Comparing Dimensionality Reduction Methods using Data Descriptor Landscapes

Bastian Rieck Heike Leitte

Interdisciplinary Center for Scientific Computing
Heidelberg University

UNIVERSITÄT HEIDELBERG
ZUKUNFT SEIT 1386

HGS MathComp

IWR

In an ideal world

Example workflow in data science

Multivariate data + Dimensionality reduction + Embedding = Insight

Disclaimer: This is *not* a definition but rather based on observing the machine learning colleagues down the hall…

An idealized metaphor

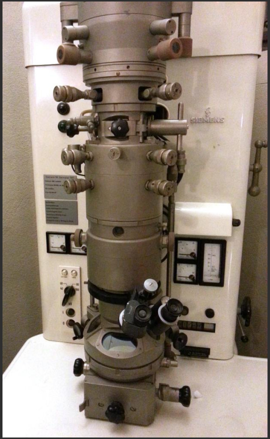


Image source: flickr user “Georg Holderied”

In practice

Opening the toolbox

Image source: flickr user “OZinOH”

In practice

Opening the toolbox

Locally linear embedding (LLE)

Image source: flickr user “OZinOH”

In practice

Opening the toolbox

Locally linear embedding (LLE)
Hessian locally linear embedding (HLLE)

Image source: flickr user “OZinOH”

In practice

Opening the toolbox

Locally linear embedding (LLE)

Hessian locally linear embedding (HLLE)

Local tangent space alignment (LTSA)

Image source: flickr user "OZinOH"

In practice

Opening the toolbox

Image source: flickr user “OZinOH”

Challenges

1. Choose a suitable tool.

Challenges

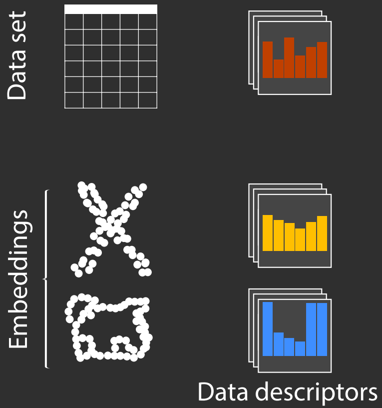
1. Choose a suitable tool.
2. Choose suitable parameters.

Challenges

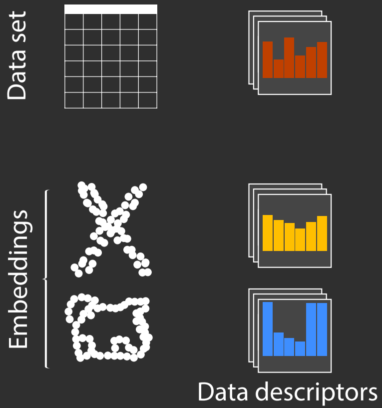
1. Choose a suitable tool.
2. Choose suitable parameters.

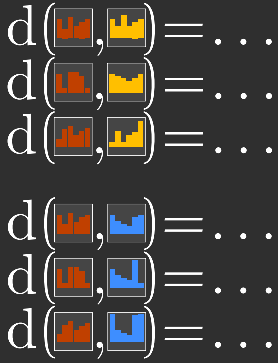
How can visualization help?

Our proposed approach

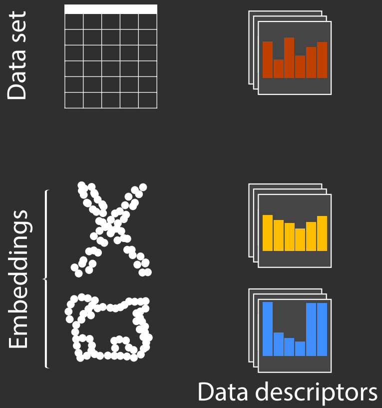


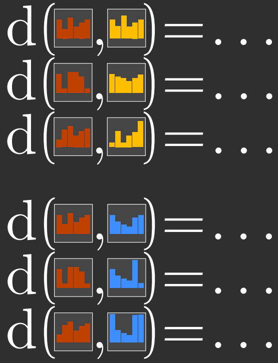
Our proposed approach





Our proposed approach





Idea: If data descriptors are *close*, the properties they measure are retained.

Data descriptors

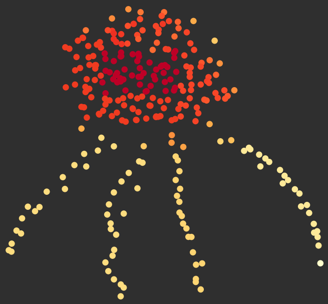
Definition & examples

Any function $f\colon \mathbb{D} \subseteq \mathbb{R}^n \to \mathbb{R}$ is an admissible data descriptor. f should measure "salient properties" of a multivariate data set.

Data descriptors

Definition & examples

Any function $f\colon \mathbb{D} \subseteq \mathbb{R}^n \to \mathbb{R}$ is an admissible data descriptor. f should measure "salient properties" of a multivariate data set.

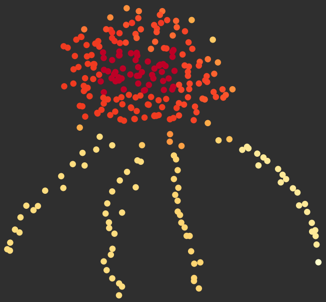


Density

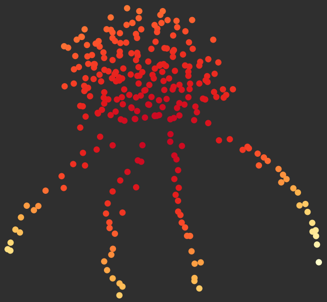
Data descriptors

Definition & examples

Any function $f\colon \mathbb{D} \subseteq \mathbb{R}^n \to \mathbb{R}$ is an admissible data descriptor. f should measure "salient properties" of a multivariate data set.



Density



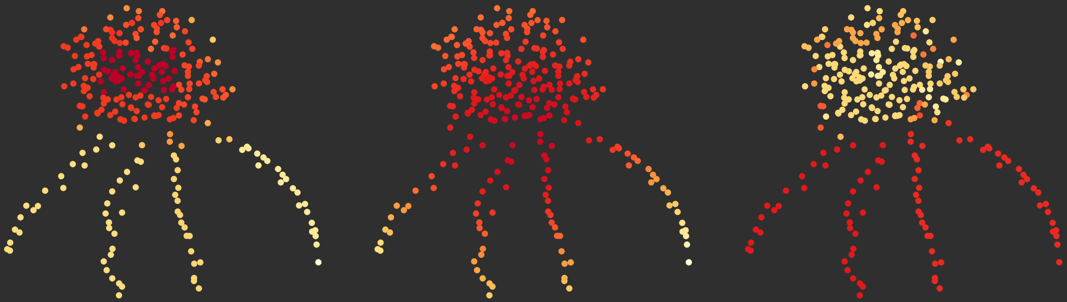
Eccentricity

Low  High

Data descriptors

Definition & examples

Any function $f\colon \mathbb{D} \subseteq \mathbb{R}^n \to \mathbb{R}$ is an admissible data descriptor. f should measure “salient properties” of a multivariate data set.



Density Eccentricity Local linearity

How to calculate data descriptor distances?

$\mathrm{d}($, $) := ?$

How to calculate data descriptor distances?

d( , ) $:= ?$

Persistent homology, because…

How to calculate data descriptor distances?

d( , ) $:= ?$

Persistent homology, because…

1. Calculation between different domains:

   Original data $\mathbb{R}^n$
   Embedding $\mathbb{R}^d$ $(d \ll n)$

How to calculate data descriptor distances?

$\mathrm{d}($  ,  $) := ?$

Persistent homology, because…

1. Calculation between different domains:

   Original data $\mathbb{R}^{n}$

   Embedding $\mathbb{R}^{d}$ $(d \ll n)$

2. Invariance to some transformations.

How to calculate data descriptor distances?

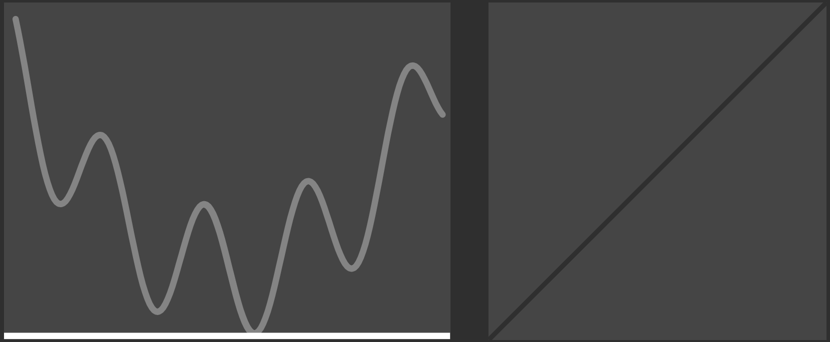
d( , ) $:= ?$

Persistent homology, because…

1. Calculation between different domains:
   Original data $\mathbb{R}^n$
   Embedding $\mathbb{R}^d$ $(d \ll n)$
2. Invariance to some transformations.
3. Robust against noise.

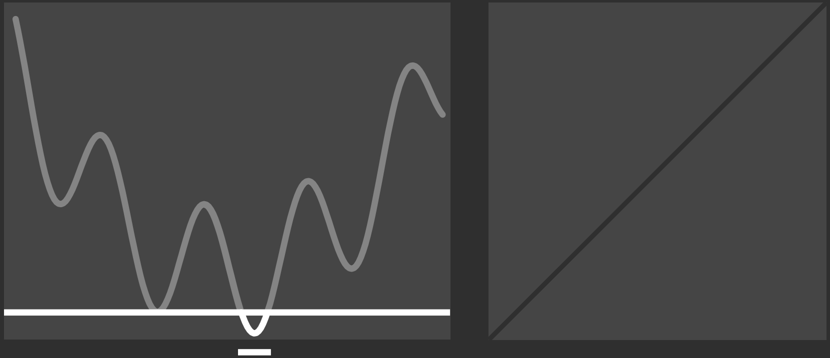
Persistent homology & persistence diagrams

One-dimensional example



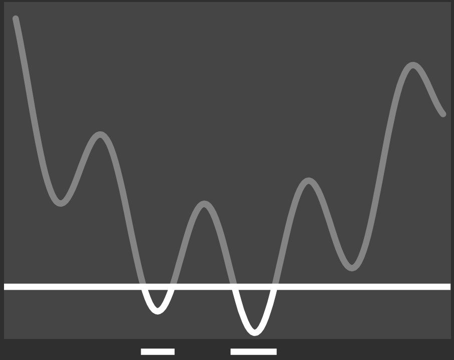
Persistent homology & persistence diagrams

One-dimensional example



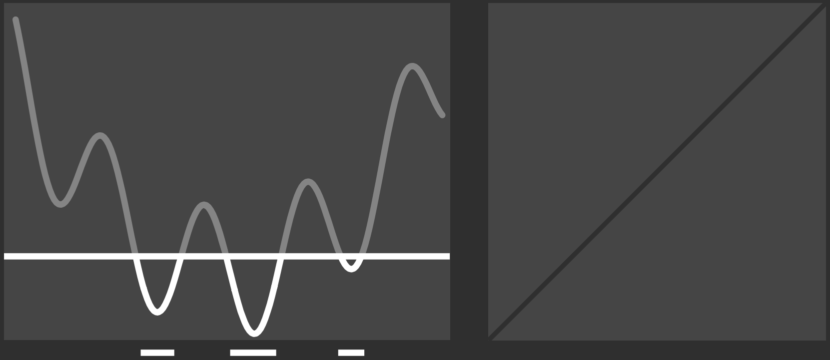
Persistent homology & persistence diagrams

One-dimensional example



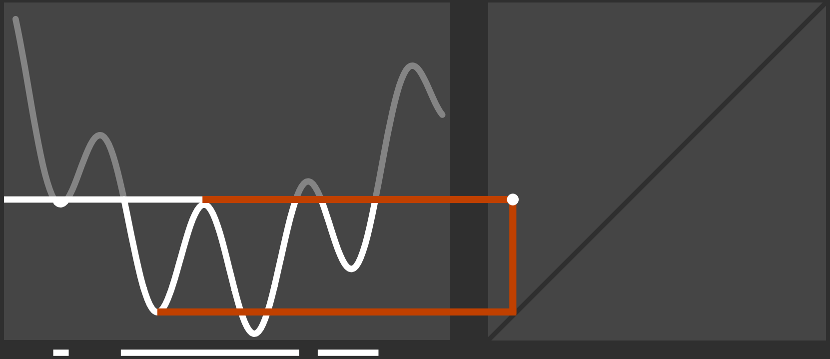
Persistent homology & persistence diagrams

One-dimensional example



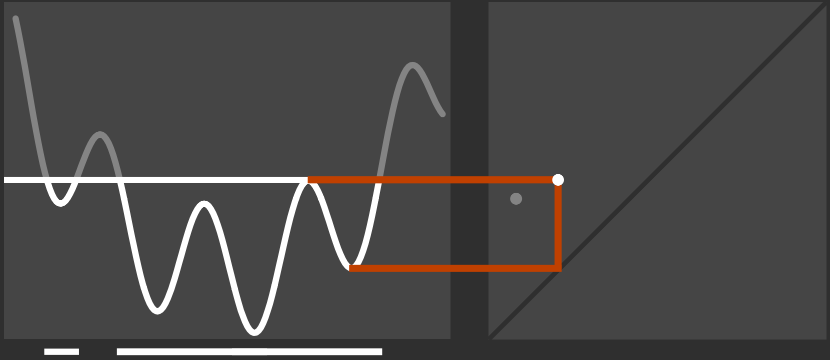
Persistent homology & persistence diagrams

One-dimensional example



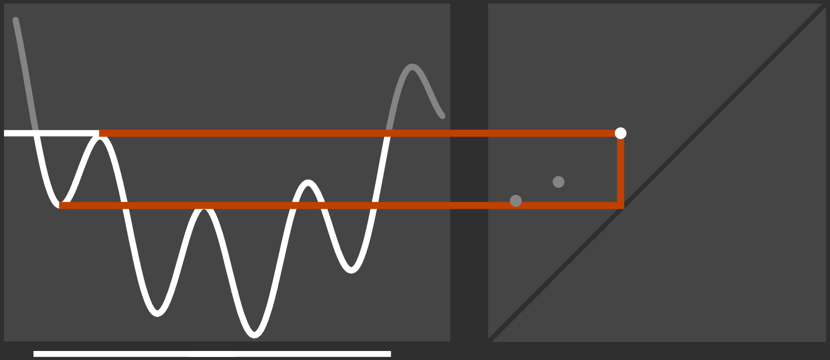
Persistent homology & persistence diagrams

One-dimensional example



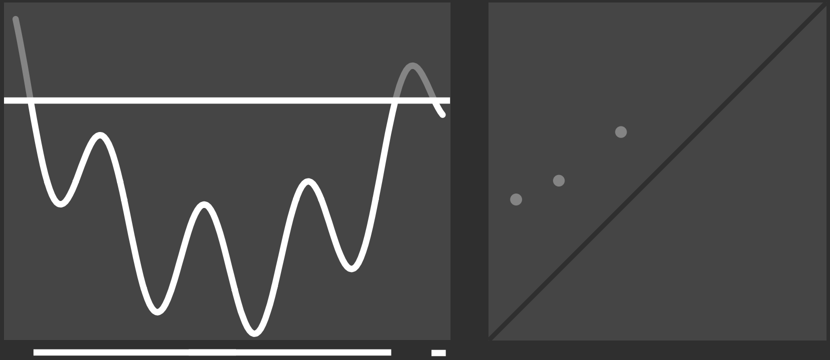
Persistent homology & persistence diagrams

One-dimensional example



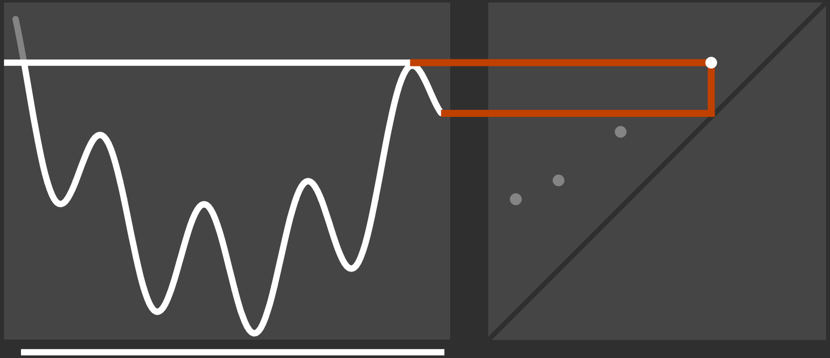
Persistent homology & persistence diagrams

One-dimensional example



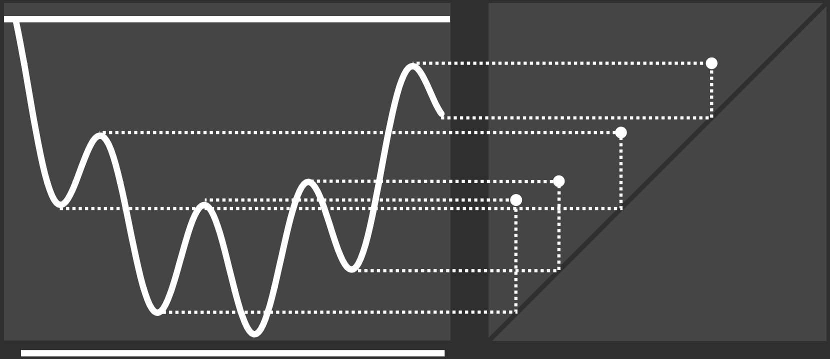
Persistent homology & persistence diagrams

One-dimensional example



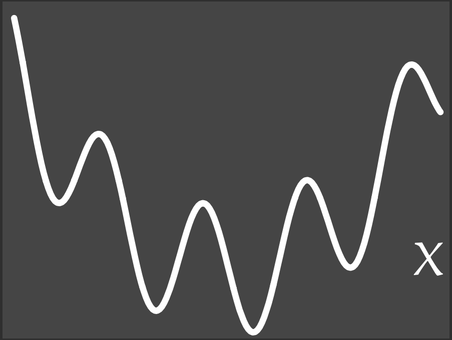
Persistent homology & persistence diagrams

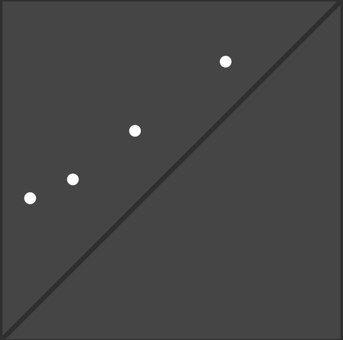
One-dimensional example



Persistent homology & persistence diagrams

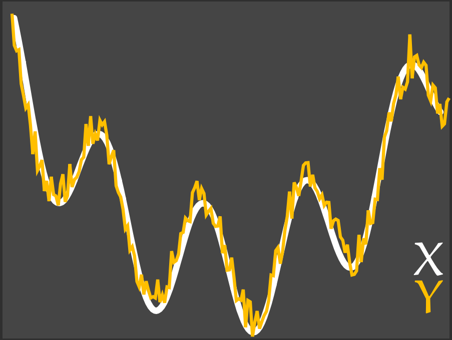
Distance calculations

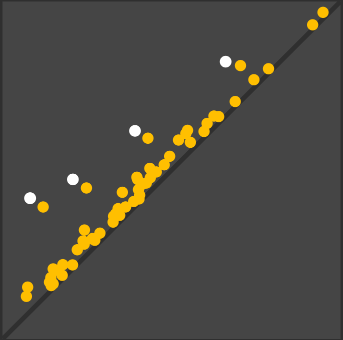




Persistent homology & persistence diagrams

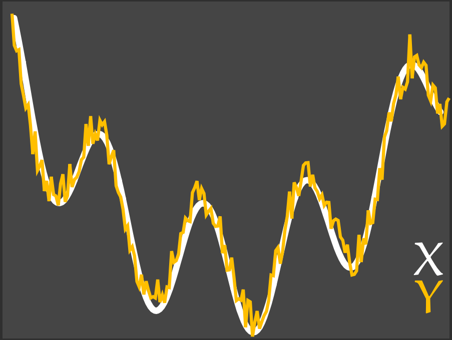
Distance calculations

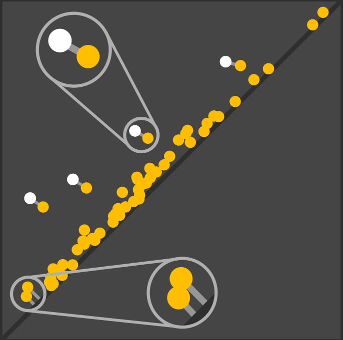




Persistent homology & persistence diagrams

Distance calculations

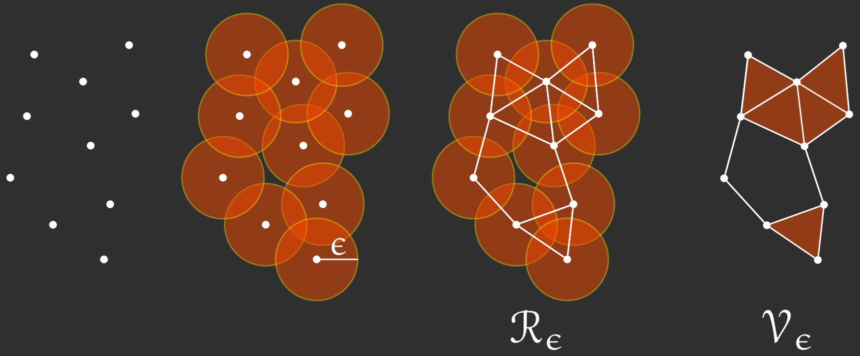




$$W_2(X, Y) = \sqrt{\inf_{\eta\colon X \to Y} \sum_{x \in X} \|x - \eta(x)\|_\infty^2}$$

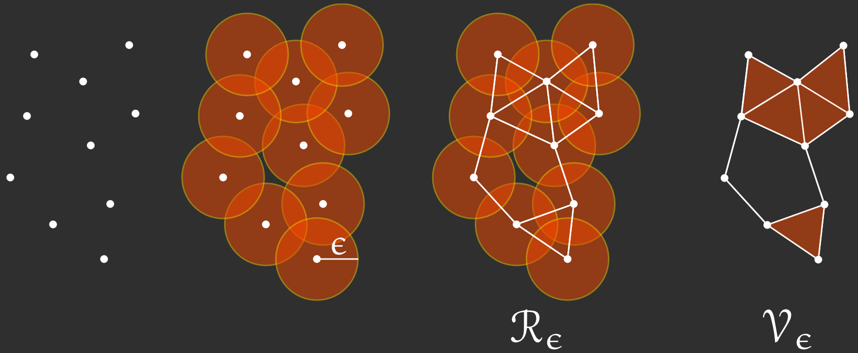
Extension to multivariate input data

Rips graphs and Vietoris–Rips complexes



Extension to multivariate input data

Rips graphs and Vietoris–Rips complexes



For $\epsilon \leqslant \epsilon'$, we have $\mathcal{R}_\epsilon \subseteq \mathcal{R}_{\epsilon'}$ and $\mathcal{V}_\epsilon \subseteq \mathcal{V}_{\epsilon'}$.

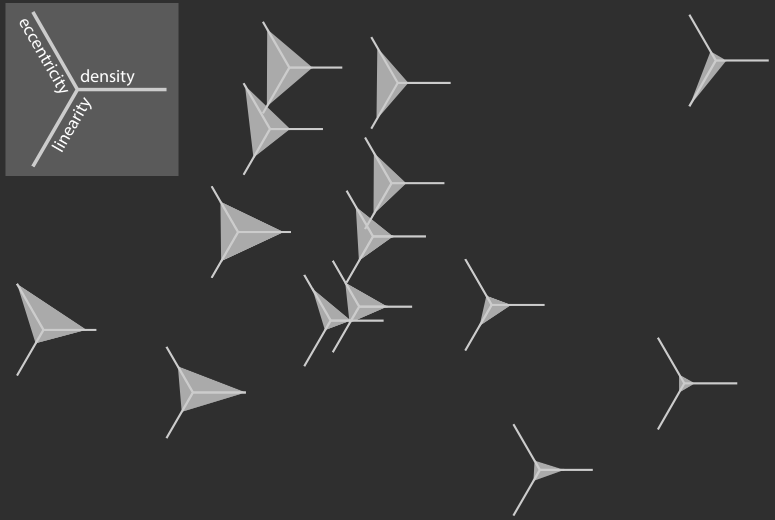
How to combine multiple data descriptors?

How to combine multiple data descriptors?

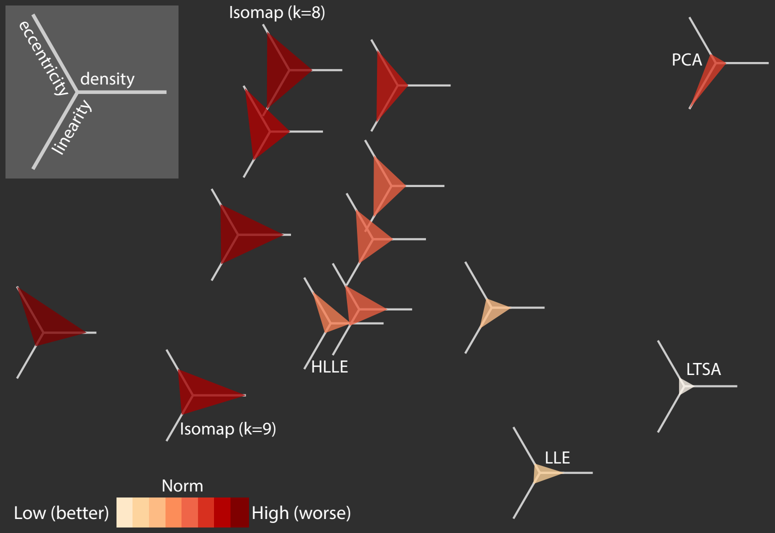
Use feature vectors as input for PCA!

Data descriptor landscapes

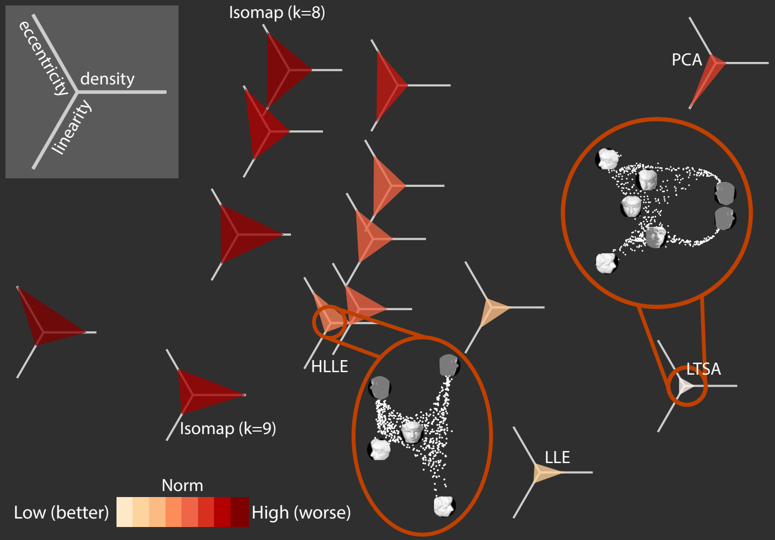
Data descriptor landscapes



Data descriptor landscapes



Data descriptor landscapes



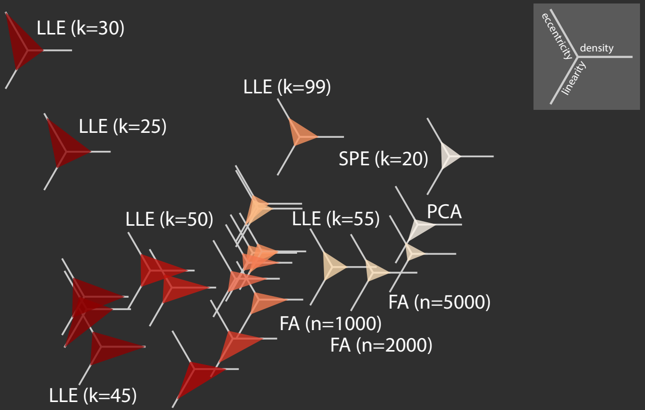
Results

Climate data

- Multivariate simulation data from *German Climate Research Centre* (DKRZ).
- Grid data from 192×96 locations with six continuous attributes over 1460 time steps.
- In total, 18432 vectors from $\mathbb{R}^6$.

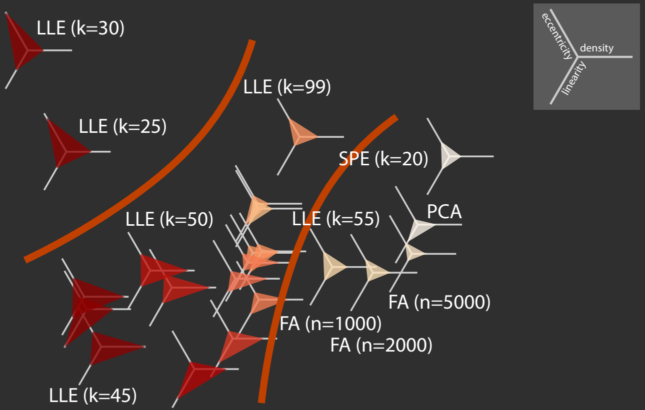
Data descriptor landscape for climate data

Linear vs. non-linear methods



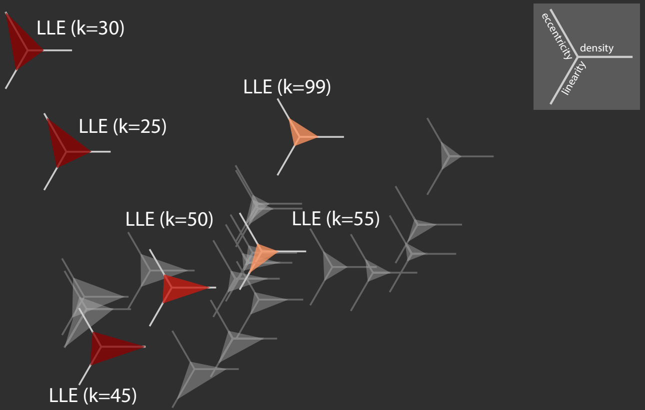
Data descriptor landscape for climate data

Linear vs. non-linear methods



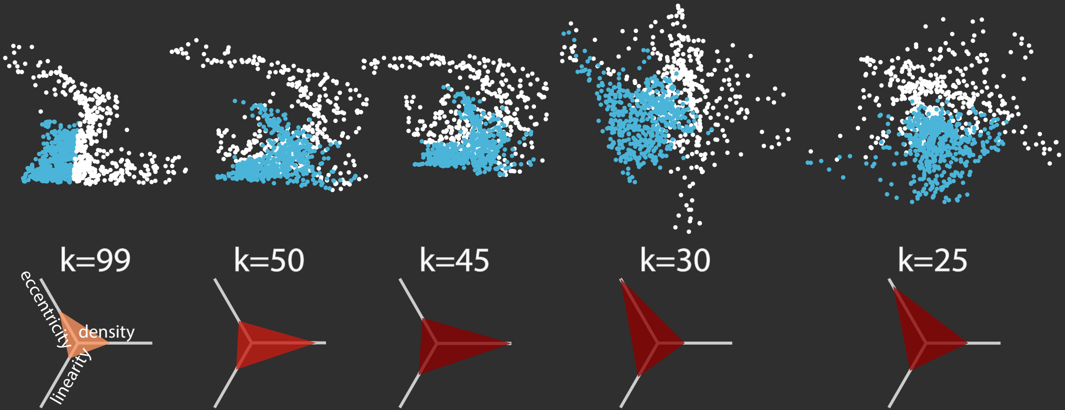
A closer look at LLE

Stability analysis

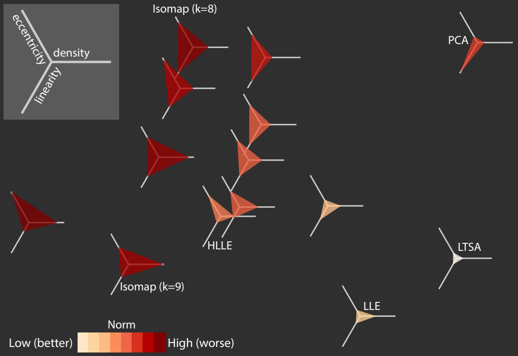


A closer look at LLE

Brushing some embeddings



Conclusion



- *Data descriptor approach*:
  - Functional description of salient properties.
  - Landscapes as a new visual metaphor for comparing multiple dimensionality reduction methods.
- Main applications: Variability and stability analysis of dimensionality reduction methods.
- Future work: Domain-specific data descriptors

Thank you for your attention!