Scalable Nonlinear Spectral Dimensionality Reduction Methods For Streaming Data

Suchismit Mahapatra

Department of Computer Science

Dissertation Defense

University at Buffalo
The State University of New York

Manifold Learning In Streams

Motivation

- Understanding the structure of multidimensional patterns is of primary importance.
- Processing data streams, potentially infinite requires adequate summarization which can handle inherent constraints and approximate characteristics well.

Manifold Learning In Streams

Challenges Involved

- Curse of dimensionality combined with lack of scalability of algorithms makes data analysis difficult/inadequate.
- Cannot use entire streams as training data motivates Out-of-Sample Extension (OOSE) techniques.
- Need to formalize "collective error" in NLSDR methods and strategies to quantify it.
- Dealing with intersecting manifolds.
- Need to handle concept drift i.e. changes in stream properties.

Thesis

Thesis Contributions

- Formulate a generalized Out-of-Sample Extension framework for streaming NLSDR.
- Provide algorithms which are specific instantiations of the above generalized framework, for Isomap and LLE.
- Provide theoretical proofs which support the basic operating principles of framework.

Additionally, provide a novel Tangent Manifold clustering strategy to deal with intersecting manifolds.

Thesis

Thesis Contributions In Detail

In particular,

- Chapter 3: S-Isomap [1], which can compute low-dimensional embeddings cheaply without affecting the quality significantly.
- Chapter 4: S-Isomap++ [2], which can deal with multimodal and/or unevenly sampled distributions.
- Chapter 5: GP-Isomap [3], which is able to detect concept drift and can embed streaming samples effectively.
- Chapter 6: A Generalized Out-of-Sample Extension Framework for streaming NLSDR [4] and subsequently discusses Streaming-LLE.

Thesis

Publications

1. "Error metrics for learning reliable manifolds from streaming data.", Proceedings of the 2017 SDM. SIAM, 2017.
2. "S-Isomap++: Multi Manifold Learning from Streaming Data.", 2017 IEEE International Conference on Big Data. IEEE, 2017.
3. "Learning manifolds from non-stationary streaming data.", arXiv preprint arXiv:1804.08833, 2018. (under submission at ECML-PKDD 2018)
4. "A Generalized Out-of-Sample Extension Framework for streaming NLSDR" (under preparation for TKDE 2018)

Thesis

Algorithmic Contributions

	Isomap	S-Isomap	S-Isomap++	GP-Isomap
Scalable Stream Processing	✗	✓	✓	✓
Handling Multiple/ Intersecting Manifolds	✗	✗	✓	✓
Handling Non-stationary Streams	✗	✗	✗	✓

Additionally,

- Formulate techniques for generalized OOSE framework for streaming NLSDR.
- Propose streaming extensions for Local Linear Embedding (LLE).

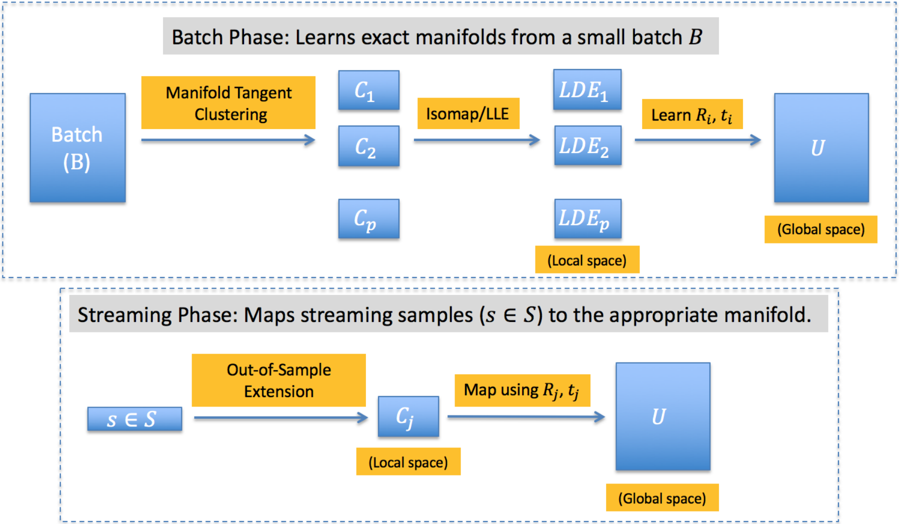
Thesis

Theoretical Contributions

- Prove that a small initial batch is sufficient for reliable learning of manifolds.
- Show equivalence between GP-Isomap prediction and S-Isomap prediction.

Methodology

A Generalized Framework For Multi-Manifold Learning



Methodology

A Generalized Framework For Multi-Manifold Learning

Input: Batch **B**, Stream **S**; Parameters ϵ, **k**, **l**, λ
Output: LDE $\mathcal{Y}_{\mathcal{S}}$

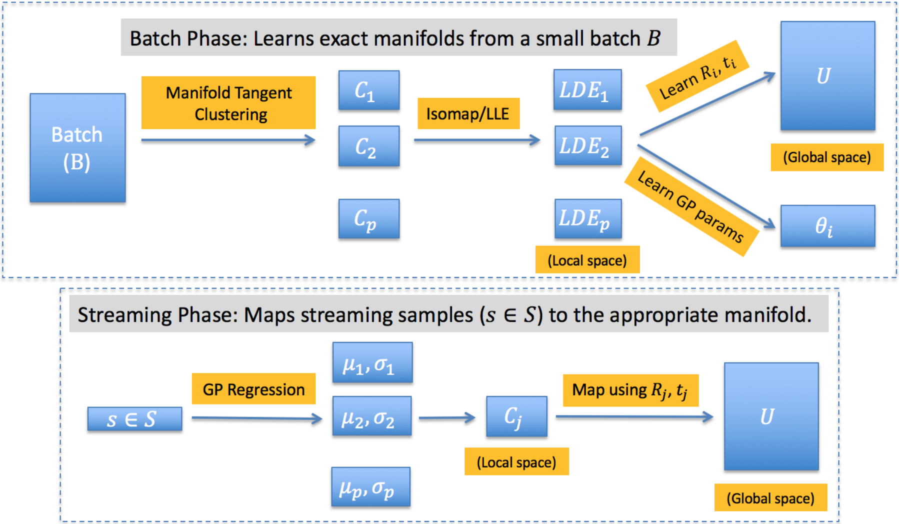
```
Partition B into clusters C_{i=1,2...p}.
Compute low dim. emb. ∀C_{i=1,2...p} using A.
Determine support ξ_s using C_{i=1,2...p}.
Compute {R_i, t_i}_{i=1,2...p} which maps M_i → U.

For each s ∈ S
Using OOS_A, project s to M_i ∀i = 1,2...p.
Using {R_i, t_i}_{i=1,2...p}, map s → U.
Embed s in M_j where j ← argmin_i |U_i(s) − μ(C_i, R_i, t_i)|.
Y_S ← Y_S ∪ y_s
```

Partition $\mathbf{B}$ into clusters $\mathbf{C}_{i=1,2\ldots p}$.
Compute low dim. emb. $\forall \mathbf{C}_{i=1,2\ldots p}$ using $\mathcal{A}$.
Determine support ξ_s using $\mathbf{C}_{i=1,2\ldots p}$.
Compute $\{\mathbf{R}_i, \mathbf{t}_i\}_{i=1,2\ldots p}$ which maps $\mathcal{M}_i \rightarrow \mathcal{U}$.

For each $\mathbf{s} \in \mathbf{S}$
Using $\mathcal{OOS}_{\mathcal{A}}$, project $\mathbf{s}$ to $\mathcal{M}_i$ $\forall i = 1, 2 \ldots p$.
Using $\{\mathbf{R}_i, \mathbf{t}_i\}_{i=1,2\ldots p}$, map $\mathbf{s} \rightarrow \mathcal{U}$.
Embed $\mathbf{s}$ in $\mathcal{M}_j$ where $j \leftarrow \text{argmin}_i \, |\mathcal{U}_i(\mathbf{s}) - \mu(\mathbf{C}_i, \mathbf{R}_i, \mathbf{t}_i)|$.
$\mathcal{Y}_{\mathcal{S}} \leftarrow \mathcal{Y}_{\mathcal{S}} \cup \mathbf{y}_{\mathbf{s}}$

Methodology

A Generalized Non-parametric Framework For Multi-Manifold Learning



Methodology

A Generalized Non-parametric Framework For Multi-Manifold Learning

Input: Batch $\mathbf{B}$, Stream $\mathbf{S}$; Parameters ϵ, $\mathbf{k}$, $\mathbf{l}$, λ, $\boldsymbol{\sigma}_t$, $\mathbf{n}_s$

Output: LDE $\mathcal{Y}_{\mathcal{S}}$

Partition $\mathbf{B}$ into clusters $\mathbf{C}_{i=1,2\ldots p}$.

Compute low dim. emb. $\forall \mathbf{C}_{i=1,2\ldots p}$ using $\mathcal{A}$.

Estimate $\phi_i^{\mathcal{GP}}$ $\forall \mathbf{C}_{i=1,2\ldots p}$ using $\mathcal{EST}_{\mathcal{A}}$.

Determine support ξ_s using $\mathbf{C}_{i=1,2\ldots p}$.

Compute $\{\mathbf{R}_i, \mathbf{t}_i\}_{i=1,2\ldots p}$ which maps $\mathcal{M}_i \rightarrow \mathcal{U}$.

For each $\mathbf{s} \in \mathbf{S}$

Using $\mathcal{GPR}_{\mathcal{A}}$, compute $\boldsymbol{\mu}_i, \boldsymbol{\sigma}_i$ for $\mathbf{s}$ $\forall i = 1, 2 \ldots p$.

$j \leftarrow \text{argmin}_i\, \boldsymbol{\sigma}_i$.

Embed $\mathbf{s}$ in $\mathcal{M}_j$ if $\boldsymbol{\sigma}_j \leq \boldsymbol{\sigma}_t$, otherwise add $\mathbf{s}$ to $\mathbf{S}_u$.

Re-run Batch Phase with $\mathbf{B} \cup \mathbf{S}_u$ when $\mathbf{S}_u \geq \mathbf{n}_s$.

$\mathcal{Y}_{\mathcal{S}} \leftarrow \mathcal{Y}_{\mathcal{S}} \cup \mathbf{y}_{\mathbf{s}}$

Methodology

S-Isomap++ - Specific Instantiation For Isomap

- Use Isomap for learning low-dimensional embeddings for $\mathbf{C}_{i=1,2\ldots p}$.
- Out-of-Sample Extension performed for streaming samples $\mathbf{s} \in \mathbf{S}$ using Streaming-Isomap.

Methodology

Streaming-LLE - Specific Instantiation For LLE

- Use LLE for learning low-dimensional embeddings for $\mathbf{C}_{i=1,2\ldots p}$.
- Out-of-Sample Extension performed for streaming samples $\mathbf{s} \in \mathbf{S}$ using OOSE-LLE.

Methodology

OOSE-LLE - Out-Of-Sample Extension For LLE

Input: $\mathbf{s}$, $\mathbf{C}_i$, $\mathbf{LDE}_i$

Output: $\mathbf{y_s}$

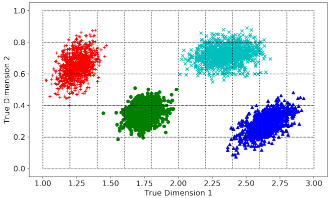
1: $\zeta_{\mathbf{s}} \leftarrow \text{KNN}(\mathbf{s}, \mathbf{C}_i)$

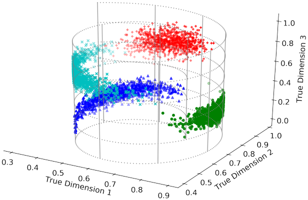
2: $\mathbf{w}^* \leftarrow \underset{w}{\text{argmin}} \left\| \left(\mathbf{s} - \sum_{x_j \in \zeta_{\mathbf{s}}} w_j x_j \right) \right\|^2$

3: **return** $\left(\sum_{y_j \in \zeta_{\mathbf{s}}} \mathbf{w}_j^* \mathbf{y}_j \right)$

Experiments
Datasets

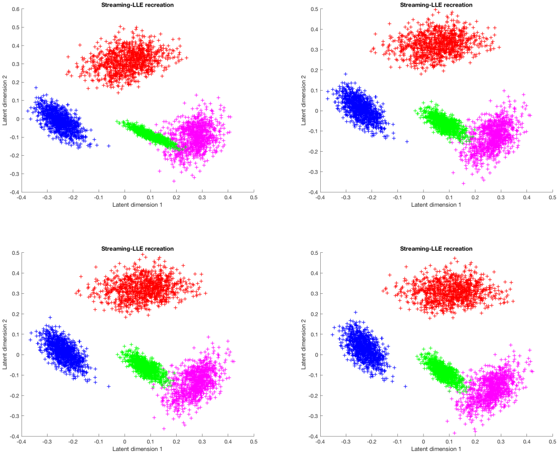
- Euler Isometric Swiss Roll - Synthetically generated dataset consisting of four $\mathbb{R}^2$ Gaussian patches embedded into $\mathbb{R}^3$ using a non-linear function $\psi(\cdot)$.
- Gas Sensor Array Dataset (GSAD) - Benchmark dataset which uses measurements from 16 chemical sensors used to discriminate between 6 gases at various concentrations.





Streaming-LLE

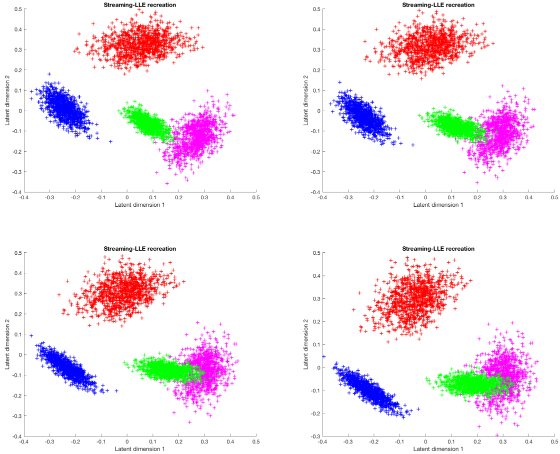
Results - Effect Of Changing **k**



Top Left: $k = 8$, Top Right: $k = 16$, Bottom Left: $k = 24$, Bottom Right: $k = 32$

Streaming-LLE

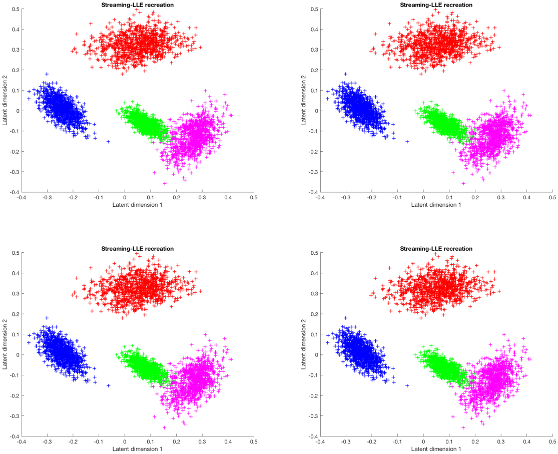
Results - Effect Of Changing l



Top Left: $l = 1$, Top Right: $l = 2$, Bottom Left: $l = 4$, Bottom Right: $l = 8$

Streaming-LLE

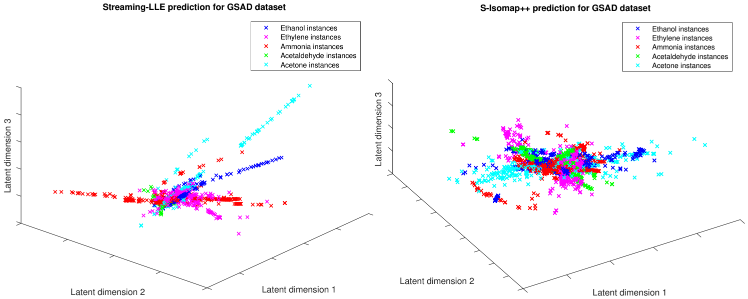
Results - Effect Of Changing λ



[Top Left: $\lambda = 0.005$, Top Right: $\lambda = 0.01$, Bottom Left: $\lambda = 0.02$, Bottom Right: $\lambda = 0.04$]

Streaming-LLE

Results - Comparison Between Streaming-LLE And S-Isomap++



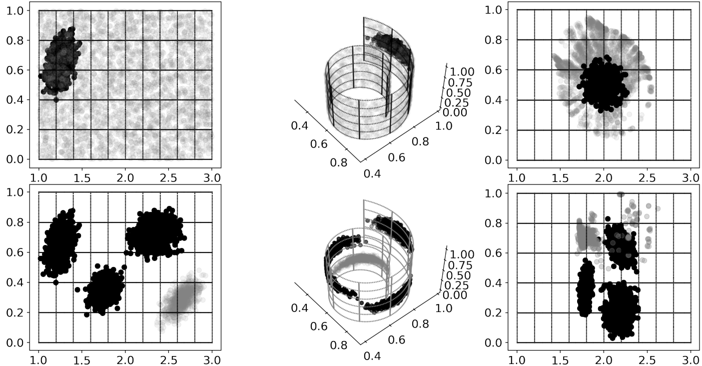
[Low-dimensional embedding uncovered by the Streaming-LLE algorithm on the Gas Sensor Array dataset. S-Isomap++ seems to uncover embeddings whose manifolds have smooth surfaces, while Streaming-LLE seems to uncover individual manifolds which are linear but disjoint and non-smooth.]

GP-Isomap

Handling Non-stationary Streams

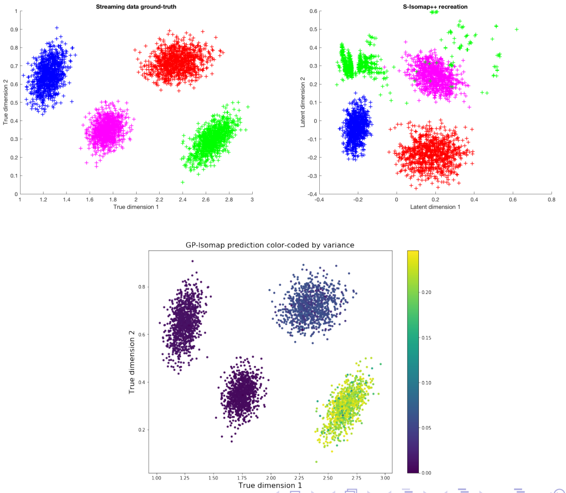
Motivation:

- S-Isomap++ cannot detect and handle changes in the stream distribution.



GP-Isomap
Motivation

- Fits a GP on batch data.
- Computes GP predictions on streaming samples.
- Uses GP variance to identify possible shifts in stream.
- Subsequently, re-trains batch to handle novel instances.



GP-Isomap
Methodology

- Uses Isomap for learning low-dimensional embeddings for $\mathbf{C}_{i=1,2\ldots p}$.
- For hyper-parameter estimation, uses low-dimensional embeddings uncovered by Isomap and Geodesic Distance based kernel.
- For Gaussian Process (GP) regression, uses low-dimensional embeddings uncovered by Isomap, Geodesic Distance based kernel and GP specific estimated hyper-parameters.

GP-Isomap

Geodesic-Distance Based Kernel

The GP-Isomap algorithm uses a novel geodesic distance based kernel function defined as:

$$k(\mathbf{y}_i, \mathbf{y}_j) = \boldsymbol{\sigma}_s^2 \exp\left(-\frac{\mathbf{b}_{i,j}}{2\ell^2}\right)$$

where $\mathbf{b}_{i,j}$ is the ij^{th} entry of the normalized geodesic distance matrix $\mathbf{B}$, σ_s^2 is the signal variance (whose value is fixed as 1.0 in this work) and ℓ is the length scale hyper-parameter.

GP-Isomap

Geodesic-Distance Based Kernel

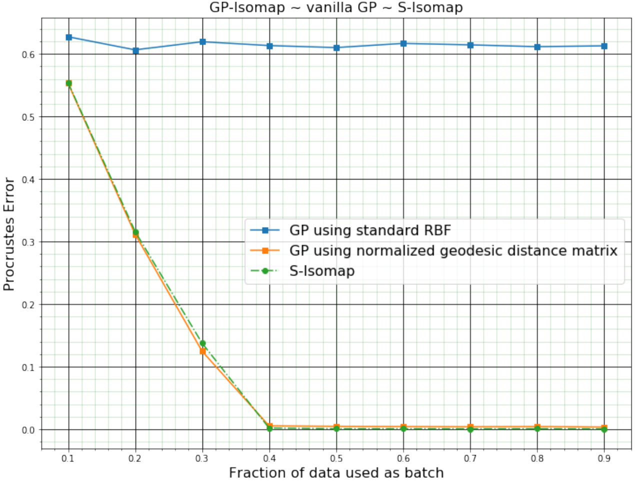
The novel kernel is positive-definite (PD) as demonstrated below :-

$$\mathbf{K}(\mathbf{x},\mathbf{y}) = \mathbf{I} + \sum_{i=1}^{d}\left[\exp\left(-\frac{\lambda_i}{2\ell^2}\right) - 1\right]\mathbf{q}_i\mathbf{q}_i^T = \mathbf{I} + \mathbf{Q}\widetilde{\mathbf{\Lambda}}\mathbf{Q}^\mathbf{T}$$

where $\widetilde{\mathbf{\Lambda}} = \begin{bmatrix} \left[\exp\left(-\frac{\lambda_1}{2\ell^2}\right) - 1\right] & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \left[\exp\left(-\frac{\lambda_d}{2\ell^2}\right) - 1\right] \end{bmatrix}$ and $\{\lambda_i, \mathbf{q}_i\}_{i=1\ldots d}$ are the eigenvalue/eigenvector pairs of **B**.

GP-Isomap

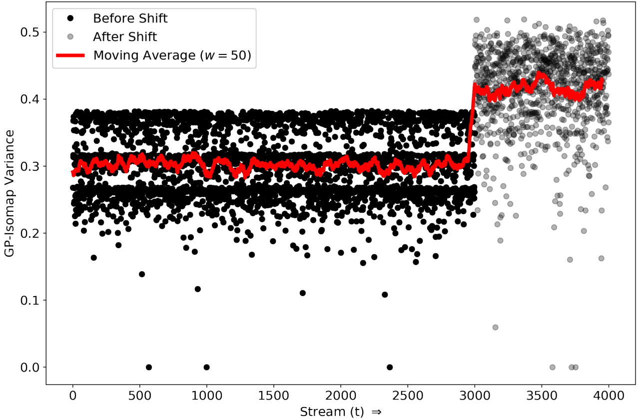
Results



[Procrustes error (PE) between the ground truth with a) GP-Isomap (blue line) with the geodesic distance based kernel, b) S-Isomap (dashed blue line with dots) and c) GP-Isomap (green line) using the Euclidean distance based kernel, for different fractions (f) of data used in the batch $\mathcal{B}$.]

GP-Isomap

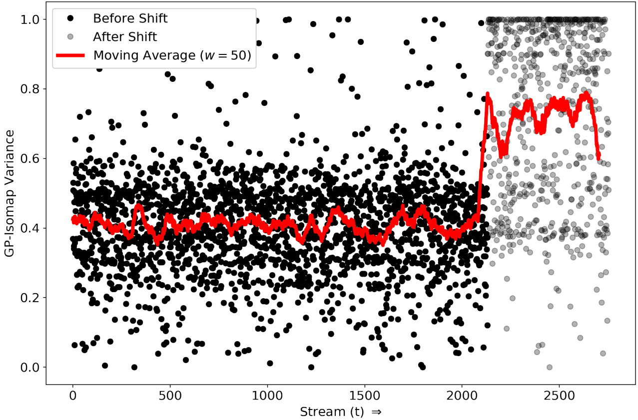
Results



[Using variance to detect *concept-drift* using the four patches dataset.Initially, when stream consists of samples generated from known modes, variance is low, later when samples from an unrecognized mode appear, variance shoots up. We can also observe the three variance "bands" above corresponding to the variance levels of the three modes for $t \leq 3000$.]

GP-Isomap

Results



[Using variance to identify *concept-drift* for the GSAD dataset. The introduction of points from an unknown mode in the stream results in variance increasing drastically as demonstrated by the mean (red line). The spread of variances for points from known modes ($t \precsim 2000$) is also smaller, compared to the spread for the points from the unknown mode ($t \succsim 2000$).]

S-Isomap

Theoretical Results

Theorem

Given uniformly sampled, unimodal distribution from which the batch dataset $\mathcal{B}$ for S-Isomap is derived from, $\exists n_0$ i.e. for $n \geq n_0$ the Procrustes Error $\epsilon_{Proc}(\boldsymbol{\tau}_{\mathcal{B}}, \boldsymbol{\tau}_{ISO})$ between $\boldsymbol{\tau}_{\mathcal{B}} = \phi^{-1}(\mathcal{B})$, the true underlying representation and $\boldsymbol{\tau}_{ISO} = \hat{\phi}^{-1}(\mathcal{B})$, the embedding uncovered by Isomap is small ($\epsilon_{Proc} \approx 0$) i.e. the batch phase of the S-Isomap algorithm converges.

Proof.

- [Bernstein *et al.*] showed that a data set $\mathcal{B}$ having samples drawn from a Poisson distribution with density function α satisfying certain conditions, leads to

$$(1 - \lambda_1) \leq \frac{d_G(x, y)}{d_M(x, y)} \leq (1 + \lambda_2) \; [\forall x, y \in \mathcal{B}] \quad (1)$$

S-Isomap
Theoretical Results

Proof.

- $\widetilde{D}_{\mathrm{G}} = \widetilde{D}_{\mathrm{M}} + \Delta\widetilde{D}_{\mathrm{M}}$
- Equating the expected sample size $(n\widetilde{\alpha})$ from a fixed distribution to the density function α, we get the threshold for n_{o} i.e.

$$\begin{aligned} n_{\mathrm{o}} &= (1/\widetilde{\alpha})\log(V/(\mu\widetilde{V}(\delta/4)))/\widetilde{V}(\delta/2) \\ &= (1/\widetilde{\alpha})\left[\log(V/\mu\eta_d(\lambda_2\epsilon/16)^d)\right]/\eta_d(\lambda_2\epsilon/8)^d \end{aligned} \tag{2}$$

where $\widetilde{D}_{\mathrm{M}}$ and $\widetilde{D}_{\mathrm{G}}$ represent the squared distance matrix corresponding to $d_{\mathrm{M}}(x,y)$ and $d_{\mathrm{G}}(x,y)$ respectively, $\widetilde{\alpha}$ is the probability of selecting a sample from $\mathcal{B}$, $V =$ volume of the manifold, $\widetilde{V}(r) = \eta_d r^d$ and $\eta_d =$ volume of unit ball in $\mathbb{R}^d$.

S-Isomap

Theoretical Results

Proof.

- [Sibson *et al*] demonstrated the robustness of MDS to small perturbations i.e. let F perturb the true squared-distance matrix B to $B + \Delta B = B + \epsilon F$. PE between the embeddings uncovered by MDS for B and $B + \Delta B$ equates to $\frac{\epsilon^2}{4} \sum_{j,k} \frac{{e_j^T F e_k}^2}{\lambda_j + \lambda_k} \approx 0$ for small perturbation matrix F.
- Substituting $\epsilon = 1$ and replacing B with $\widetilde{D}_M$ and ΔB with $\Delta\widetilde{D}_M$ above, we get our result, since the entries of $\Delta\widetilde{D}_M$ are very small i.e. $\{0 \leq \Delta\widetilde{D}_M(i,j) \leq \lambda^2\}_{1 \leq i,j \leq n}$ where $\lambda = \max(\lambda_1, \lambda_2)$ for small λ_1, λ_2.

GP-Isomap

Theoretical Results

Theorem

The prediction τ_{GP} of GP-Isomap is equivalent to the prediction τ_{ISO} of S-Isomap upto translation, rotation and scaling factors i.e. the Procrustes Error $\epsilon_{Proc}(\tau_{GP}, \tau_{ISO})$ between τ_{GP} and τ_{ISO} is 0.

Proof.

- Want to show $\epsilon_{\text{Proc}}(\tau_{\text{GP}}, \tau_{\text{ISO}}) = 0$.
- Subsequently, demonstrate that τ_{GP} is a scaled, translated, rotated version of τ_{ISO}.

GP-Isomap

Theoretical Results

Proof.

- The 1st dimension for S-Isomap prediction can be written as

$$\tau_{\text{ISO}_1} = \frac{\sqrt{\lambda_1}}{2} \sum_{i=1}^{n} \mathbf{q}_{1,i} \left(\gamma - \mathbf{g}^2_{i,n+1} \right) \tag{3}$$

- The 1st dimension for GP-Isomap prediction can be written as

$$\tau_{\text{GP}_1} = \frac{\alpha \sqrt{\lambda_1}}{1 + \alpha \mathbf{c}_1} \sum_{i=1}^{n} \mathbf{q}_{1,i} \left(1 - \frac{\mathbf{g}^2_{i,n+1}}{2\ell^2} \right) \tag{4}$$

where $\gamma = \left(\frac{1}{n} \sum_{j} \mathbf{g}^2_{i,j} \right)$, $\lambda_1 = 1^{st}$ eigenvalue of $\mathbf{B}$ and $\mathbf{q}_1$ the corresponding eigenvector, $\alpha = \frac{1}{\left(1 + {\sigma_n}^2\right)}$ and $\mathbf{c}_1 = \left[\exp\left(-\frac{\lambda_1}{2\ell^2} \right) - 1 \right]$.

GP-Isomap

Theoretical Results

Proof.

- (3) is a scaled, translated, rotated version of (4).
- Similarly, for each of the dimensions ($1 \leq i \leq d$), τ_{GP_i} can be shown to be a scaled, translated, rotated version of τ_{ISO_i}.
- We consolidate these individual scaling, translation and rotation factors together into single collective factors and demonstrate the required result.

∎

Thesis

Conclusions & Future Work

- Can work with only a fraction of the data and still be able to learn, while processing the remaining data "cheaply".
- Demonstrate theoretically that a "point of transition" exists for certain algorithms.
- Provide error metrics to practically identify them.
- Formulate a generalized OOSE framework for streaming NLSDR.
- Including other NLSDR methods as part of this framework and understanding relationships with other members of the NLDR family are future research directions.

Thesis

Acknowledgements

- Dr Varun Chandola
- Dr Jaroslaw Zola
- Dr Nils Napp
- Dr Bina Ramamurthy
- Dr Xin He
- Dr Haimonti Dutta
- NSF for supporting my PhD.