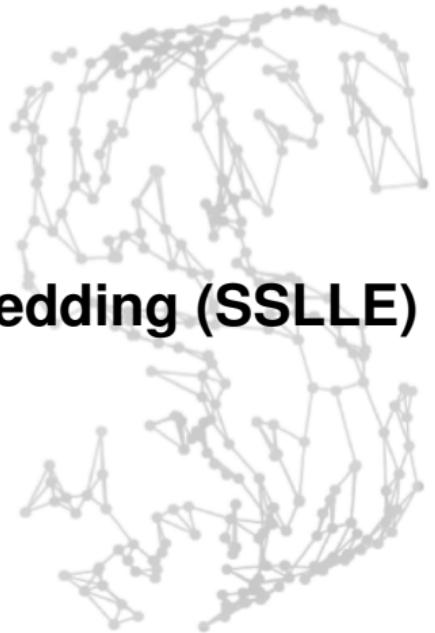


# Semi-Supervised Locally Linear Embedding (SSLLE)

Application & Sensitivity Analysis of Critical Parameters



# 0 AGENDA

---

- 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

# 1 PROBLEM MANIFOLD LEARNING

---

**Situation:** rapidly increasing amount of data thanks to novel applications and data sources

**Problem:** high data dimensionality detrimental to

- 1 model functionality
- 2 interpretability

**Manifold assumption:** data in high-dimensional observation space truly sampled from low-dimensional manifold



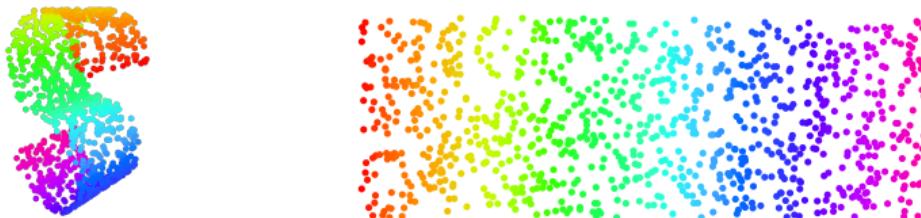
**How to find a meaningful, structure-preserving embedding?**

# 1 PROBLEM MANIFOLD LEARNING

---

**Formal goal of manifold learning:**

- **Given:** data  $\mathcal{X} = (\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N)$ , with  $\mathbf{x}_i \in \mathbb{R}^D \forall i \in \{1, 2, \dots, N\}$  and  $N, D \in \mathbb{N}$ , supposedly lying on  $d$ -dimensional manifold  $\mathcal{M}$   
⇒  $\psi : \mathcal{M} \rightarrow \mathbb{R}^d$  with  $d \ll D, d \in \mathbb{N}$   
⇒  $\mathcal{X} \sim \mathcal{M} \subset \mathbb{R}^D$
- **Goal:** find  $d$ -dimensional Euclidean representation  
⇒  $\mathcal{Y} = (\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_N)$ , with  $\mathbf{y}_i = \psi(\mathbf{x}_i) \in \mathbb{R}^d \forall i \in \{1, 2, \dots, N\}$ .



---

## 2 LGML

---

## 2 LGML TAXONOMY

**Landscape:** various approaches, many of which may be translated into one another

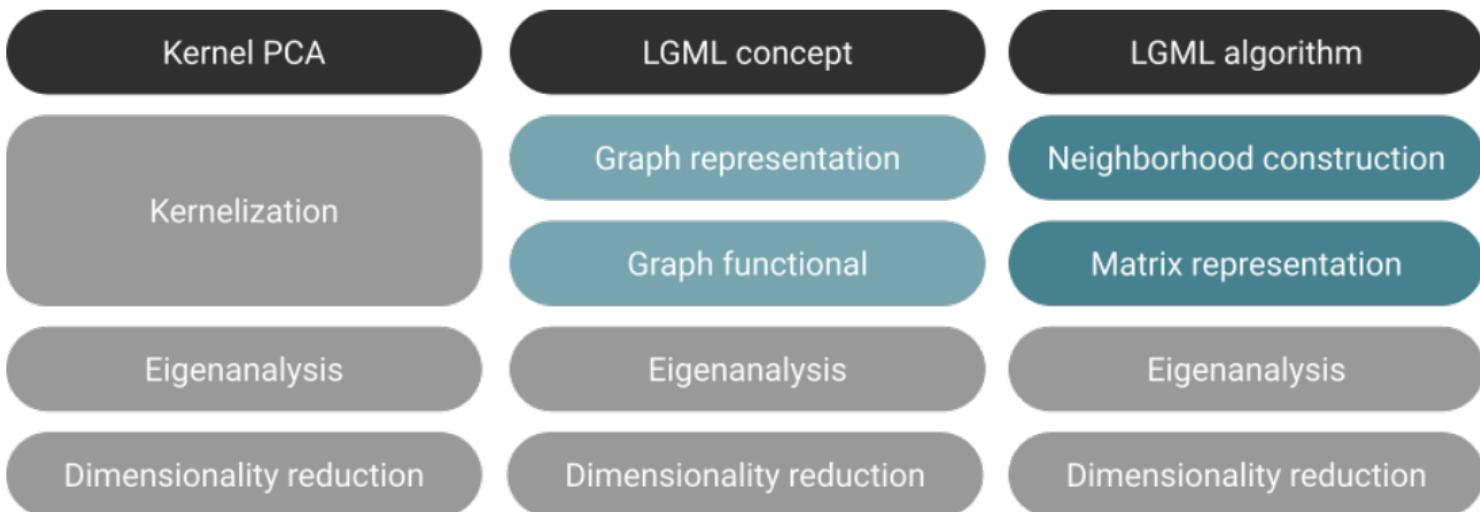


**LEM** Laplacian eigenmaps  
**LLE** locally linear embedding  
**HLLE** Hessian LLE  
**SSLLE** semi-supervised LLE

## 2 LGML CONCEPT

---

**Idea:** capture intrinsic geometry, find principal axes of variability, retain most salient ones



**Achievements: non-linearity & locality**

## 2 LGML CONCEPT

---

**Graph representation:** constructing a skeletal model of the manifold in  $\mathbb{R}^D$

**Vertices:** given by observations

**Edges:** present between neighboring points

- Typically,  $k$ -neighborhoods
- Edge weights determined by nearness

**Graph functional:** belief about intrinsic manifold properties at the heart of each method

- Smoothness LEM
- Curviness HLLLE
- Local linearity LLE SSLLE



---

# 3 TECHNIQUES

---

## 3.1 UNSUPERVISED LEM

---

**Proposal:** Belkin and Niyogi (2001)

**Idea:** forcing nearby inputs to be mapped to nearby outputs

- Notion of smoothness in mapping function
- Second-order penalty on gradient

**Solution:** eigenanalysis of graph Laplacian  $L$

- Derived from weight matrix encoding nearness of inputs
- Discrete approximation of Laplace-Beltrami operator  $\mathcal{L}(f)$
- Generalized eigenvalue problem

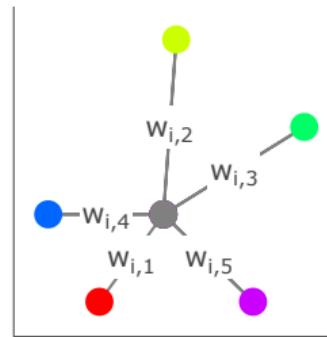
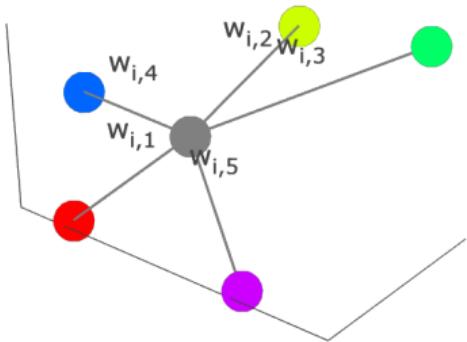
**Solution: bottom  $d + 1$  eigenvectors**

## 3.1 UNSUPERVISED LLE

**Proposal:** Roweis and Saul (2000)

**Idea:** preserving locally linear reconstructions

- Linear reconstruction of points in  $\mathbb{R}^D$  by their neighbors
- Reconstruction weights = topological properties
- Neighborhood patches invariant to dimensionality reduction



### 3.1 UNSUPERVISED LLE

---

**Reconstruction loss minimization:** finding optimal reconstruction weights

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

**Embedding loss minimization:** finding optimal embedding coordinates

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

**Eigenvalue problem:** define  $\mathbf{E} = (\mathbf{I} - \mathbf{W})^T (\mathbf{I} - \mathbf{W})$  and set  $\tilde{\mathcal{Y}} = \mathcal{Y}^T$ , such that

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

**Solution: bottom  $d + 1$  eigenvectors**

### 3.1 UNSUPERVISED HLLE

---

**Proposal:** Donoho and Grimes (2003)

**Idea:** finding a truly linear mapping while preserving local isometry

- Notion of curviness in mapping function
- Second-order penalty on Hessian
- Strong convergence guarantees but rather complex computations

**Solution:** eigenanalysis of empirical Hessian functional  $\mathcal{H}$

- Quadratic form of Hessian estimators in linear neighborhood patches
- Discrete approximation of continuous Hesssian functional  $\mathcal{H}(f)$
- Null space problem

**Solution: bottom  $d + 1$  eigenvectors + scaling**

## 3.2 SEMI-SUPERVISED SSLLE

---

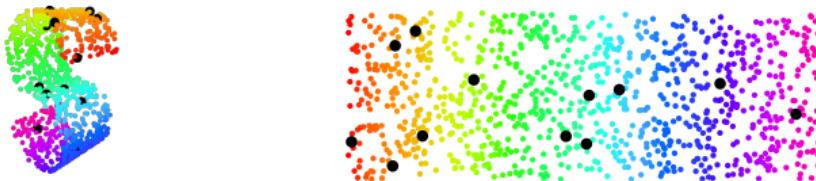
**Proposal:** Yang et al. (2006)

**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

- More active than semi-supervised learning?
- Information available or to be obtained by querying the oracle
- Maximum information at little expense ⇒ careful choice of prior points



## 3.2 SEMI-SUPERVISED SSLLE

---

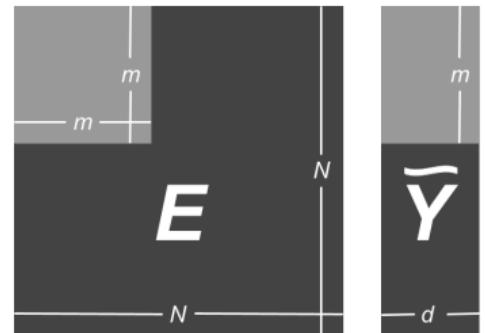
**Types of prior information:** exact vs inexact

→ Level of confidence encoded in parameter  $\beta$

**Algorithmic impact:** recall LLE eigenvalue problem

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

⇒ partitioning of  $\mathbf{E}$  and  $\tilde{\mathcal{Y}}$



## 3.2 SEMI-SUPERVISED SSLLE

---

**Modified optimization problem:** exact information

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

$$\Leftrightarrow \tilde{\mathcal{Y}}_2^T = M_{22}^{-1} M_{12} \tilde{\mathcal{Y}}_1^T \quad (5)$$

**Modified optimization problem:** inexact information

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

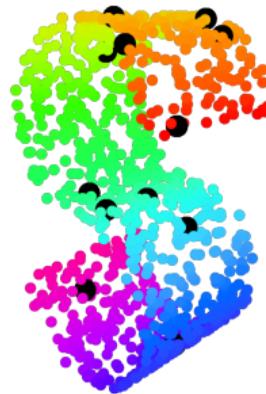
$$\Leftrightarrow \begin{bmatrix} M_{11} + \beta I & M_{12} \\ M_{21} & M_{22} \end{bmatrix} \begin{bmatrix} \tilde{\mathcal{Y}}_1^T \\ \tilde{\mathcal{Y}}_2^T \end{bmatrix} = \begin{bmatrix} \hat{\mathcal{Y}}_1^T \\ \mathbf{0} \end{bmatrix} \quad (7)$$

## 3.3 CHALLENGES CRITICAL PARAMETERS

---

**Choice of landmark points:** basically, three options

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



### 3.3 CHALLENGES CRITICAL PARAMETERS

---

**Number & location of prior points:** utility of prior knowledge ANALYSIS

- Exploration vs labeling cost

**Noise level:** quality of prior knowledge ANALYSIS

- Support vs harm through prior knowledge

**Confidence parameter:** strength of belief in prior knowledge

- Rather robust

**Further hyperparameters:** also critical in unsupervised case

- 1 Intrinsic dimensionality
- 2 Neighborhood size
- 3 Regularization constant

---

# 4 SENSITIVITY ANALYSIS

---

## 4.1 SETUP DATA

---

**Data:** two data sets,  $N = 1000$  observations each

**Swiss roll:** *the standard synthetic manifold*

- 1 Sample  $\mathbf{u}_1, \mathbf{u}_2 \sim U(0, 1)$  iid with  $|\mathbf{u}_1| = |\mathbf{u}_2| = N$
- 2 Compute  $\mathbf{t} = 1.5\pi(1 + 2\mathbf{u}_1)$  and  $\mathbf{s} = 2\mathbf{u}_2$
- 3  $\mathcal{X}_{\text{swiss}} = [\mathbf{x}_1 \quad \mathbf{x}_2 \quad \mathbf{x}_3] = [\mathbf{t} \cos \mathbf{t} \quad \mathbf{s} \quad \mathbf{t} \sin \mathbf{t}]$



**Incomplete tire:** examined in Yang et al. (2006)

- 1 Sample  $\mathbf{u}_1, \mathbf{u}_2 \sim U(0, 1)$  iid with  $|\mathbf{u}_1| = |\mathbf{u}_2| = N$
- 2 Compute  $\mathbf{t} = \frac{5\pi}{3}\mathbf{u}_1$  and  $\mathbf{s} = \frac{5\pi}{3}\mathbf{u}_2$
- 3  $\mathcal{X}_{\text{tire}} = [\mathbf{x}_1 \quad \mathbf{x}_2 \quad \mathbf{x}_3]$   
 $= [(3 + \cos \mathbf{s}) \cos \mathbf{t} \quad (3 + \cos \mathbf{s}) \sin \mathbf{t} \quad \sin \mathbf{s}]$



## 4.1 SETUP SCENARIOS

---

### Sensitivity analysis I: landmark coverage $\times$ number of landmark points

- Landmark coverage  $\in \{\text{poor, random, maximum}\}$
- Number of landmark points  $\in \{2, 4, 6, 8, 10, 12\}$

### Sensitivity analysis II: noise level $\times$ number of landmark points

- Landmark coverage kept at optimal configuration
- Simulation of inexact prior information through additive Gaussian noise
- Corruption of landmark  $\mathbf{p}$  as  $\tilde{\mathbf{p}} = \mathbf{p} + \boldsymbol{\epsilon} = (p_t, p_s) + (\epsilon_t, \epsilon_s)$   
with  $\epsilon_i \sim N(0, (\alpha \cdot s_i)^2)$  iid, 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  $\in \{2, 4, 6, 8, 10, 12\}$

## 4.1 SETUP EVALUATION

---

**Evaluation criterion:**  $\text{AUC}(R_{NX})$  (Kraemer et al. (2019), Lueks et al. (2011))

- Area under the  $R_{NX}$  curve
- Based on co-ranking matrix

**Co-ranking matrix:** comparing distance ranks in observation & embedding spaces

- Rank distance matrices  $(r)_{ij}^{\text{obs}}, (r)_{ij}^{\text{emb}} \in \mathbb{R}^{N \times N}$
- Co-ranking matrix  $\mathbf{Q} = (q)_{\ell m} \in \mathbb{R}^{N \times N}$  with  $q_{\ell m} = |\{(i, j) : r_{ij}^{\text{emb}} = \ell \wedge r_{ij}^{\text{obs}} = m\}|$
- 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 points collapsed

## 4.1 SETUP EVALUATION

---

### Co-ranking-based metrics:

- Number of points remaining in  $k$ -neighborhood after projection:

$$Q_{NX}(k) = \frac{1}{kN} \sum_{\ell=1}^k \sum_{m=1}^k q_{\ell m}$$

$$\rightarrow R_{NX}(k) = \frac{(N-1)Q_{NX}(k) - k}{N-1-k}$$

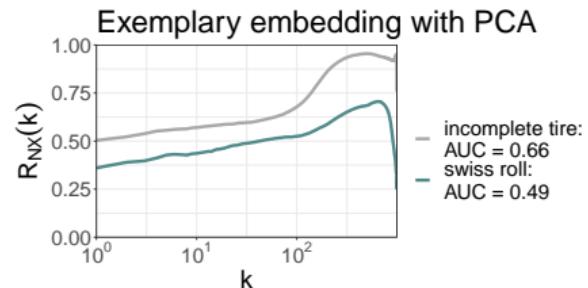
### AUC measure:

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

- Interpretation:

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

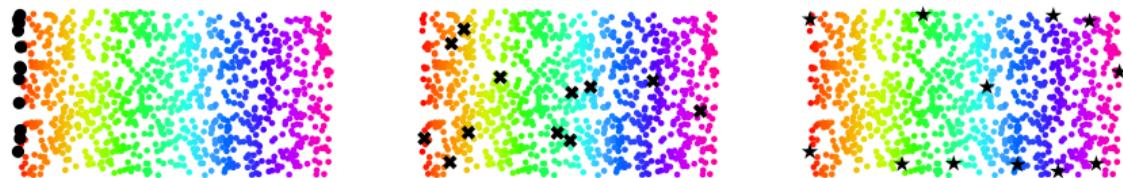
2  $\text{AUC}(R_{NX}) = 1 \Rightarrow$  optimal embedding



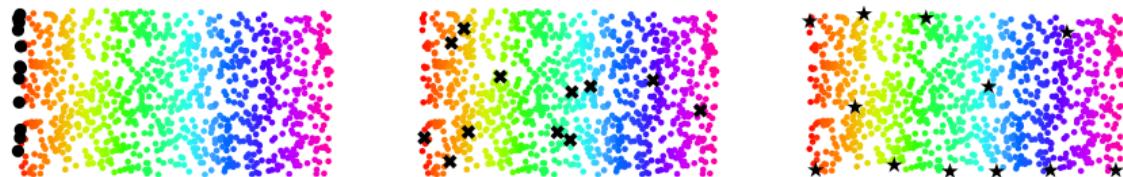
## 4.2 RESULTS SENSITIVITY ANALYSIS I

**Key variation:** poor, random, maximum coverage

Swiss roll



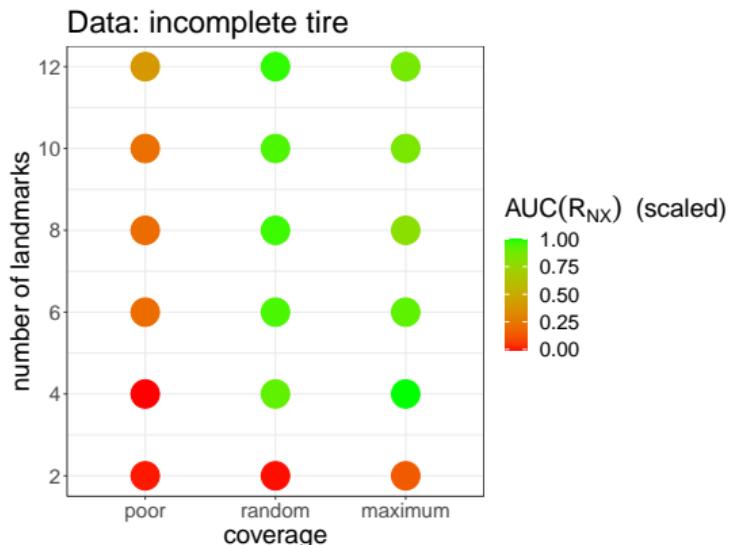
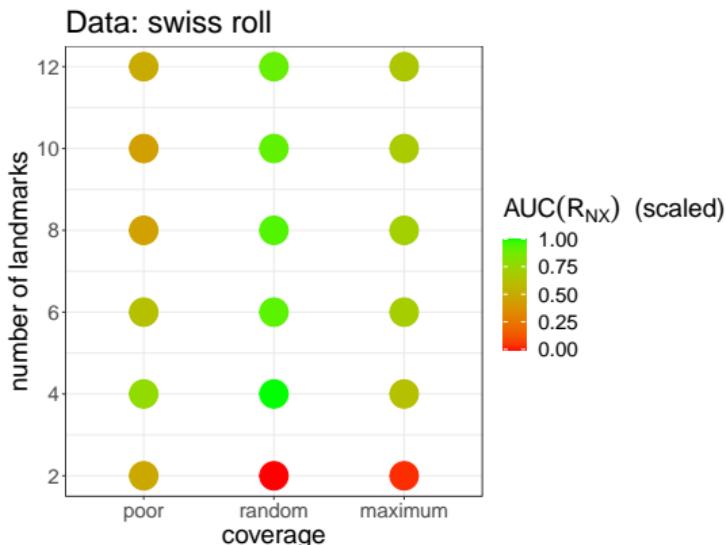
Incomplete tire



- poor coverage    × random coverage    ★ maximum coverage

## 4.2 RESULTS SENSITIVITY ANALYSIS I

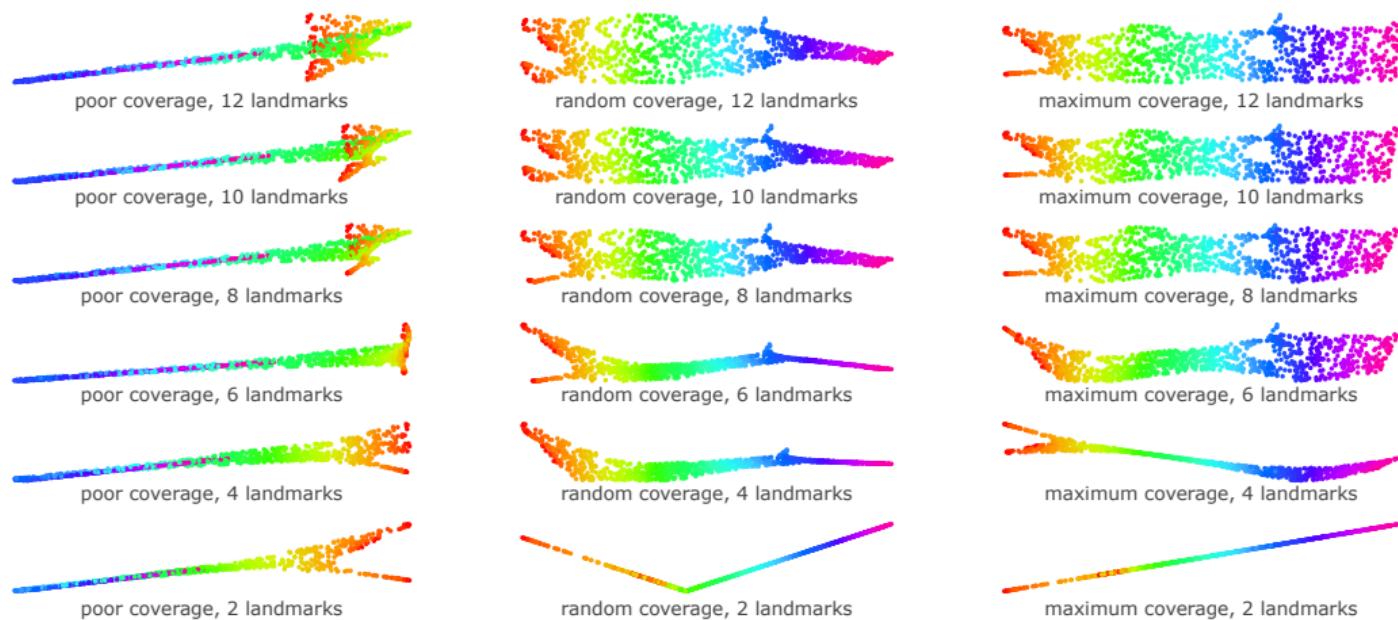
**Quantitative results:** seemingly better performance of random coverage



$AUC(R_{NX})$  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 –  $AUC(R_{NX}) \in [0.2772, 0.6231]$

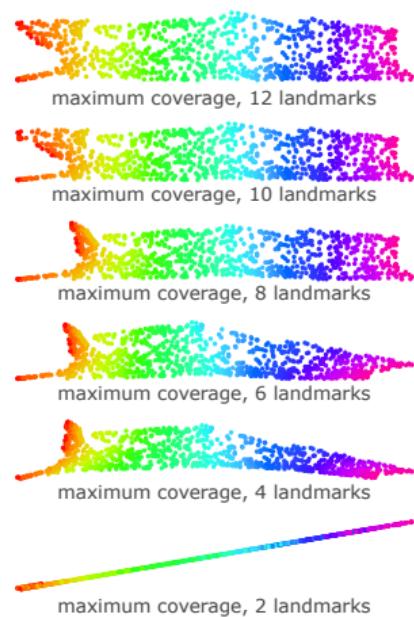
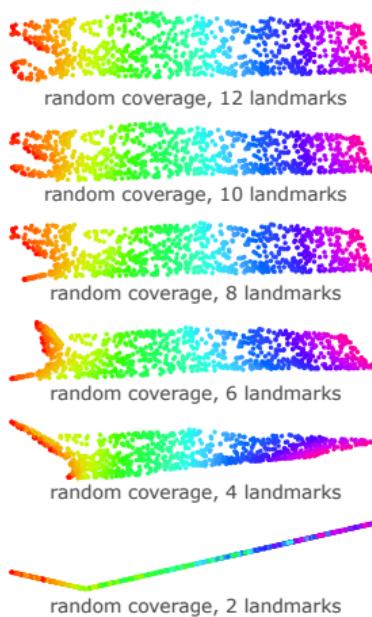
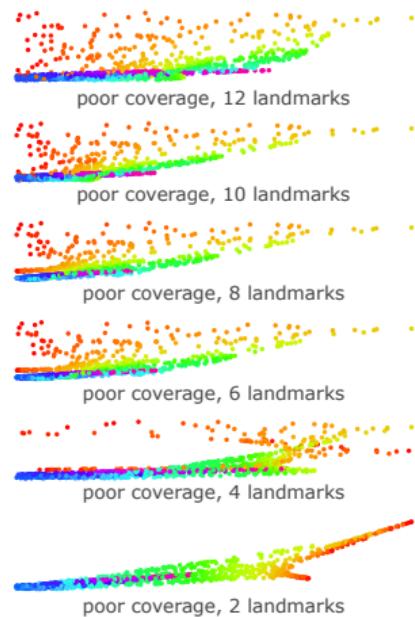
## 4.2 RESULTS SENSITIVITY ANALYSIS I

### Qualitative results: swiss roll



## 4.2 RESULTS SENSITIVITY ANALYSIS I

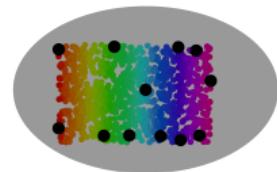
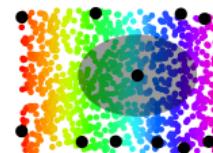
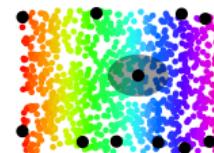
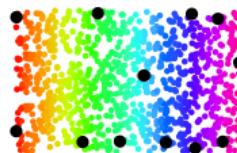
Qualitative results: incomplete tire



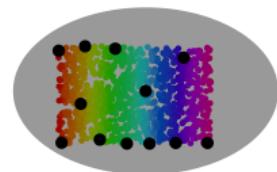
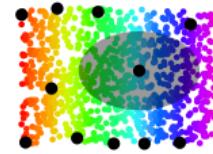
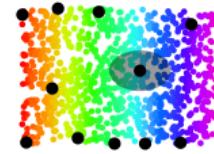
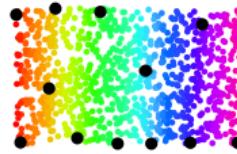
## 4.2 RESULTS SENSITIVITY ANALYSIS II

**Key variation:** noise level  $\alpha \in \{0.1, 0.5, 1.0, 3.0\}$

Swiss roll



Incomplete tire



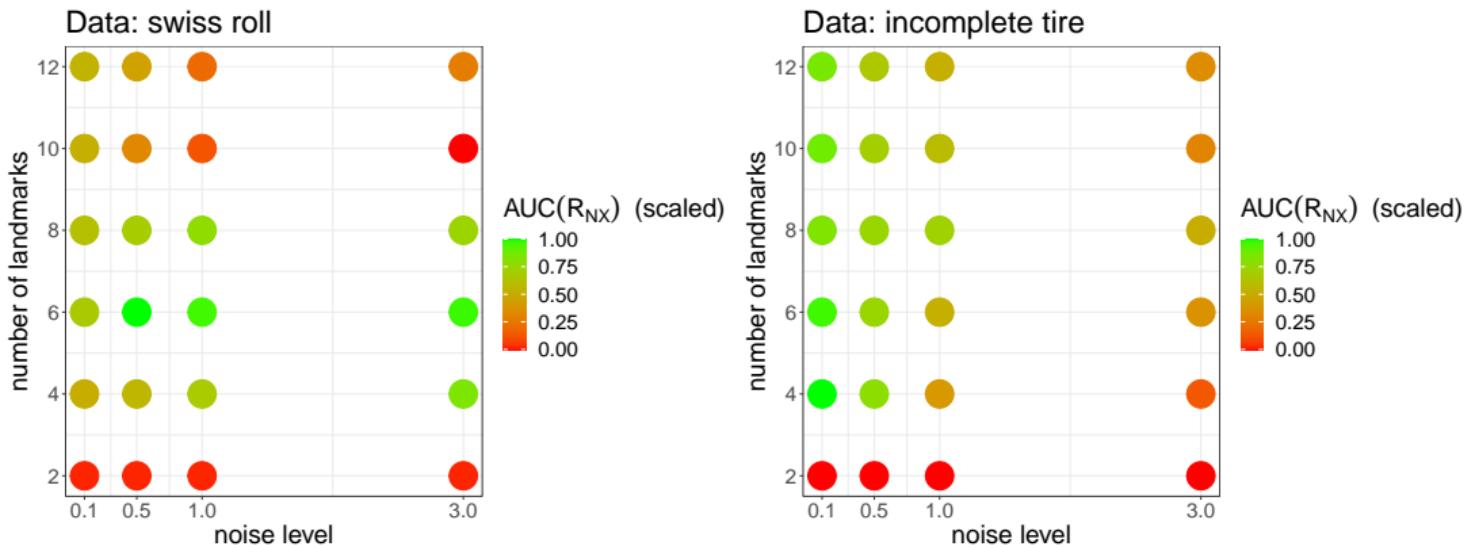
Potential displacement by random noise, exemplified at one prior point location.

Ellipses have semi-axes of length = one standard deviation of the Gaussian noise variable, i.e., noise level scaled by standard deviation  $s_i$  in  $t$  and  $s$  direction, respectively:  $\alpha \cdot s_i$  with  $i \in \{t, s\}$ .

Landmarks have been found via maximum coverage.

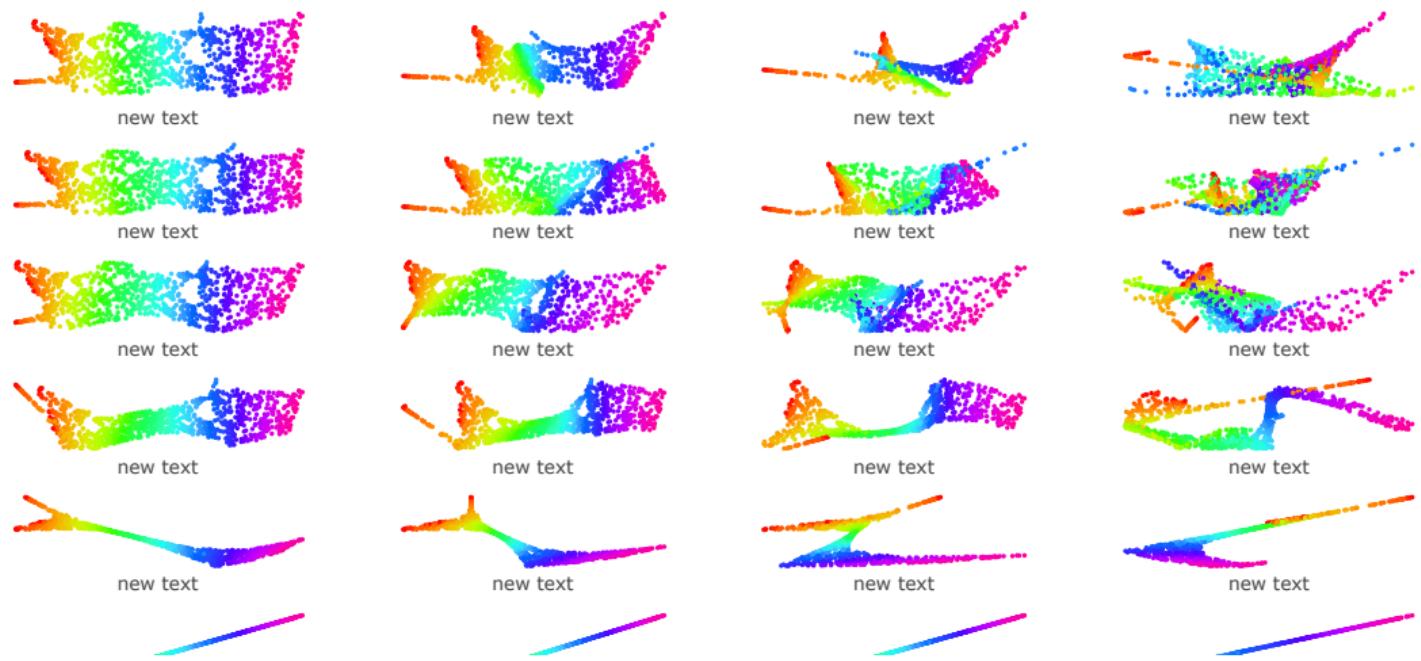
## 4.2 RESULTS SENSITIVITY ANALYSIS II

**Quantitative results:** some compensation of noise by larger number of landmarks



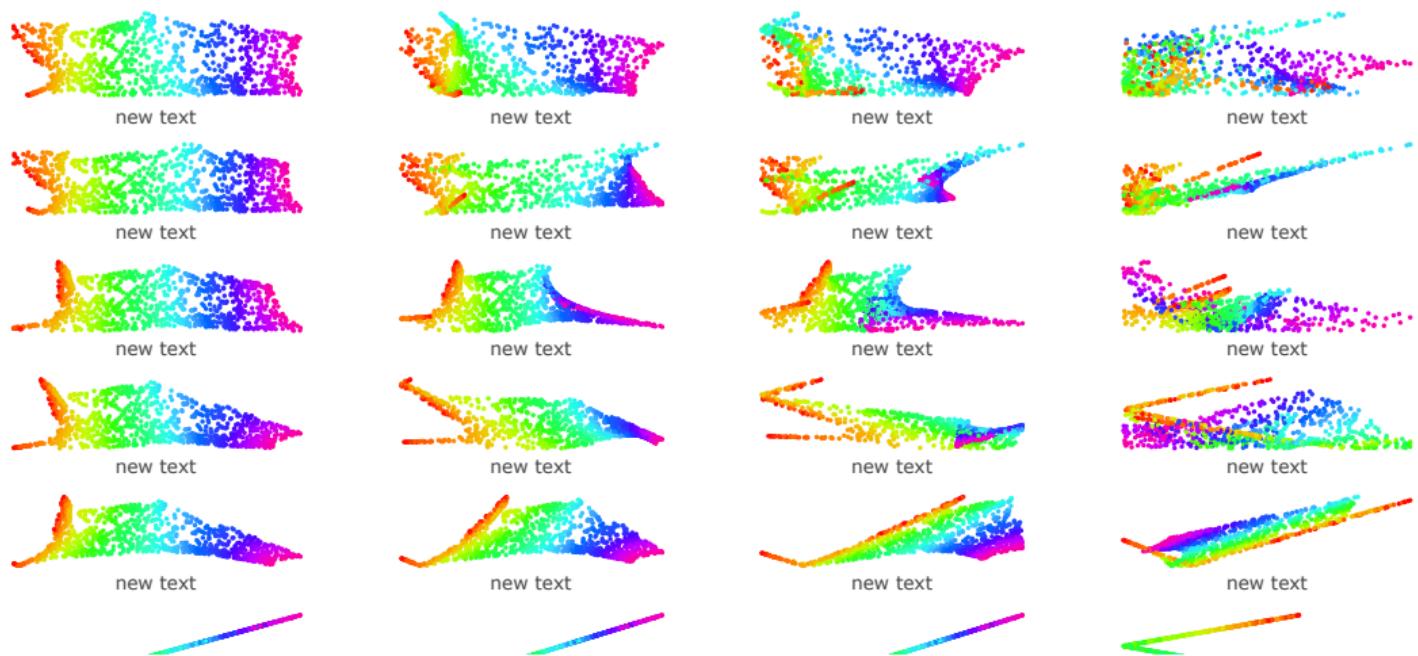
## 4.2 RESULTS SENSITIVITY ANALYSIS II

Qualitative results: swiss roll



## 4.2 RESULTS SENSITIVITY ANALYSIS II

Qualitative results: incomplete tire

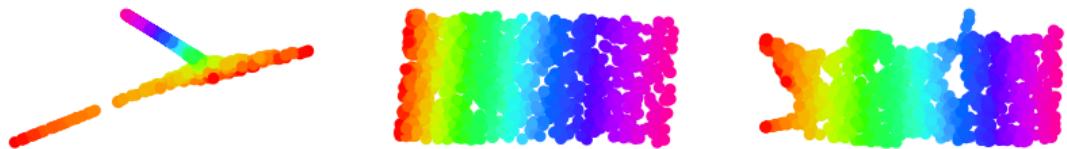


## 4.2 RESULTS CONCLUDING COMPARISON

---

### Comparison: LLE vs HLLE vs SSLLE

Swiss roll



Incomplete tire



SSLLE: own implementation (see Github), LLE & HLLE: implementation in R's `dimRed` package (Kraemer, 2019).

SSLLE with maximum coverage, 12 landmarks and exact prior information, LLE and HLLE with number of neighbors as deemed optimal by SSLLE implementation (no other hyperparameters to set).

---

# 5 DISCUSSION

---

## 5 DISCUSSION

---

	Strengths	Drawbacks
General	Reasonably simple Few parameters Tractable computations	Vital dependency on graph approximation No indication of intrinsic dimensionality Fundamental weakness in optimization problem Possibly very tight eigenvalue spectrum Meaningfulness of embedding not guaranteed Weak preservation of geometric properties
Semi-supervised extension	Simple but impactful improvement Better handling of more complex manifolds Small number of landmarks sufficient	Potentially high labeling cost Label noise problematic Prior point location crucial

**SSLLE: imperfect but potentially powerful approach to LGML**

---

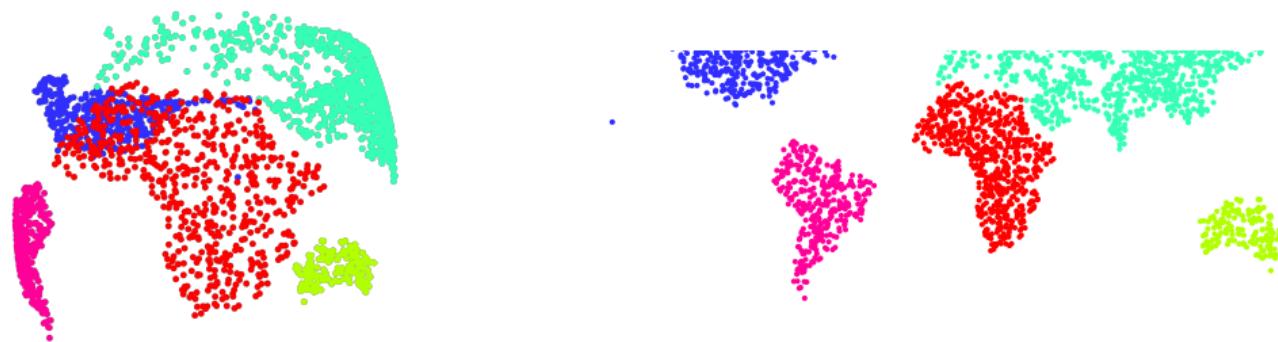
# APPENDIX

---

# APPLICATION TO WORLD DATA SET

---

**Additional:** world data in 3D and true 2D embedding



---

# 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.
- Bengio, Y., Delalleau, O., Roux, N. L., Paiement, J.-F., Vincent, P. and Ouimet, M. (2004). Learning eigenfunction links spectral embedding and kernel pca, *Neural Computation* **16**: 2197–2219.
- Cayton, L. (2005). Algorithms for manifold learning, *Technical Report CS2008-0923*, University of California, San Diego (UCSD).
- 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.
- Ghojogh, B., Ghodsi, A., Karray, F. and Crowley, M. (2020). Locally linear embedding and its variants: Tutorial and survey.
- Kraemer, G. (2019). *dimRed: A Framework for Dimensionality Reduction*. R package version 0.2.3.  
**URL:** <https://CRAN.R-project.org/package=dimRed>
- Kraemer, G., Reichstein, M. and Mahecha, M. D. (2019). dimred and coranking — unifying dimensionality reduction in r, *Technical report*.
- Lueks, W., Mokbel, B., Biehl, M. and Hammer, B. (2011). How to evaluate dimensionality reduction? improving the co-ranking matrix, *arXiv: Machine Learning* .
- Roweis, S. T. and Saul, L. K. (2000). Nonlinear dimensionality reduction by locally linear embedding, *Science* **290**(5500): 2323–2326.
- Saul, L. K., Weinberger, K. Q., Sha, F., Ham, J. and Lee, D. D. (2006). Spectral methods for dimensionality reduction, in O. Chapelle, B. Scholkopf and A. Zien (eds), *Semi-Supervised Learning*, MIT Press Scholarship Online, chapter 1.
- van der Maaten, L., Postma, E. and van den Herik, J. (2009). Dimensionality reduction: A comparative review, *Technical Report TiCC TR 2009-005*, Tilburg University.
- Yang, X., Fu, H., Zha, H. and Barlow, J. (2006). Semi-supervised nonlinear dimensionality reduction, *Proceedings of the 23rd International Conference on Machine Learning*.