Error Metrics for Learning Reliable Manifolds from Streaming Data

Frank Schoeneman* Suchismit Mahapatra*
Varun Chandola Nils Napp Jaroslaw Zola

Department of Computer Science
University at Buffalo

SIAM International Conference on Data Mining 2017

Motivation

Massive amounts of data

- Huge amounts of data is coming from high-performance high-fidelity numerical simulations, high-resolution scientific instruments or Internet of Things feeds.
- Real-world data is typically a result of complex non-linear processes, but can often be described by a low-dimensional manifold.

Motivation

Massive amounts of data

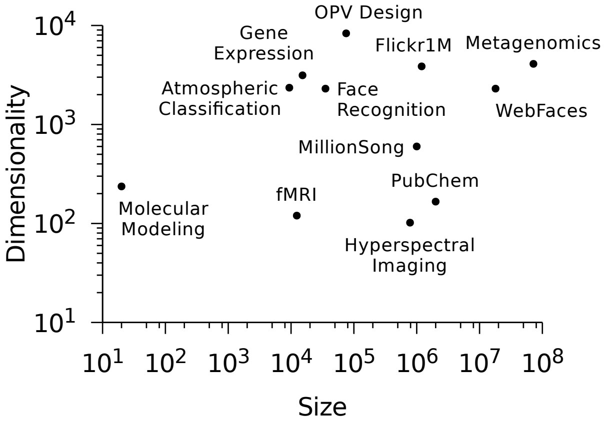

Figure: Topology of high-dimensional, massive datasets

Motivation

Nonlinear Process Dynamics

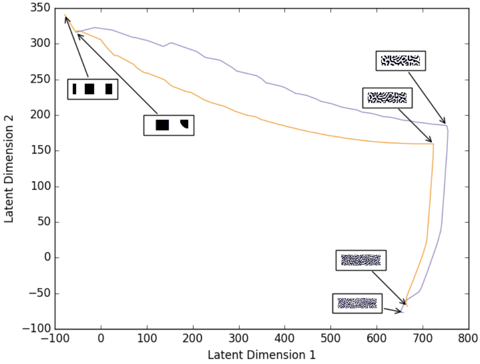

Figure: Morphological parametric trajectories for a nonlinear process.

Nonlinear Dimension Reduction (NLDR)

Formulation

Definition

Given $X = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n]^\top$, such that each $\mathbf{x}_i \in \mathbb{R}^D$, the task is to find a corresponding low-dimensional representation, $\mathbf{y}_i \in \mathbb{R}^d$, for each $\mathbf{x}_i$, where $d \ll D$.

- We assume there exists a function $\phi : \mathbb{R}^d \to \mathbb{R}^D$ that maps each data sample $\mathbf{y}_i \in \mathbb{R}^d$ to $\mathbf{x}_i \in \mathbb{R}^D$.
- The goal is to learn the inverse mapping, ϕ^{-1}, that can be used to map high-dimensional $\mathbf{x}_i$ to low-dimensional $\mathbf{y}_i$, i.e. $\mathbf{y}_i = \phi^{-1}(\mathbf{x}_i)$.

Nonlinear Dimension Reduction (NLDR)

Formulation

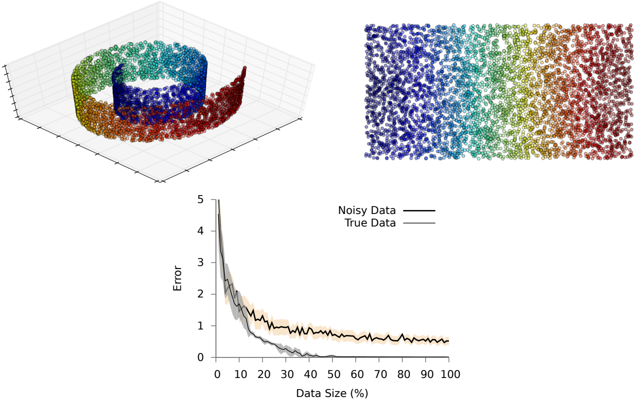

Figure: Procrustes error between true and approximate mapping learnt on increasing number of data points, with and without sampling error.

Nonlinear Dimension Reduction (NLDR)

Overview & Workflow

- NLDR techniques i.e. Isomap, Diffusion Maps, Laplacian Eigenmaps, Locally Linear Embedding rely on the spectral decomposition of the feature matrix that captures properties of the underlying sub-manifold.

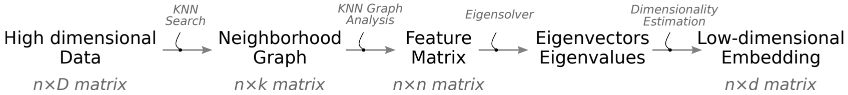

Figure: General NLDR workflow

Nonlinear Dimension Reduction (NLDR)

Isomap

Definition

A manifold $\mathcal{M}$ is a metric space with the following property: if $x \in \mathcal{M}$, then there exists some neighborhood $\mathcal{U}$ of x and $\exists n$ such that $\mathcal{U}$ is homeomorphic to $\mathbb{R}^n$.

Nonlinear Dimension Reduction (NLDR)

Isomap

Definition

A manifold $\mathcal{M}$ is a metric space with the following property: if $x \in \mathcal{M}$, then there exists some neighborhood $\mathcal{U}$ of x and $\exists n$ such that $\mathcal{U}$ is homeomorphic to $\mathbb{R}^n$.

- Isomap is a non-linear generalization of the classical Multi Dimensional Scaling(MDS) algorithm.
- The intuition is to perform MDS, not in the input space, but rather in the geodesic space of the non-linear data manifold.
- But there are plenty of challenges to manifold learning.

Nonlinear Dimension Reduction (NLDR)

Challenges

- Widely used manifold learning methods have been designed for off-line or batch processing.
- Standard methods are computationally expensive or impractical to apply to high-throughput data streams.
- Error in manifold learning is not yet completely understood, making error measurement on streaming data all the more complex.
- Applying Isomap to data streams and formulating the notion of collective error has not been well studied.

S-Isomap

Procrustes Analysis

- To measure the notion of error, we use Procrustes analysis.
- The idea is to align two matrices, $\mathcal{A}$ and $\mathcal{B}$, by finding the optimal translation t, rotation $\mathcal{R}$, and scaling s that minimizes the Frobenius norm between the two aligned matrices i.e.:

$$\epsilon_{proc}(\mathcal{A}, \mathcal{B}) = \min_{\mathcal{R},t,s} \|s\mathcal{R}\mathcal{B} + t - \mathcal{A}\|_F. \quad (1)$$

- The above has a closed form solution obtained by performing SVD on $\mathcal{A}\mathcal{B}^T$.
- We determine how well $LDE_{\mathcal{X}}$ represents the low-dimensional ground truth $GT_{\mathcal{X}}$ using the above error metric $\epsilon_{proc}(LDE_{\mathcal{X}}, GT_{\mathcal{X}})$.

S-Isomap

Reference Sample Method

- Allows us to measure error in the absence of low-dimensional ground truth.

1. Given dataset $\mathcal{X}$, choose $\mathcal{F} \subset \mathcal{X}$, a reference set, and two equal sized sample sets $\mathcal{R}_1, \mathcal{R}_2 \subset \mathcal{X}$ and create two data sets, $\mathcal{D}_1$ and $\mathcal{D}_2$, such that $\mathcal{D}_i = \mathcal{F} \cup \mathcal{R}_i$ for $i = 1, 2$.
2. Perform NLDR on each of $\mathcal{D}_i$ to get different approximations of $\mathcal{F}$. (Intuitively we learnt mappings $\hat{\phi}_1^{-1}$ and $\hat{\phi}_2^{-1}$ for same $\mathcal{F}$)
3. Compute reference sample error as:

$$\epsilon_{rs} = \epsilon_{proc}(\hat{\phi}_1^{-1}(\mathcal{F}), \hat{\phi}_2^{-1}(\mathcal{F})). \tag{2}$$

S-Isomap

Experiments using MNIST, Corel, Swiss Roll datasets

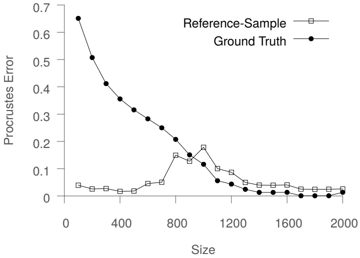

Figure: Demonstration of behavior of error of the Reference Sample method, as well as the Procrustes Analysis as we increase number of samples. Notice the similar asymptotic behavior of error.

S-Isomap

Experiments using MNIST, Corel, Swiss Roll datasets

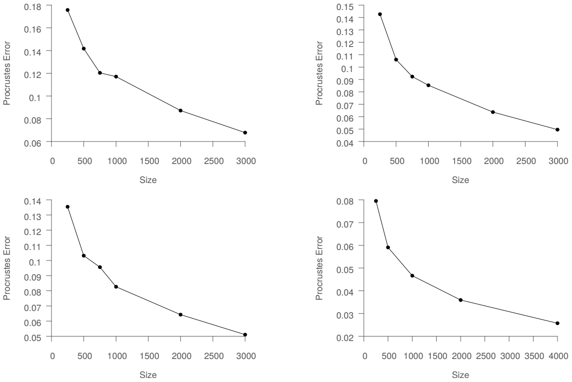

Figure: We actually need a much smaller dataset to adequately form a robust manifold structure !!

S-Isomap

Algorithm Design

- This key intuition allowed us to formulate a much cheaper means for mapping streaming points to the manifold.
- Choose an initial batch set $\mathcal{B}$ based on error analysis.
- Perform exact Isomap (or other NLDR) on this $\mathcal{B}$ to get the manifold $\mathcal{M} = LDE_{\mathcal{B}}$.
- Subsequently, map streaming points $s \in \mathcal{S}$ by matching their inner products with $LDE_{\mathcal{B}}$ to the computed geodesic distances with the k nearest neighbors of s.

S-Isomap

Proposed Algorithm

Algorithm 1 input: G_b, X_b, Y_b, $\mathbf{x}_s$, k

$\mathbf{kNN}$, $\mathbf{kDist} \leftarrow \mathrm{KNN}(\mathbf{x}_s, X_b, k)$

for $1 \le i \le n$ **do**

  $\mathbf{g}_i \leftarrow \min_{1\le j\le k}\{\mathbf{kDist}_j + G_{b_{\mathbf{kNN}_j,i}}\}$

end for

$\mathbf{c} \leftarrow \frac{1}{2}(\bar{g} \cdot \mathbf{1}_n - \mathbf{g} - \bar{\bar{G}}_b \cdot \mathbf{1}_n + \bar{\mathbf{G}}_b)$

$\mathbf{p} \leftarrow (Y_b^\top Y_b)^{-1} Y_b^\top \mathbf{c}$

$\hat{Y} \leftarrow [Y_b; \mathbf{p}]$

$\mathbf{y}_s \leftarrow \mathbf{p} - \bar{\hat{Y}}$

return $\mathbf{y}_s$

S-Isomap

Performance analysis

Method	Time Complexity
OOSE (non-incremental)	$\mathcal{O}(m * (n^2 \log(n) + n^2 k))$
OOSE (incremental)	$\mathcal{O}(\sum_{i=1}^{m+n}(iD + i^2 \log(i) + i^2 k))$
S-Isomap	$\mathcal{O}(n^3 + mn(D + d^2 + k))$

Table: $n = |B|$, $m = |S|$, $n \ll m$

OOSE above refers to the out-of-sample-extension technique proposed by Law and Jain (2006). S-Isomap requires $\mathcal{O}(max(n^2, nd))$ space for operation.

S-Isomap

Results for Euler Isometric Swiss Roll

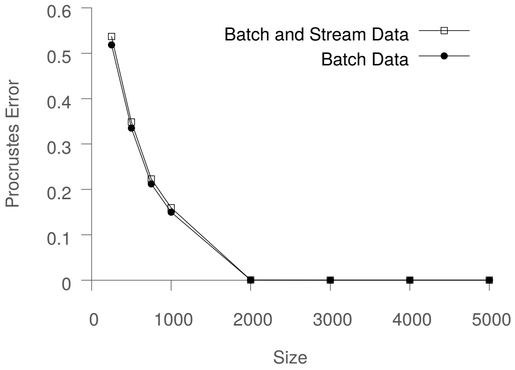

Figure: The results illustrate that the error due to streaming points is low as well as similar asymptotic behavior.

S-Isomap

Timing results

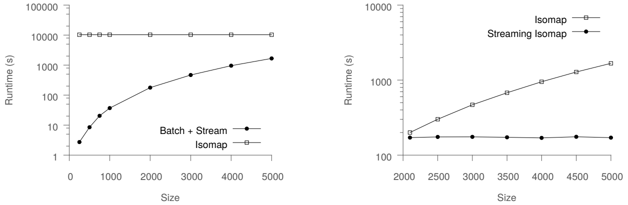

Figure: Timing results for S-Isomap. Results are in log scale and demonstrate the performance gain achieved.

Summary & Future work

Summary

- We studied & formulated the notion of error metrics for manifold learning techniques and quantify them, as well as we devise a technique to deal with scenarios wherein ground truth is unavailable to help quantify the error.
- We demonstrate that it is possible to learn a robust, stable manifold using only a subset of dataset.
- Consequently, we propose a novel efficient algorithm, suitable for high-volume and high-throughput stream processing, to incorporate streamed data into a stable manifold.

Summary & Future work

Future work

- S-Isomap is able to deal with uniform, unimodal distributions. We are currently working to extend this work to deal with non-uniform as well as potentially multi-modal distributions.
- Theoretical analysis to provide bounds on $|\mathcal{B}|$.
- Multi-manifold extensions which can work in parallel and thus improve efficiency.