Semi-Supervised Locally Linear Embedding (SSLLE)

Application & Sensitivity Analysis of Critical Hyperparameters

0 AGENDA

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

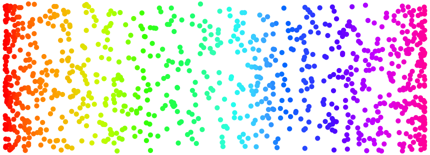
How to find a meaningful, structure-preserving embedding?

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 $d \ll D, d \in \mathbb{N}$
$\Rightarrow \mathcal{X} \sim \mathcal{M} \subset \mathbb{R}^D$

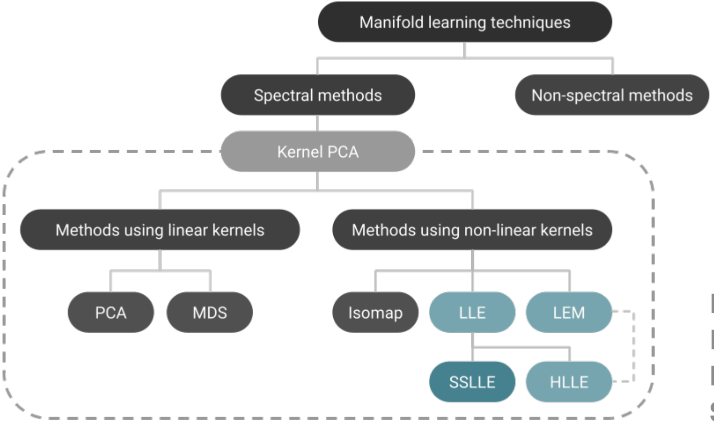
→ **Goal.** Find d-dimensional Euclidean representation
$\Rightarrow \mathcal{Y} = (\boldsymbol{y}_1, \boldsymbol{y}_2, ..., \boldsymbol{y}_N)$, with $\boldsymbol{y}_i = \psi(\boldsymbol{x}_i) \in \mathbb{R}^d \; \forall i \in \{1, 2, ..., N\}$.



2 LGML

TAXONOMY

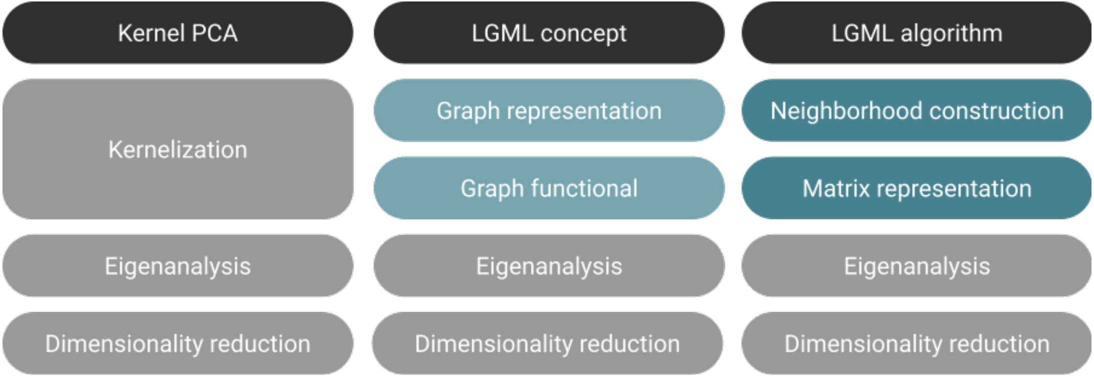
Landscape. Various approaches, many of which may be translated into one another



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



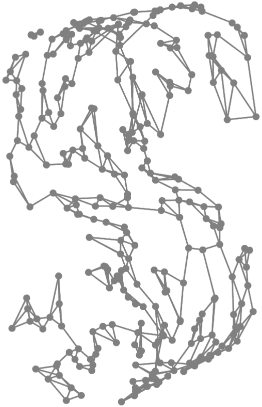
CONCEPT

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

$\rightarrow$ ...

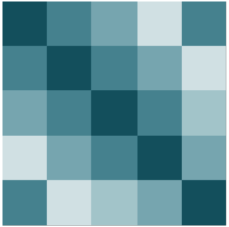
Achievements: non-linearity & locality

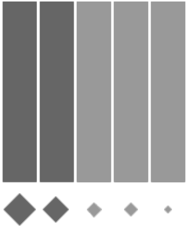
CONCEPT

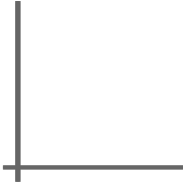
Eigenanalysis. Finding axes of variability in intrinsic manifold structure

- → 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

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

Graph Laplacian. Discrete approximation of Laplace-Beltrami operator

$\rightarrow$ Weight matrix. $\boldsymbol{W} = (w)_{ij} \in \mathbb{R}^{N \times N}$, where $w_{ij} = w_{ij}(\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2)$

$\rightarrow$ Graph Laplacian. $\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{W} \in \mathbb{R}^{N \times N}$, $\boldsymbol{D} = diag(\sum_j w_{ij}) \in \mathbb{R}^{N \times N}$

Generalized eigenvalue problem.

$$\min_{\mathcal{Y}} trace(\mathcal{Y}^T \boldsymbol{L} \mathcal{Y}), \quad \text{s.t. } \mathcal{Y}^T \boldsymbol{D} \mathcal{Y} = \boldsymbol{I} \tag{1}$$

Solution: bottom $d + 1$ eigenvectors

LLE

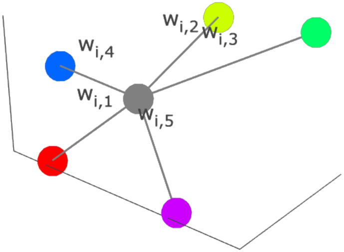
Proposal. Roweis and Saul (2000)

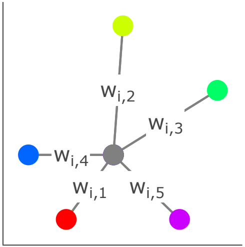
Idea. Preserving locally linear reconstructions

→ 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

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

Embedding loss minimization. Finding optimal embedding coordinates

$$\min_{\mathcal{Y}} \Phi(\mathcal{Y}) = \min_{\mathcal{Y}} \sum_i \left\| \boldsymbol{y}_i - \sum_j w_{ij} \boldsymbol{y}_j \right\|^2, \quad \text{s.t. } \frac{1}{N} \sum_i \boldsymbol{y}_i \boldsymbol{y}_i^T = \boldsymbol{I}, \quad \sum_i \boldsymbol{y}_i = \boldsymbol{0} \quad \forall i \in \{1, 2, ..., N\} \tag{3}$$

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

$$\min_{\mathcal{Y}} \mathit{trace}(\mathcal{Y}^T \boldsymbol{E} \mathcal{Y}), \quad \text{s.t. } \frac{1}{N} \mathcal{Y}^T \mathcal{Y} = \boldsymbol{I}, \quad \mathcal{Y}^T \boldsymbol{1} = \boldsymbol{0}. \tag{4}$$

Solution: bottom $d + 1$ eigenvectors

3.1 UNSUPERVISED **HLLE**

Proposal. Donoho and Grimes (2003)

Idea. Finding a truly locally 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

Hessian functional. Measuring average curviness over $\mathcal{M}$

$\rightarrow$ Continuous functional. $\mathscr{H}(f) = \int_{\mathcal{M}} \left\| \boldsymbol{H}_f^{\text{loc}}(\boldsymbol{p}) \right\|_F^2 d\boldsymbol{p}$

$\rightarrow$ Hessian estimators $\boldsymbol{H}_\ell$ derived from locally linear neighborhood patches

$\rightarrow$ Empirical approximator. $\mathcal{H}_{ij} = \sum_\ell \sum_m (\boldsymbol{H}_\ell)_{m,i} (\boldsymbol{H}_\ell)_{m,j}$

$\rightarrow$ Finding null space of $\mathcal{H}$

Solution: bottom $d + 1$ eigenvectors + scaling

SSLLE

Proposal. Yang et al. (2006)

Problem. Embedding found by unsupervised methods not always meaningful

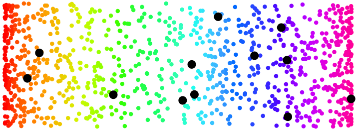
Idea. Improving LLE by use of prior knowledge

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

$\rightarrow$ 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



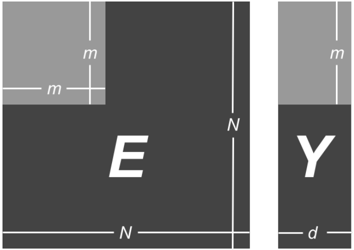
Types of prior information. Exact vs inexact

$\rightarrow$ 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

$$\min_{\mathcal{Y}_2} \begin{bmatrix} \mathcal{Y}_1 & \mathcal{Y}_2 \end{bmatrix} \begin{bmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{bmatrix} \begin{bmatrix} \mathcal{Y}_1^T \\ \mathcal{Y}_2^T \end{bmatrix} \tag{5}$$

$$\Leftrightarrow \mathcal{Y}_2^T = M_{22}^{-1} M_{12} \mathcal{Y}_1^T \tag{6}$$

Modified optimization problem. Inexact information

$$\min_{\mathcal{Y}} \begin{bmatrix} \mathcal{Y}_1 & \mathcal{Y}_2 \end{bmatrix} \begin{bmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{bmatrix} \begin{bmatrix} \mathcal{Y}_1^T \\ \mathcal{Y}_2^T \end{bmatrix} + \beta \left\| \mathcal{Y}_1^T - \hat{\mathcal{Y}}_1^T \right\|_F^2 \tag{7}$$

$$\Leftrightarrow \begin{bmatrix} M_{11} + \beta \boldsymbol{I} & M_{12} \\ M_{21} & M_{22} \end{bmatrix} \begin{bmatrix} \mathcal{Y}_1^T \\ \mathcal{Y}_2^T \end{bmatrix} = \begin{bmatrix} \hat{\mathcal{Y}}_1^T \\ \mathbf{0} \end{bmatrix} \tag{8}$$

SSLLE

Choice of landmark points. Basically, three options

- $\rightarrow$ Pre-existing prior information $\Rightarrow$ worst case: poor coverage
- $\rightarrow$ Random sampling
- $\rightarrow$ Maximum coverage

Maximum coverage. Points scattered across manifold surface

- $\rightarrow$ Goodness of solution depending on condition number $\kappa(M_{22})$
- $\rightarrow$ $\kappa(M_{22})$ minimal at maximization of minimum pairwise distances between prior points

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

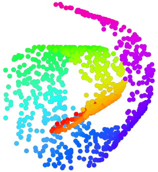
$\rightarrow$ Rather robust

4 SENSITIVITY ANALYSIS

DATA

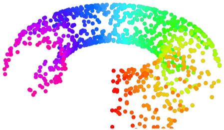
Data. Two data sets, $N = 1000$ observations each

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 $\boldsymbol{t} = 1.5\pi(1 + 2\boldsymbol{u}_1)$
3. Set $\mathcal{X}_{\text{swiss}} = \begin{bmatrix} \boldsymbol{t} \cos \boldsymbol{t} & 21\boldsymbol{u}_2 & \boldsymbol{t} \sin \boldsymbol{t} \end{bmatrix}$

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. Set $\mathcal{X}_{\text{tire}} = \begin{bmatrix} (3 + \cos \boldsymbol{s}) \cos \boldsymbol{t} & (3 + \cos \boldsymbol{s}) \sin \boldsymbol{t} & \sin \boldsymbol{s} \end{bmatrix}$

SCENARIOS

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

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

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

$\Rightarrow$ Best case: maximum coverage & 12 landmarks

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

$\rightarrow$ Simulation of inexact prior information through perturbation with Gaussian noise

$\rightarrow$ Noise level $\in \{0.1, 0.5, 1, 3, 5\}$ $\Rightarrow$ standard deviation

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

$\Rightarrow$ Best case: noise level 0.1 & 12 landmarks

4.1 SETUP EVALUATION

Evaluation criterion I. Residual variance

$\rightarrow$ foo

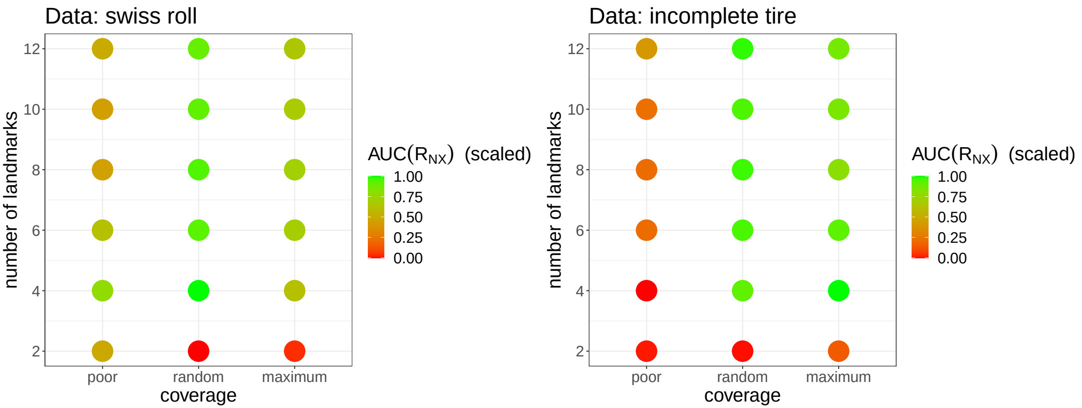
Evaluation criterion II. Area under the xy curve

$\rightarrow$ foo

$\Rightarrow$ Both semi-reliable
$\Rightarrow$ Additional visual inspection

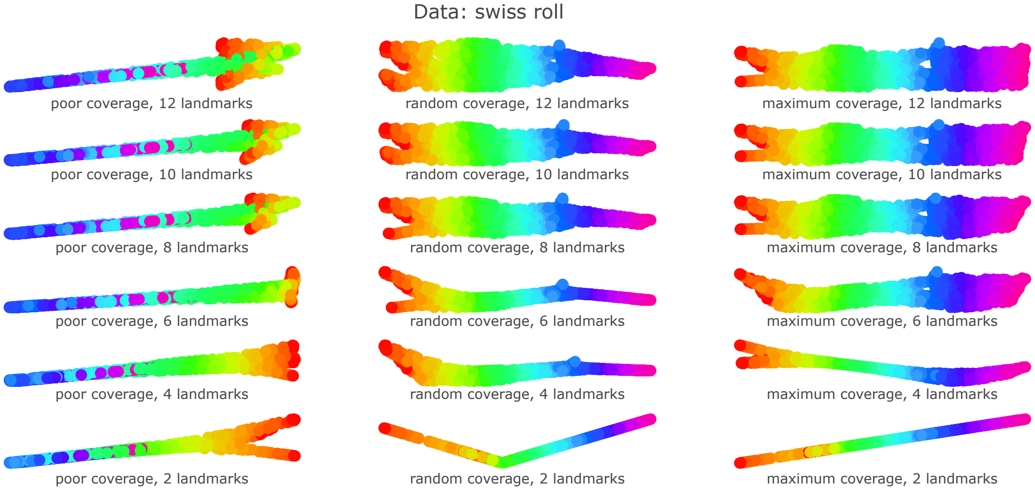
SENSITIVITY ANALYSIS I

Quantitative results. Seemingly better performance of random coverage



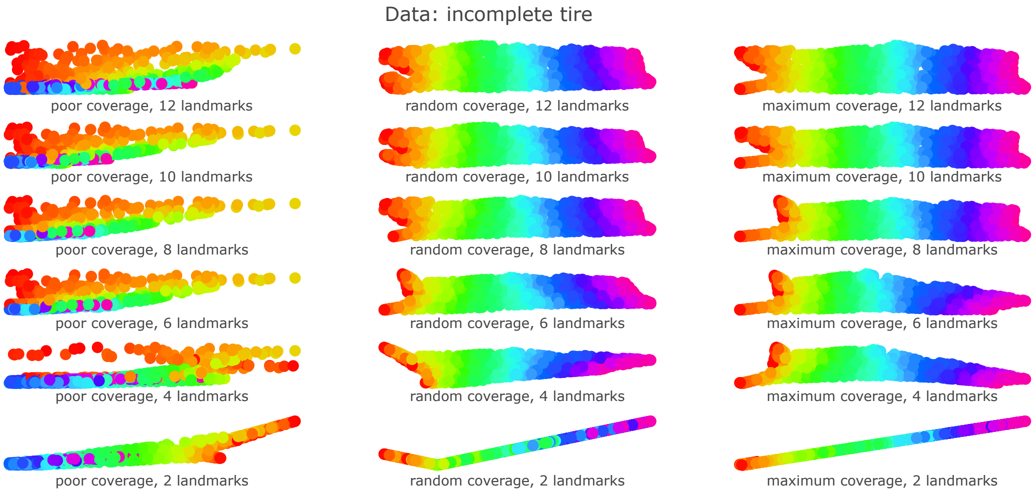
SENSITIVITY ANALYSIS I

Qualitative results. Somewhat mixed picture



SENSITIVITY ANALYSIS I

Qualitative results. Somewhat mixed picture



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.

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*.