Semi-Supervised Locally Linear Embedding (SSLLE)

Application & Sensitivity Analysis of Critical Hyperparameters

1 Problem
2 Local graph-based manifold learning (LGML)
3 Techniques
    1 Unsupervised
    2 Semi-supervised SSLLE
    3 Challenges
4 Sensitivity analysis
    1 Setup
    2 Results
5 Discussion

MANIFOLD LEARNING

Situation. Rapidly increasing amount of data thanks to novel applications and data sources

Problem. High data dimensionality detrimental to

- $\rightarrow$ Model functionality
- $\rightarrow$ Interpretability
- $\rightarrow$ Generalization ability

Manifold assumption. Data in high-dimensional observation space truly sampled from low-dimensional manifold

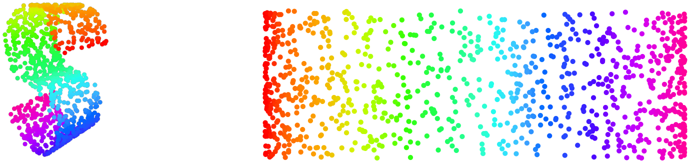
How to find a meaningful, structure-preserving embedding?

MANIFOLD LEARNING

Formal goal of manifold learning.

→ **Given.** Data $\mathcal{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_N)$, with $\boldsymbol{x}_i \in \mathbb{R}^D \; \forall i \in \{1, 2, ..., N\}$ and $N, D \in \mathbb{N}$, supposedly lying on d-dimensional manifold $\mathcal{M}$
$\Rightarrow \psi : \mathcal{M} \rightarrow \mathbb{R}^d$ with $d \ll D, d \in \mathbb{N}$
$\Rightarrow \mathcal{X} \sim \mathcal{M} \subset \mathbb{R}^D$

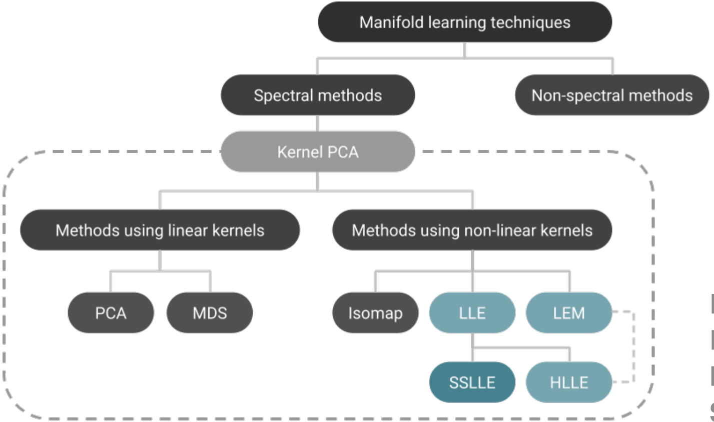
→ **Goal.** Find d-dimensional Euclidean representation
$\Rightarrow \mathcal{Y} = (\boldsymbol{y}_1, \boldsymbol{y}_2, ..., \boldsymbol{y}_N)$, with $\boldsymbol{y}_i = \psi(\boldsymbol{x}_i) \in \mathbb{R}^d \; \forall i \in \{1, 2, ..., N\}$.



2 LGML

TAXONOMY

Landscape. Various approaches, many of which may be translated into one another



LEM Laplacian eigenmaps
LLE Locally linear embedding
HLLE Hessian LLE
SSLLE Semi-supervised LLE

CONCEPT

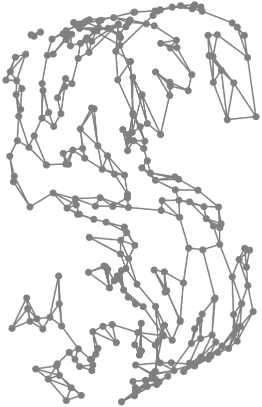
Idea. Capture intrinsic geometry, find principal axes of variability, retain most salient ones

Kernel PCA	LGML concept	LGML algorithm
Kernelization	Graph representation	Neighborhood construction
	Graph functional	Matrix representation
Eigenanalysis	Eigenanalysis	Eigenanalysis
Dimensionality reduction	Dimensionality reduction	Dimensionality reduction

Graph representation. Constructing a skeletal model of the manifold in $\mathbb{R}^D$

Vertices. Given by observations
Edges. Present between neighboring points

- $\rightarrow$ Typically, *k*-neighborhoods
- $\rightarrow$ Edge weights determined by nearness



Graph functional. Belief about intrinsic manifold properties at the heart of each method

- $\rightarrow$ Smoothness LEM
- $\rightarrow$ Local linearity LLE SSLLE
- $\rightarrow$ Curviness HLLE
- $\rightarrow$ ...

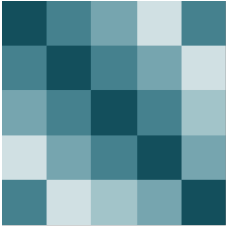
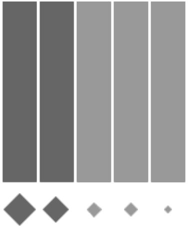
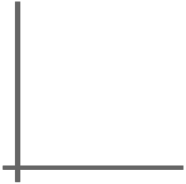
Achievements: non-linearity & locality

2 LGML CONCEPT

Eigenanalysis. Finding axes of variability in intrinsic manifold structure

- → Matrix representation of manifold properties
- → Assessment through eigenanalysis
  - → Directions of variability ⇒ eigenvectors
  - → Respective degrees of variability ⇒ eigenvalues

Dimensionality reduction. Projection into subspace spanned by *d* principal eigenvectors

3 TECHNIQUES

LEM

Proposal. Belkin and Niyogi (2001)

Idea. Forcing nearby inputs to be mapped to nearby outputs

$\rightarrow$ Notion of smoothness in mapping function

$\rightarrow$ Second-order penalty on gradient

Solution. Eigenanalysis of graph Laplacian $\boldsymbol{L}$

$\rightarrow$ Derived from weight matrix encoding nearness of inputs

$\rightarrow$ Discrete approximation of Laplace-Beltrami operator $\mathcal{L}(f)$

$\rightarrow$ Generalized eigenvalue problem

Solution: bottom $d + 1$ eigenvectors

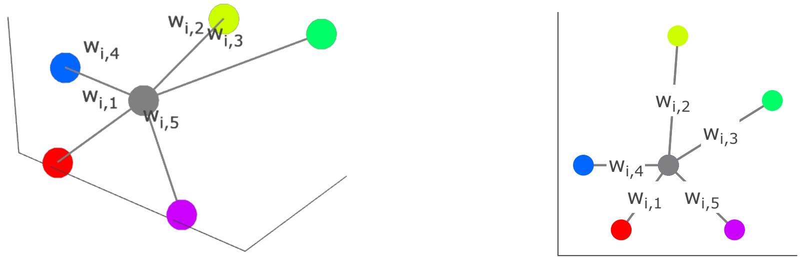
Proposal. Roweis and Saul (2000)

Idea. Preserving locally linear reconstructions

$\rightarrow$ Linear reconstruction of points in $\mathbb{R}^D$ by their neighbors

$\rightarrow$ Reconstruction weights = topological properties

$\rightarrow$ Neighborhood patches invariant to dimensionality reduction



3.1 UNSUPERVISED **LLE**

Reconstruction loss minimization. Finding optimal reconstruction weights

$$\min_{\boldsymbol{W}} \varepsilon(\boldsymbol{W}) = \min_{\boldsymbol{W}} \sum_i \left\| \boldsymbol{x}_i - \sum_j w_{ij} \boldsymbol{x}_j \right\|^2, \quad \text{s.t. } \boldsymbol{1}^T \boldsymbol{w}_i = 1 \quad \forall i \in \{1, 2, ..., N\} \tag{1}$$

Embedding loss minimization. Finding optimal embedding coordinates

$$\min_{\mathcal{Y}} \Phi(\mathcal{Y}) = \min_{\mathcal{Y}} \sum_i \left\| \boldsymbol{y}_i - \sum_j w_{ij} \boldsymbol{y}_j \right\|^2, \quad \text{s.t. } \frac{1}{N} \sum_i \boldsymbol{y}_i \boldsymbol{y}_i^T = \boldsymbol{I}, \quad \sum_i \boldsymbol{y}_i = \boldsymbol{0} \quad \forall i \in \{1, 2, ..., N\} \tag{2}$$

Eigenvalue problem. Define $\boldsymbol{E} = (\boldsymbol{I} - \boldsymbol{W})^T (\boldsymbol{I} - \boldsymbol{W})$, such that

$$\min_{\mathcal{Y}} \mathit{trace}(\mathcal{Y}^T \boldsymbol{E} \mathcal{Y}), \quad \text{s.t. } \frac{1}{N} \mathcal{Y}^T \mathcal{Y} = \boldsymbol{I}, \quad \mathcal{Y}^T \boldsymbol{1} = \boldsymbol{0}. \tag{3}$$

Solution: bottom $d + 1$ eigenvectors

HLLE

Proposal. Donoho and Grimes (2003)

Idea. Finding a truly linear mapping while preserving local isometry

$\rightarrow$ Notion of curviness in mapping function

$\rightarrow$ Second-order penalty on Hessian

$\rightarrow$ Strong convergence guarantees but rather complex computations

Solution. Eigenanalysis of empirical Hessian functional $\mathcal{H}$

$\rightarrow$ Derived as a quadratic form of Hessian estimators in linear neighborhood patches

$\rightarrow$ Discrete approximation of continuous Hesssian functional $\mathscr{H}(f)$

$\rightarrow$ Null space problem

$\boxed{\textbf{Solution: bottom } d + 1 \textbf{ eigenvectors + scaling}}$

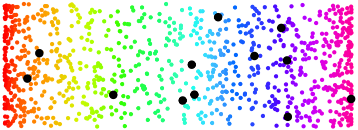
Proposal. Yang et al. (2006)

Problem. Embedding found by unsupervised methods not always meaningful

Idea. Improving LLE by use of prior knowledge

Semi-supervision. Anchoring embedding at some prior points with known coordinates

- $\rightarrow$ More active than semi-supervised learning?
- $\rightarrow$ Setting. Information available or to be obtained by querying the oracle
- $\rightarrow$ Goal. Maximum information at little expense $\Rightarrow$ careful choice of prior points



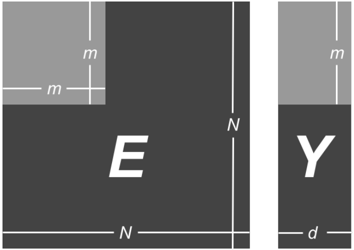
Types of prior information. Exact vs inexact

$\rightarrow$ Level of confidence encoded in parameter β

Algorithmic impact. Recall LLE eigenvalue problem

$$\min_{\mathcal{Y}} \textit{trace}(\mathcal{Y}^T \boldsymbol{E} \mathcal{Y}), \quad \text{s.t. } \frac{1}{N} \mathcal{Y}^T \mathcal{Y} = \boldsymbol{I}, \quad \mathcal{Y}^T \mathbf{1} = \mathbf{0}.$$

$\Rightarrow$ Partitioning of $\boldsymbol{E}$ and $\mathcal{Y}$



Modified optimization problem. Exact information

$$\min_{\mathcal{Y}_2} \begin{bmatrix} \mathcal{Y}_1 & \mathcal{Y}_2 \end{bmatrix} \begin{bmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{bmatrix} \begin{bmatrix} \mathcal{Y}_1^T \\ \mathcal{Y}_2^T \end{bmatrix} \tag{4}$$

$$\Leftrightarrow \mathcal{Y}_2^T = M_{22}^{-1} M_{12} \mathcal{Y}_1^T \tag{5}$$

Modified optimization problem. Inexact information

$$\min_{\mathcal{Y}_1, \mathcal{Y}_2} \begin{bmatrix} \mathcal{Y}_1 & \mathcal{Y}_2 \end{bmatrix} \begin{bmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{bmatrix} \begin{bmatrix} \mathcal{Y}_1^T \\ \mathcal{Y}_2^T \end{bmatrix} + \beta \left\| \mathcal{Y}_1^T - \hat{\mathcal{Y}}_1^T \right\|_F^2 \tag{6}$$

$$\Leftrightarrow \begin{bmatrix} M_{11} + \beta \boldsymbol{I} & M_{12} \\ M_{21} & M_{22} \end{bmatrix} \begin{bmatrix} \mathcal{Y}_1^T \\ \mathcal{Y}_2^T \end{bmatrix} = \begin{bmatrix} \hat{\mathcal{Y}}_1^T \\ \mathbf{0} \end{bmatrix} \tag{7}$$

SSLLE

Choice of landmark points. Basically, three options

- $\rightarrow$ Pre-existing prior information $\Rightarrow$ worst case: poor coverage
- $\rightarrow$ (Uniform) random sampling
- $\rightarrow$ Maximum coverage $\Rightarrow$ minimization of condition number $\kappa(M_{22})$

CRITICAL PARAMETERS

Intrinsic dimensionality. True sources of variability

$\rightarrow$ Considered known with availability of prior information

Neighborhood size. Global vs local structure

$\rightarrow$ Tunable (expensive)

Regularization constant. Singularity for $D < k$

$\rightarrow$ Heuristics

Number & location of prior points. Utility of prior knowledge ANALYSIS

$\rightarrow$ Exploration vs labeling cost

Noise level. Quality of prior knowledge ANALYSIS

$\rightarrow$ How exact must prior information be?

Confidence parameter. Strength of belief in prior knowledge

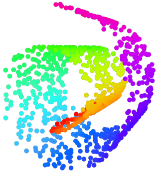
$\rightarrow$ Rather robust

4 SENSITIVITY ANALYSIS

DATA

Data. Two data sets, $N = 1000$ observations each

Swiss roll. *The* standard synthetic manifold

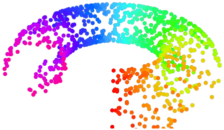


1 Sample $\boldsymbol{u}_1, \boldsymbol{u}_2 \sim U(0, 1)$ *iid* with $|\boldsymbol{u}_1| = |\boldsymbol{u}_1| = N$

2 Compute $\boldsymbol{t} = 1.5\pi(1 + 2\boldsymbol{u}_1)$ and $\boldsymbol{s} = 21\boldsymbol{u}_2$

3 $\mathcal{X}_{\text{swiss}} = \begin{bmatrix} \boldsymbol{x}_1 & \boldsymbol{x}_2 & \boldsymbol{x}_3 \end{bmatrix} = \begin{bmatrix} \boldsymbol{t}\cos\boldsymbol{t} & \boldsymbol{s} & \boldsymbol{t}\sin\boldsymbol{t} \end{bmatrix}$

Incomplete tire. Examined in Yang et al. (2006)



1 Sample $\boldsymbol{u}_1, \boldsymbol{u}_2 \sim U(0, 1)$ *iid* with $|\boldsymbol{u}_1| = |\boldsymbol{u}_1| = N$

2 Compute $\boldsymbol{t} = \frac{5\pi}{3}\boldsymbol{u}_1$ and $\boldsymbol{s} = \frac{5\pi}{3}\boldsymbol{u}_2$

3 $\mathcal{X}_{\text{tire}} = \begin{bmatrix} \boldsymbol{x}_1 & \boldsymbol{x}_2 & \boldsymbol{x}_3 \end{bmatrix}$
$= \begin{bmatrix} (3 + \cos\boldsymbol{s})\cos\boldsymbol{t} & (3 + \cos\boldsymbol{s})\sin\boldsymbol{t} & \sin\boldsymbol{s} \end{bmatrix}$

SCENARIOS

Sensitivity analysis I. Landmark coverage $\times$ number of landmark points

→ Landmark coverage $\in \{\text{poor}, \text{random}, \text{maximum}\}$

→ Number of landmark points $\in \{2, 4, 6, 8, 10, 12\}$

Sensitivity analysis II. Noise level $\times$ number of landmark points

→ Simulation of inexact prior information through additive Gaussian noise

→ Corruption of landmark $\boldsymbol{p}$ as $\tilde{\boldsymbol{p}} = \boldsymbol{p} + \boldsymbol{\epsilon} = (p_t, p_s) + (\epsilon_t, \epsilon_s)$
with $\epsilon_i \sim N(0, (\alpha \cdot s_i)^2)$ scaled by empirical variance s_i, $i \in \{t, s\}$

→ Noise level $\alpha \in \{0.1, 0.5, 1.0, 3.0\}$

→ Number of landmark points $\in \{2, 4, 6, 8, 10, 12\}$

EVALUATION

Evaluation criterion. AUC(R_{NX}) (Kraemer et al., 2019)

$\rightarrow$ Area under the R_{NX} curve

$\rightarrow$ Based on co-ranking matrix

Co-ranking matrix. Comparing distance ranks in observation & embedding spaces

$\rightarrow$ Rank distance matrices. $(r)_{ij}^{\text{obs}}$ and $(r)_{ij}^{\text{emb}}$

$\rightarrow$ Co-ranking matrix. $\boldsymbol{Q} = (q)_{\ell m}$ with $q_{\ell m} = |\{(i, j) : r_{ij}^{\text{emb}} = \ell \wedge r_{ij}^{\text{obs}} = m\}|$

$\rightarrow$ Interpretation.

1. All non-zero entries on diagonal $\Rightarrow$ optimal embedding
2. Most non-zero entries on upper triangle $\Rightarrow$ close points torn apart
3. Most non-zero entries on lower triangle $\Rightarrow$ faraway poits collapsed

EVALUATION

Co-ranking-based metrics. Comparing distance ranks in observation & embedding spaces

$\rightarrow$ Number of points belonging to k-neighborhood in both spaces.
$Q_{NX}(k) = \frac{1}{kN} \sum_{\ell=1}^{k} \sum_{m=1}^{k} q_{\ell m}$

$\rightarrow$ Adjustment for random embeddings and normalization.
$R_{NX}(k) = \frac{(N-1)Q_{NX}(k)-k}{N-1-k}$

AUC measure. Parameter-free

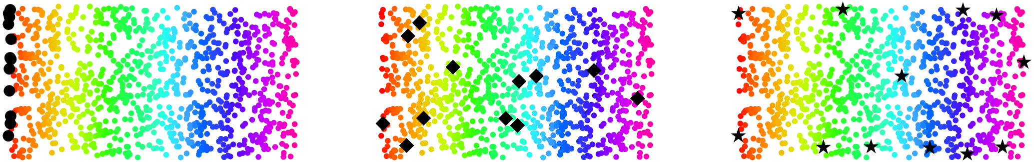
$\rightarrow$ $\text{AUC}(R_{NX}) = \frac{\sum_{k=1}^{N-2} R_{NX}(k)}{\sum_{k=1}^{N-2} 1/k} \in [0, 1]$

$\rightarrow$ Interpretation.

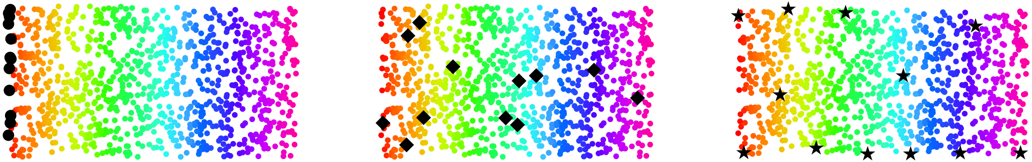
1. $\text{AUC}(R_{NX}) = 0 \Rightarrow$ random embedding
2. $\text{AUC}(R_{NX}) = 1 \Rightarrow$ optimal embedding

SENSITIVITY ANALYSIS I

Key variation. Swiss roll – poor, random, maximum coverage

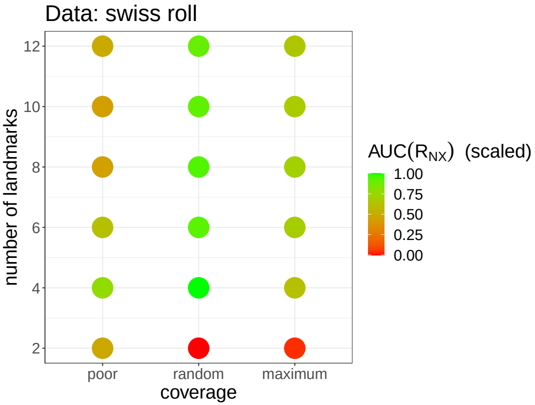


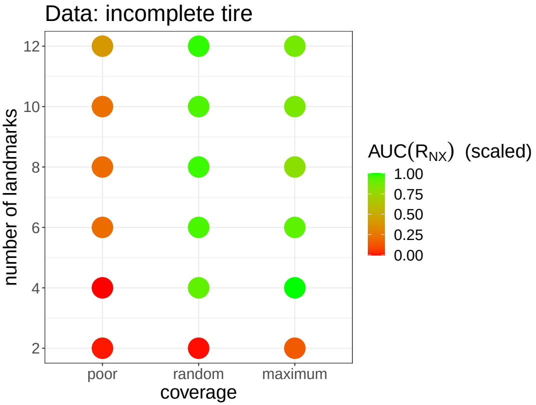
Key variation. Incomplete tire – poor, random, maximum coverage



SENSITIVITY ANALYSIS I

Quantitative results. Seemingly better performance of random coverage

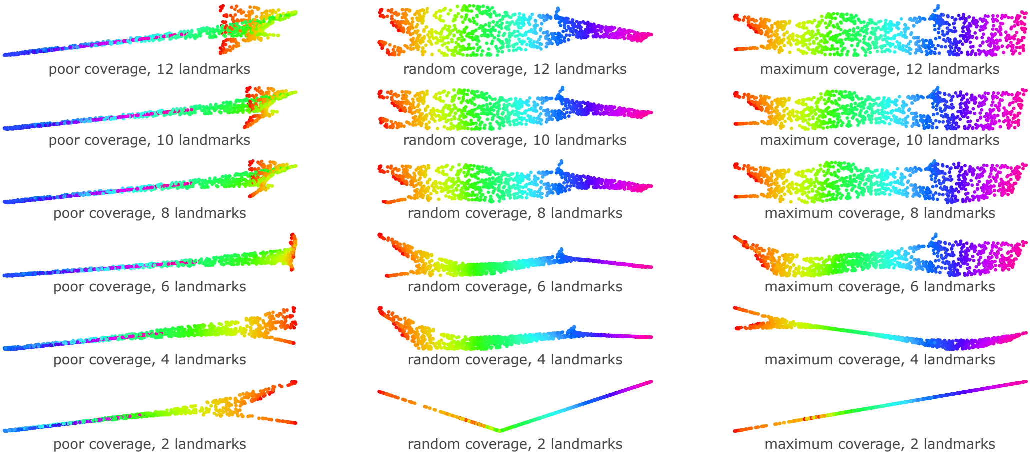




AUC(R_{NX}) has been scaled to take on a minimum of 0 and maximum of 1 in both figures for better visibility of differences. Original scales. Swiss roll: AUC$(R_{NX}) \in [0.2655, 0.4086]$, incomplete tire: AUC$(R_{NX}) \in [0.2772, 0.6231]$

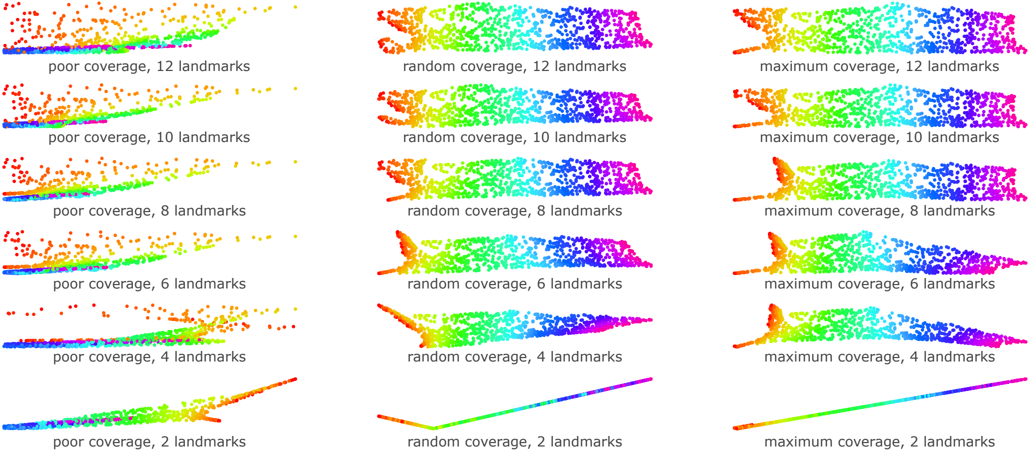
SENSITIVITY ANALYSIS I

Qualitative results. Swiss roll



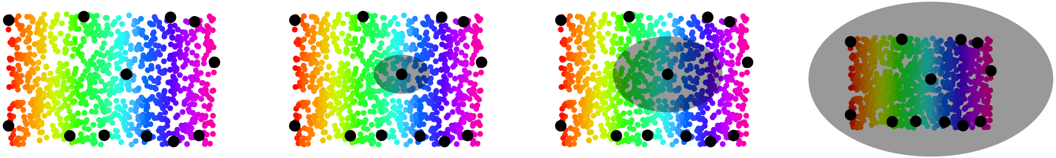
SENSITIVITY ANALYSIS I

Qualitative results. Incomplete tire

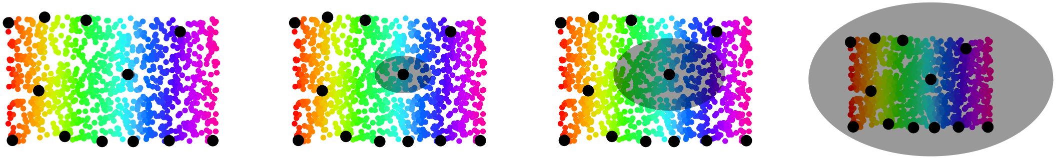


SENSITIVITY ANALYSIS II

Key variation. Swiss roll – $\alpha \in \{0.1, 0.5, 1.0, 3.0\}$

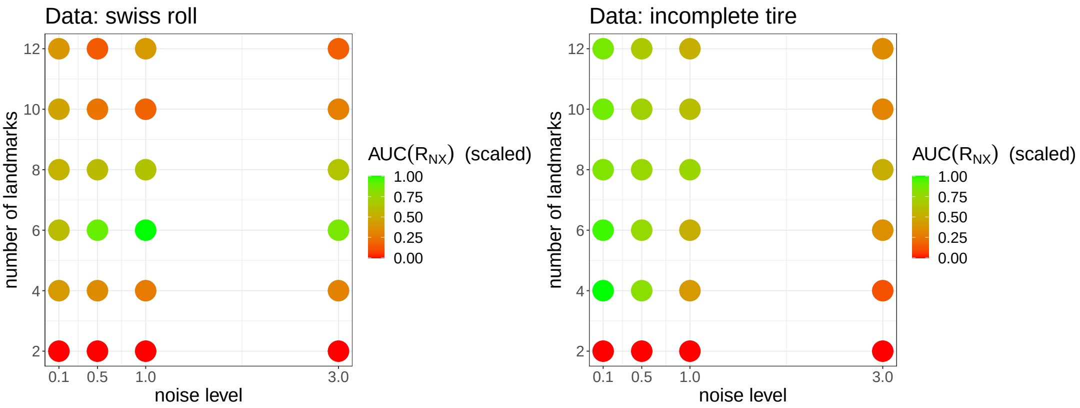


Key variation. Incomplete tire – $\alpha \in \{0.1, 0.5, 1.0, 3.0\}$



SENSITIVITY ANALYSIS II

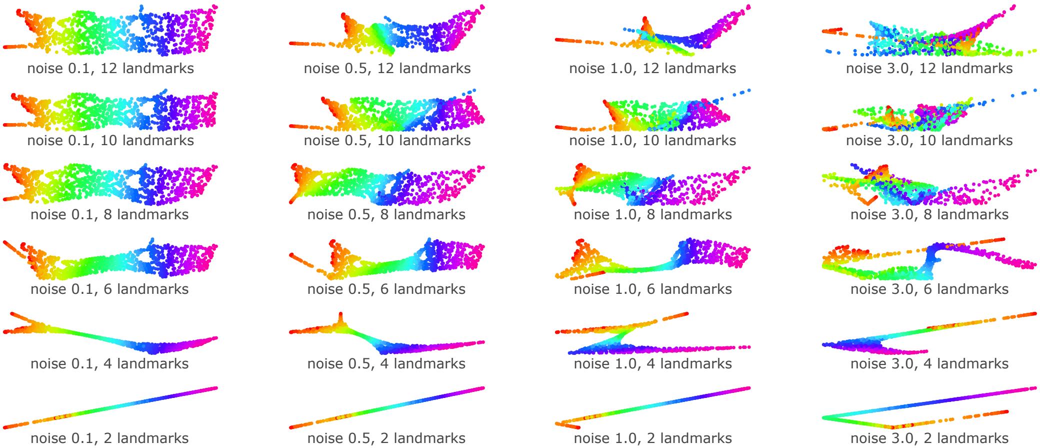
Quantitative results. Some compensation of noise by larger number of landmarks



$\mathrm{AUC}(R_{NX})$ has been scaled to take on a minimum of 0 and maximum of 1 in both figures for better visibility of differences. Original scales. Swiss roll: $\mathrm{AUC}(R_{NX}) \in [0.2720, 0.4167]$, incomplete tire: $\mathrm{AUC}(R_{NX}) \in [0.3171, 0.6172]$

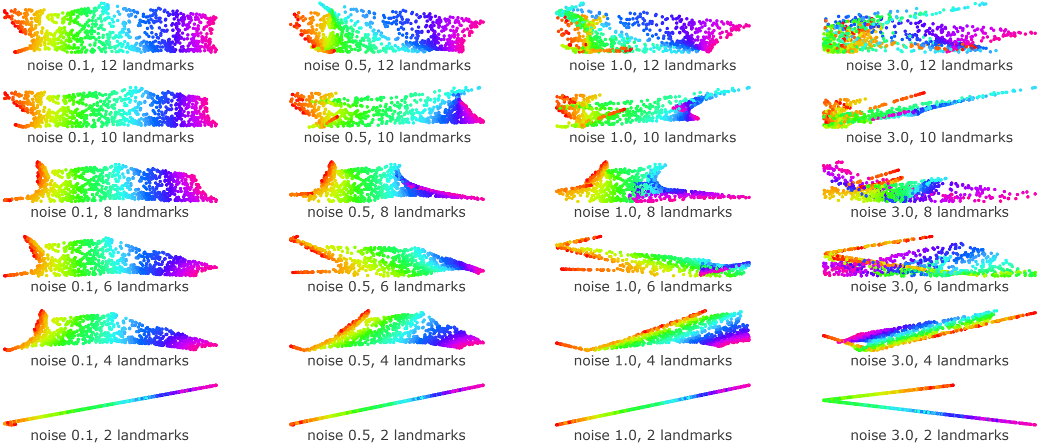
SENSITIVITY ANALYSIS II

Qualitative results. Swiss roll



4.2 RESULTS SENSITIVITY ANALYSIS II

Qualitative results. Incomplete tire



5 DISCUSSION

FOO

woteva

MAIN REFERENCES

Belkin, M. and Niyogi, P. (2001). Laplacian eigenmaps and spectral technique for embedding and clustering, *Proceedings of the 14th International Conference on Neural Information Processing Systems: Natural and Synthetic*, p. 585–591.

Donoho, D. L. and Grimes, C. (2003). Hessian eigenmaps: Locally linear embedding techniques for high-dimensional data, *Proceedings of the National Academy of Sciences of the United States of America* **100**(10): 5591–5596.

Kraemer, G., Reichstein, M. and Mahecha, M. D. (2019). dimred and coranking -— unifying dimensionality reduction in r, *Technical report*.

Roweis, S. T. and Saul, L. K. (2000). Nonlinear dimensionality reduction by locally linear embedding, *Science* **290**(5500): 2323–2326.

Yang, X., Fu, H., Zha, H. and Barlow, J. (2006). Semi-supervised nonlinear dimensionality reduction, *Proceedings of the 23rd International Conference on Machine Learning*.