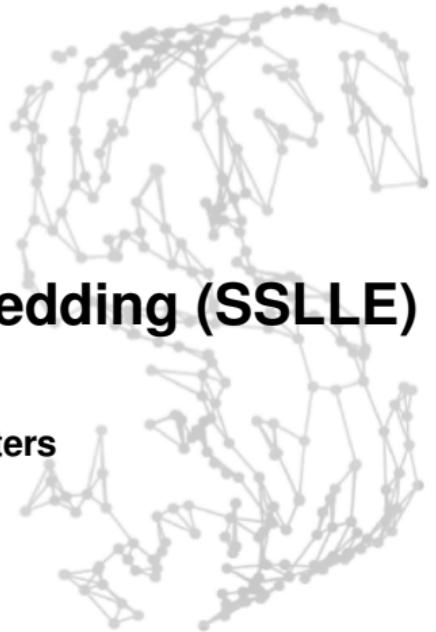


Semi-Supervised Locally Linear Embedding (SSLLE)

Application & Sensitivity Analysis of Critical Hyperparameters



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

- Model functionality
- Interpretability
- Generalization ability

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}
 - $\Rightarrow \psi : \mathcal{M} \rightarrow \mathbb{R}^d$ with $d \ll D, d \in \mathbb{N}$
 - $\Rightarrow \mathcal{X} \sim \mathcal{M} \subset \mathbb{R}^D$
- **Goal.** Find d -dimensional Euclidean representation
 - $\Rightarrow \mathcal{Y} = (\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_N)$, 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

Kernel PCA

LGML concept

LGML algorithm

Kernelization

Graph representation

Neighborhood construction

Graph functional

Matrix representation

Eigenanalysis

Eigenanalysis

Eigenanalysis

Dimensionality reduction

Dimensionality reduction

Dimensionality reduction

2 LGML CONCEPT

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

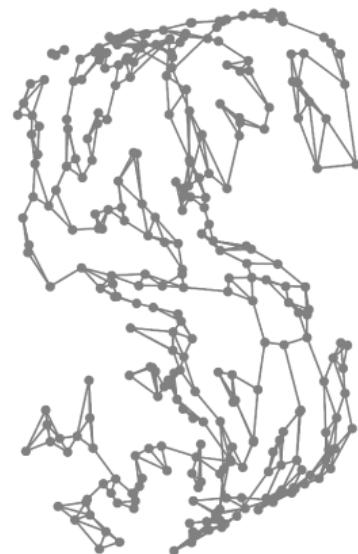
Vertices. Given by observations

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
- Local linearity LLE SSLLE
- Curviness HLLE
- ...



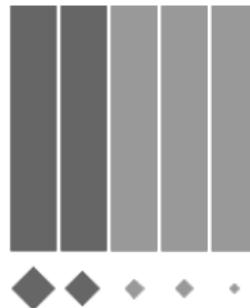
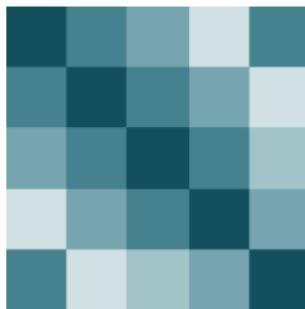
Achievements: non-linearity & locality

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

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

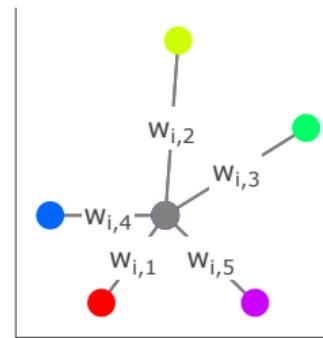
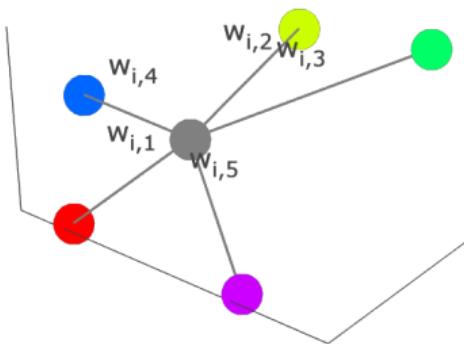
Solution: bottom $d + 1$ eigenvectors

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 $\mathbf{E} = (\mathbf{I} - \mathbf{W})^T (\mathbf{I} - \mathbf{W})$, such that

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

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}

- Derived as a 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

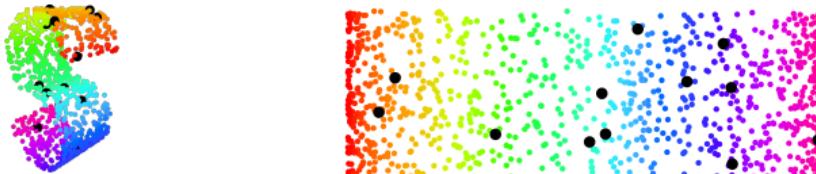
Proposal. Yang et al. (2006)

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?
- Setting. Information available or to be obtained by querying the oracle
- Goal. Maximum information at little expense \Rightarrow careful choice of prior points



3.2 SEMI-SUPERVISED SSLLE

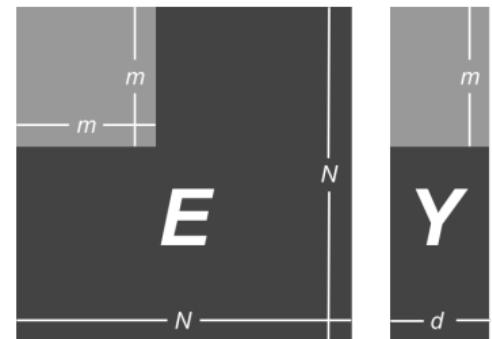
Types of prior information. Exact vs inexact

→ Level of confidence encoded in parameter β

Algorithmic impact. Recall LLE eigenvalue problem

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

⇒ Partitioning of \mathbf{E} and \mathcal{Y}



3.2 SEMI-SUPERVISED SSLLE

Modified optimization problem. Exact information

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

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

Modified optimization problem. Inexact information

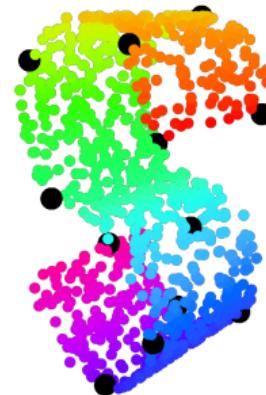
$$\min_{\mathcal{Y}_1, \mathcal{Y}_2} \begin{bmatrix} \mathcal{Y}_1 & \mathcal{Y}_2 \end{bmatrix} \begin{bmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{bmatrix} \begin{bmatrix} \mathcal{Y}_1^T \\ \mathcal{Y}_2^T \end{bmatrix} + \beta \left\| \mathcal{Y}_1^T - \hat{\mathcal{Y}}_1^T \right\|_F^2 \quad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

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

3.2 SEMI-SUPERVISED SSLLE

Choice of landmark points. Basically, three options

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



3.3 CHALLENGES CRITICAL PARAMETERS

Intrinsic dimensionality. True sources of variability

- Considered known with availability of prior information

Neighborhood size. Global vs local structure

- Tunable (expensive)

Regularization constant. Singularity for $D < k$

- Heuristics

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

- Exploration vs labeling cost

Noise level. Quality of prior knowledge ANALYSIS

- How exact must prior information be?

Confidence parameter. Strength of belief in prior knowledge

- Rather robust

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 $t = 1.5\pi(1 + 2\mathbf{u}_1)$
- 3 Set $\mathcal{X}_{\text{swiss}} = [\mathbf{t} \cos \mathbf{t} \quad 2\mathbf{u}_2 \quad \mathbf{t} \sin \mathbf{t}]$



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 Set $\mathcal{X}_{\text{tire}} = [(3 + \cos \mathbf{s}) \cos \mathbf{t} \quad (3 + \cos \mathbf{s}) \sin \mathbf{t} \quad \sin \mathbf{s}]$



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 $\in \{2, 4, 6, 8, 10, 12\}$
- ⇒ Best case: maximum coverage & 12 landmarks

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

- Simulation of inexact prior information through perturbation with Gaussian noise
 - Noise level $\in \{0.1, 0.5, 1, 3, 5\} \Rightarrow$ standard deviation
 - Number of landmark points $\in \{2, 4, 6, 8, 10, 12\}$
- ⇒ Best case: noise level 0.1 & 12 landmarks

4.1 SETUP EVALUATION

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

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

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

- Rank distance matrices. $(r)_{ij}^{\text{obs}}$ and $(r)_{ij}^{\text{emb}}$
- Co-ranking matrix. $\mathbf{Q} = (q)_{\ell m}$ 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. Comparing distance ranks in observation & embedding spaces

- Number of points belonging to k -neighborhood in both spaces.

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

- Adjustment for random embeddings and normalization.

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

AUC measure. Parameter-free

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

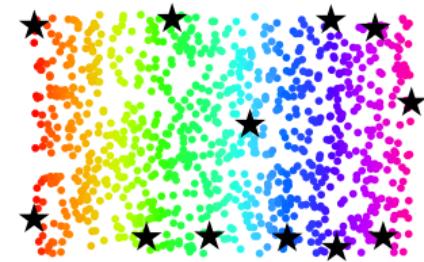
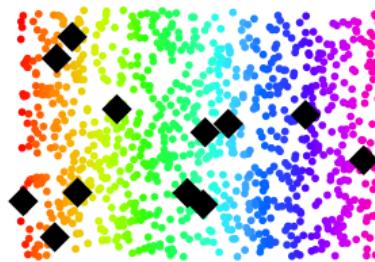
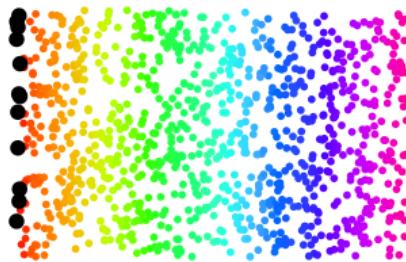
- Interpretation.

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

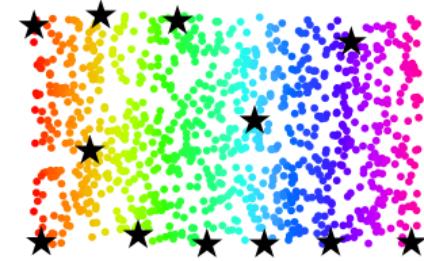
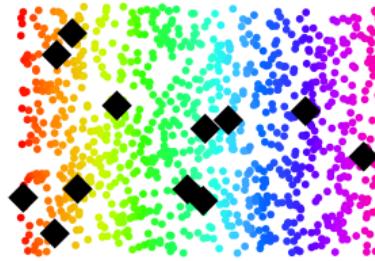
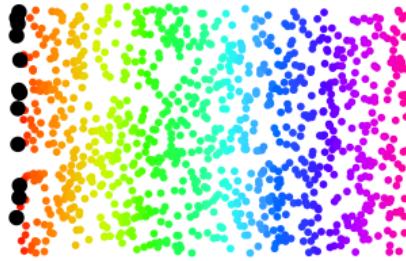
2 $\text{AUC}(R_{NX}) = 1 \Rightarrow$ optimal embedding

4.2 RESULTS SENSITIVITY ANALYSIS I

Key variation. Swiss roll

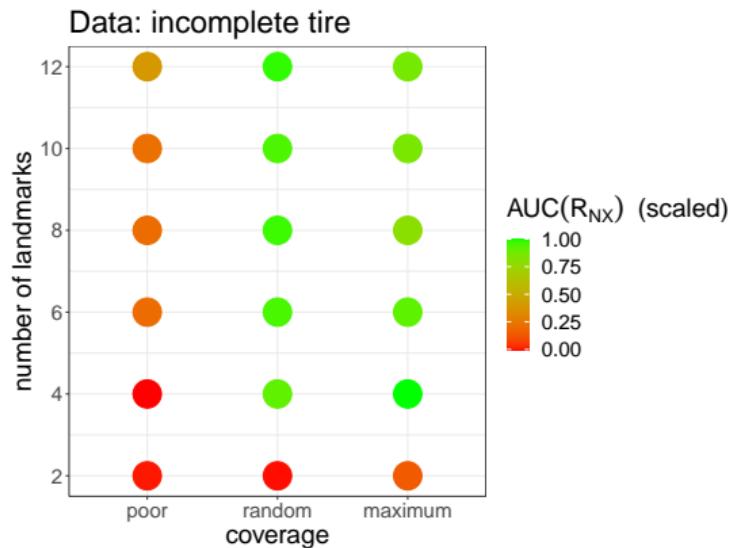
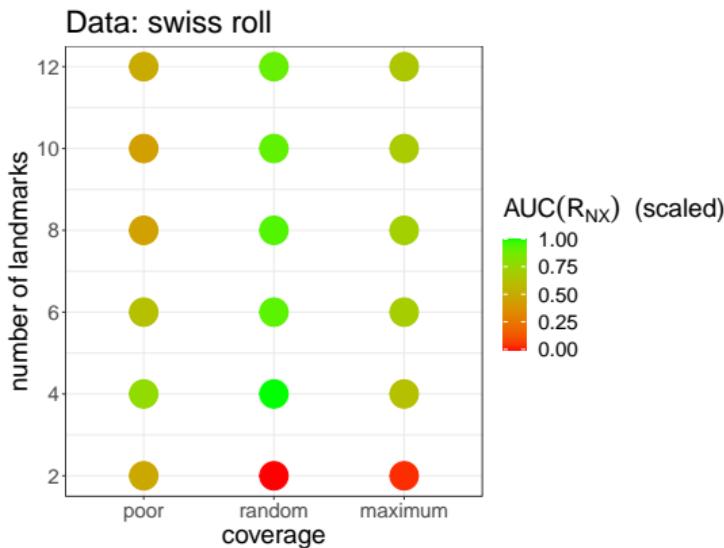


Key variation. Incomplete tire



4.2 RESULTS SENSITIVITY ANALYSIS I

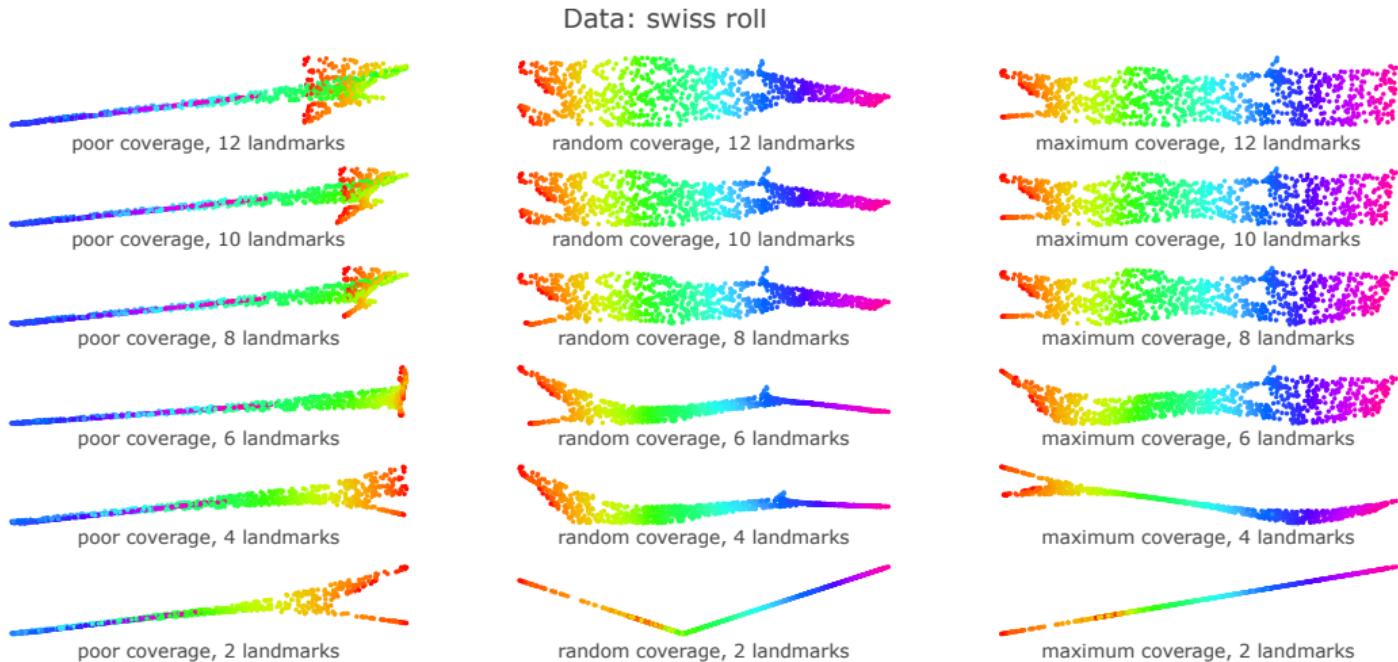
Quantitative results. Seemingly better performance of random coverage



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}) $\in [0.2772, 0.6231]$

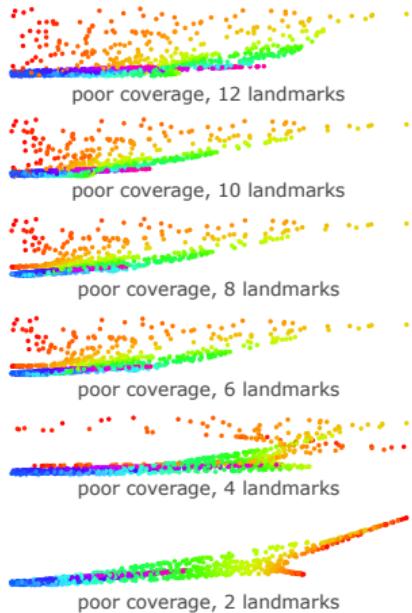
4.2 RESULTS SENSITIVITY ANALYSIS I

Qualitative results. Somewhat mixed picture

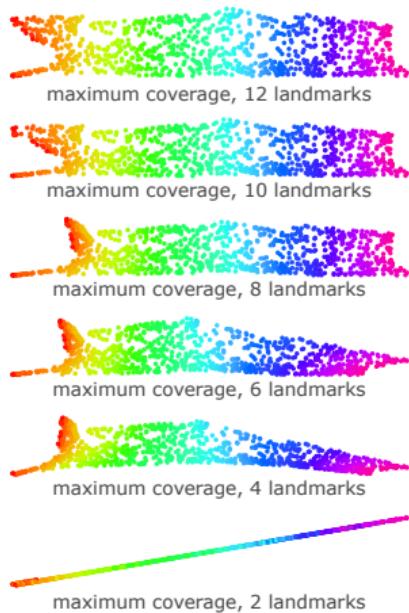
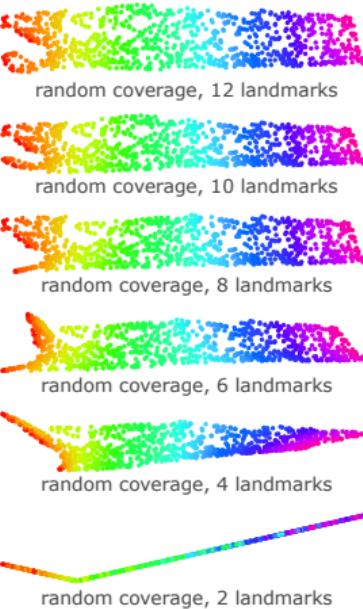


4.2 RESULTS SENSITIVITY ANALYSIS I

Qualitative results. Somewhat mixed picture

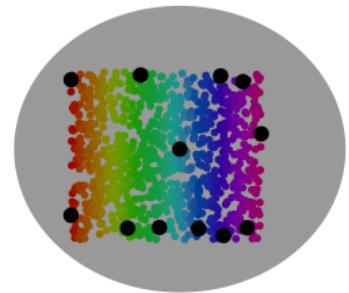
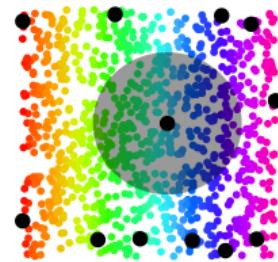
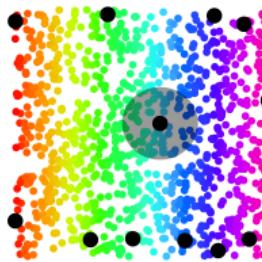
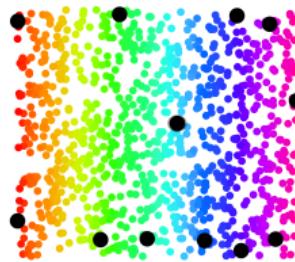


Data: incomplete tire

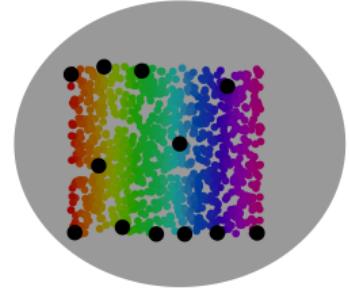
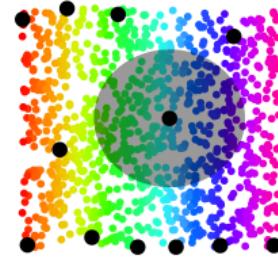
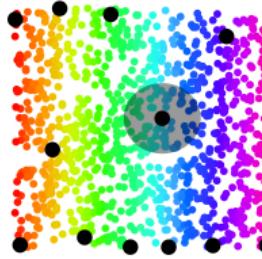
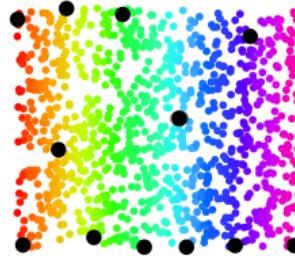


4.2 RESULTS SENSITIVITY ANALYSIS II

Key variation. Swiss roll

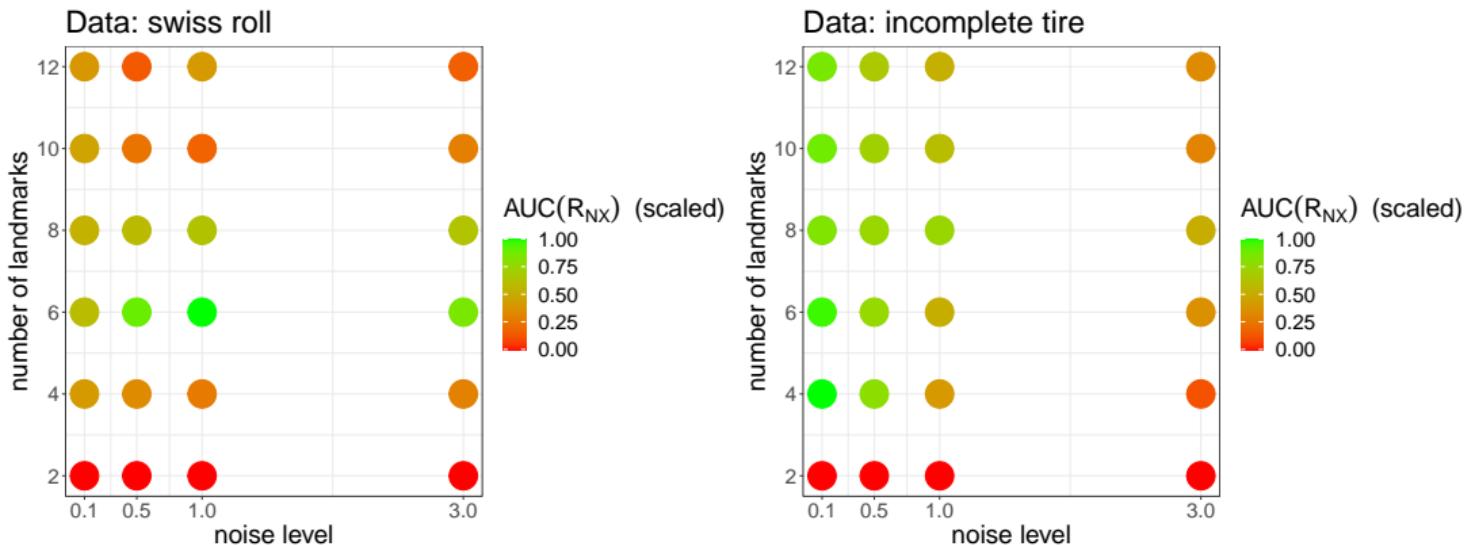


Key variation. Incomplete tire



4.2 RESULTS SENSITIVITY ANALYSIS II

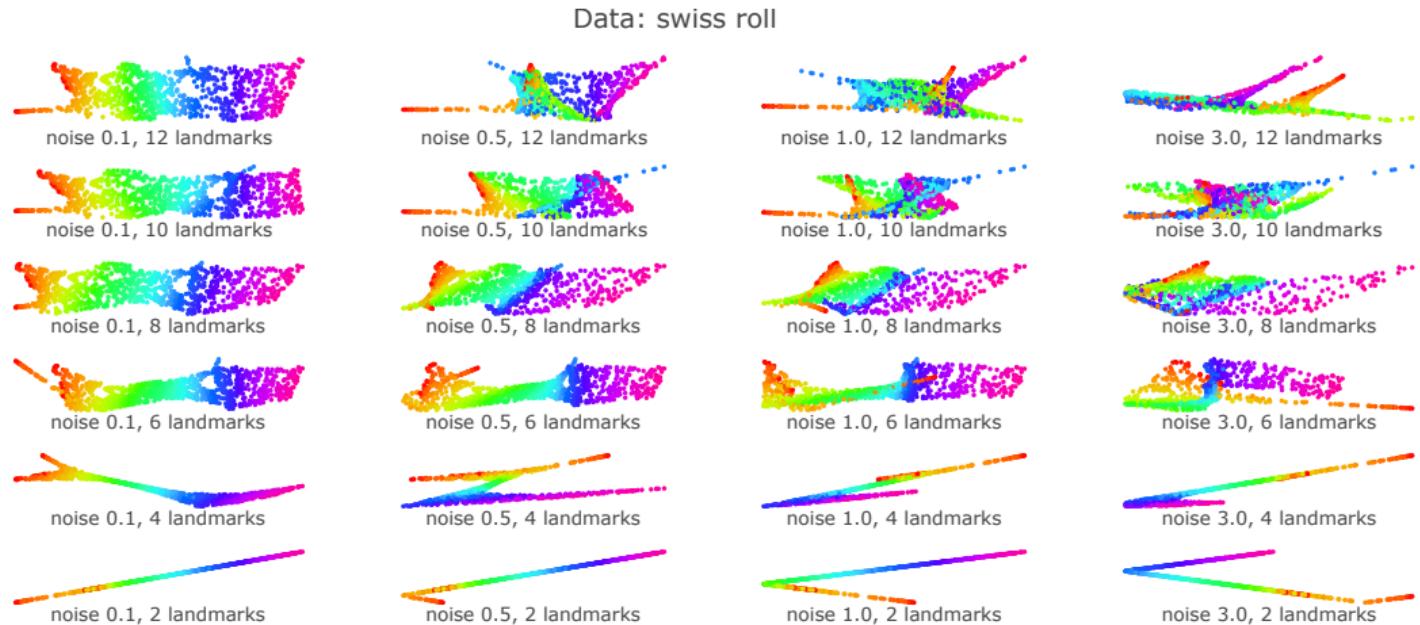
Quantitative results. Some compensation of noise by larger number of landmarks



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}) $\in [0.2720, 0.4167]$, incomplete tire: AUC(R_{NX}) $\in [0.3171, 0.6172]$

4.2 RESULTS SENSITIVITY ANALYSIS II

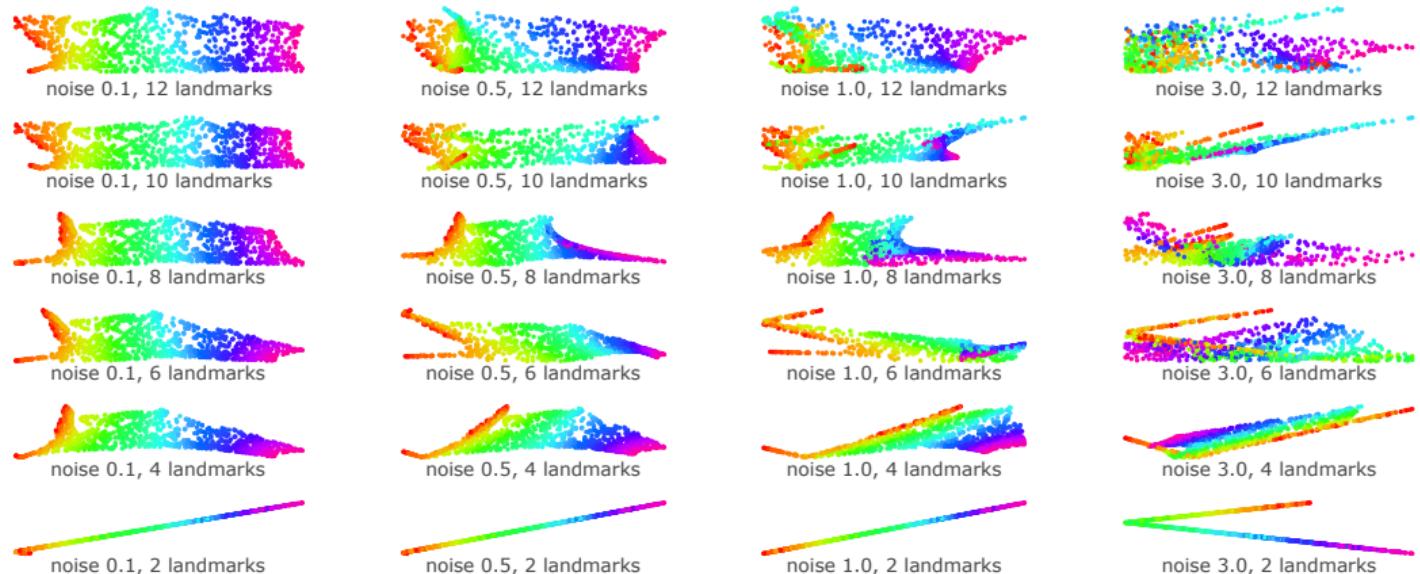
Qualitative results. Impact different across data sets



4.2 RESULTS SENSITIVITY ANALYSIS II

Qualitative results. Impact different across data sets

Data: incomplete tire



5 DISCUSSION

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*.