G(k)\right) = \left(\frac{1}{d}\sum_{k=0}^{d-1}(g(k) - y_G(k))^2\right)^{\frac{1}{2}}$$

$$g(k) = e^{-\left(\frac{k}{2\sigma}\right)^2}$$

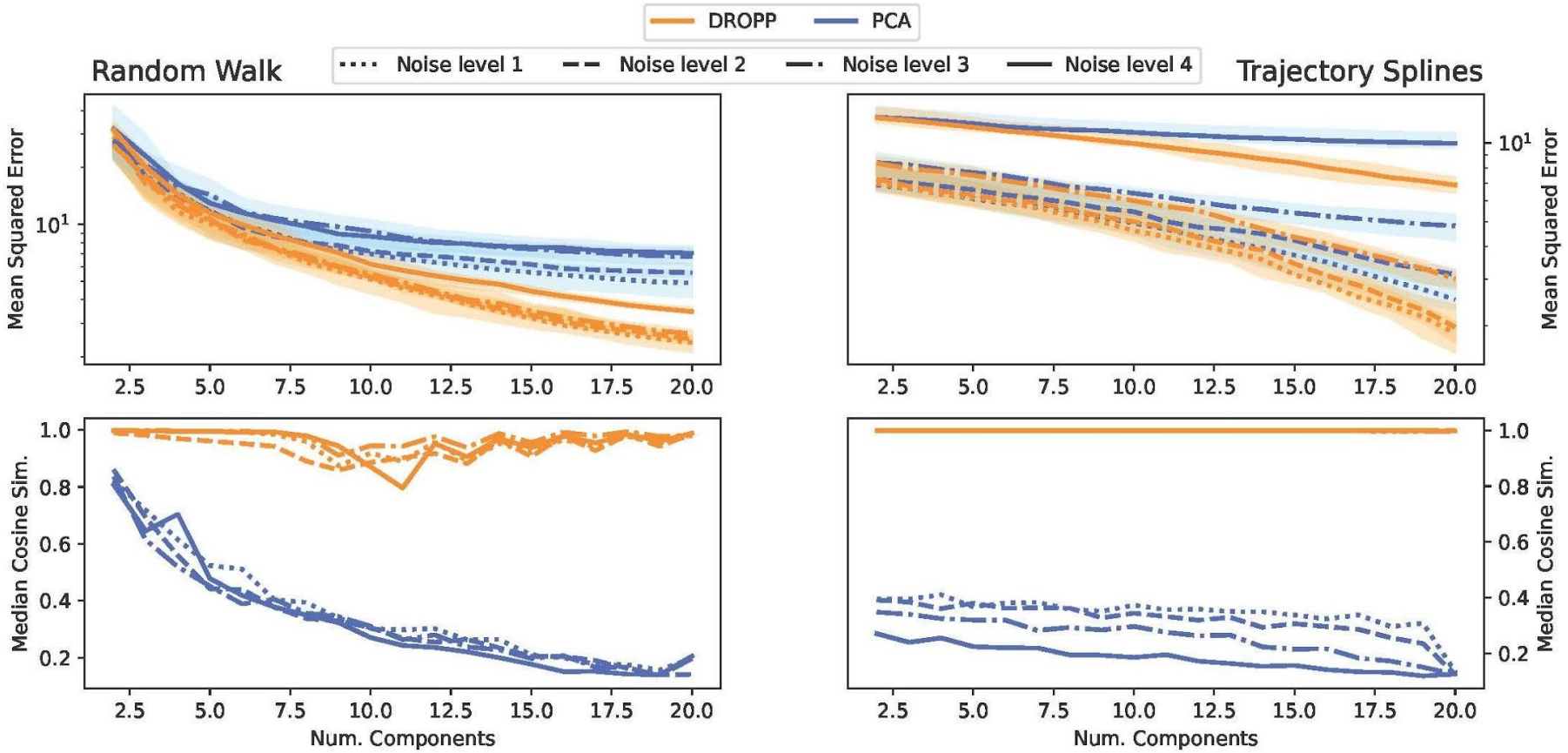
$$\mathcal{G}_{ij} = g(|i - j|)$$

Adaptions for tensor data

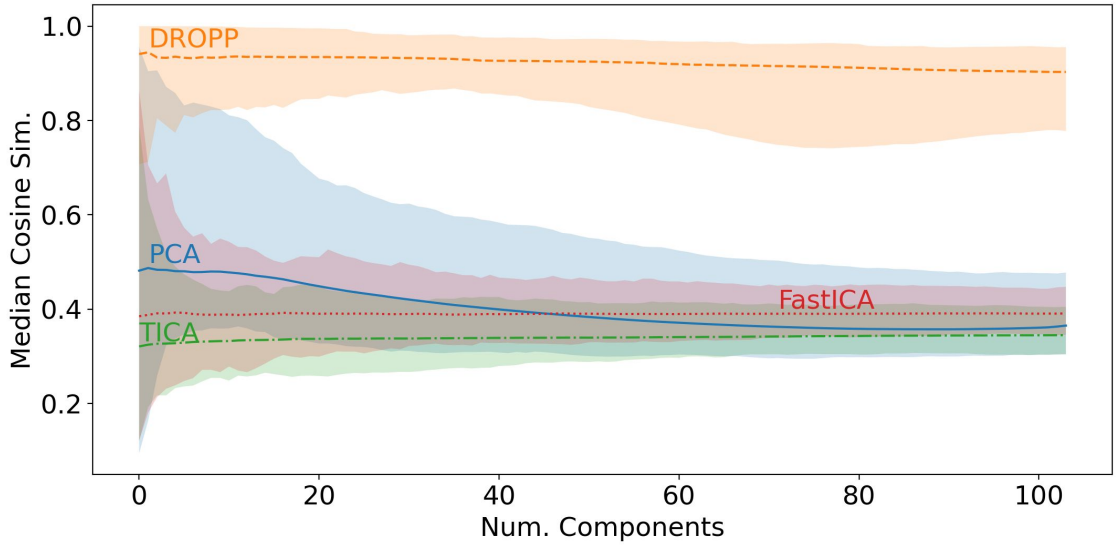
$$\widehat{\mathcal{C}}_{ij} = \begin{cases} 0 & \text{if } i \not\equiv j \mod 3 \\ \frac{1}{3}\sum_{c=1}^{3} Cov(\mathcal{X}_{:,:,c})_{\lfloor i/3 \rfloor, \lfloor j/3 \rfloor} & \text{else} \end{cases}$$

$$\widehat{\mathcal{G}}_{ij} = \begin{cases} 0 & \text{if } i \not\equiv j \mod 3 \\ g(|\lfloor i/3 \rfloor - \lfloor j/3 \rfloor|) & \text{else} \end{cases}$$

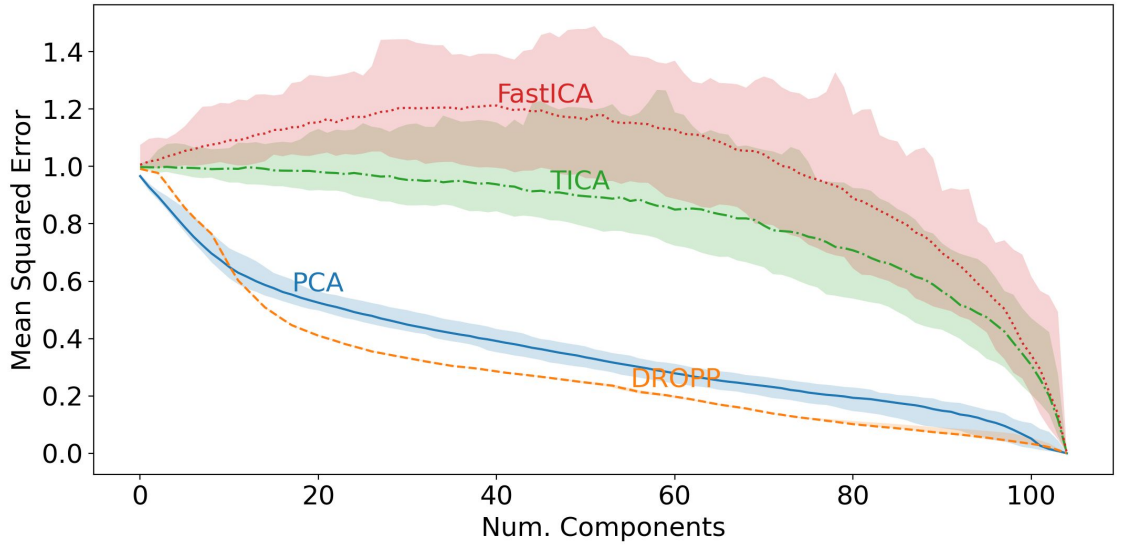
EXPERIMENTS ON SYNTHETIC DATA



EXPERIMENTS ON MD DATA



Similarity of components



Reconstruction error on different instantiations