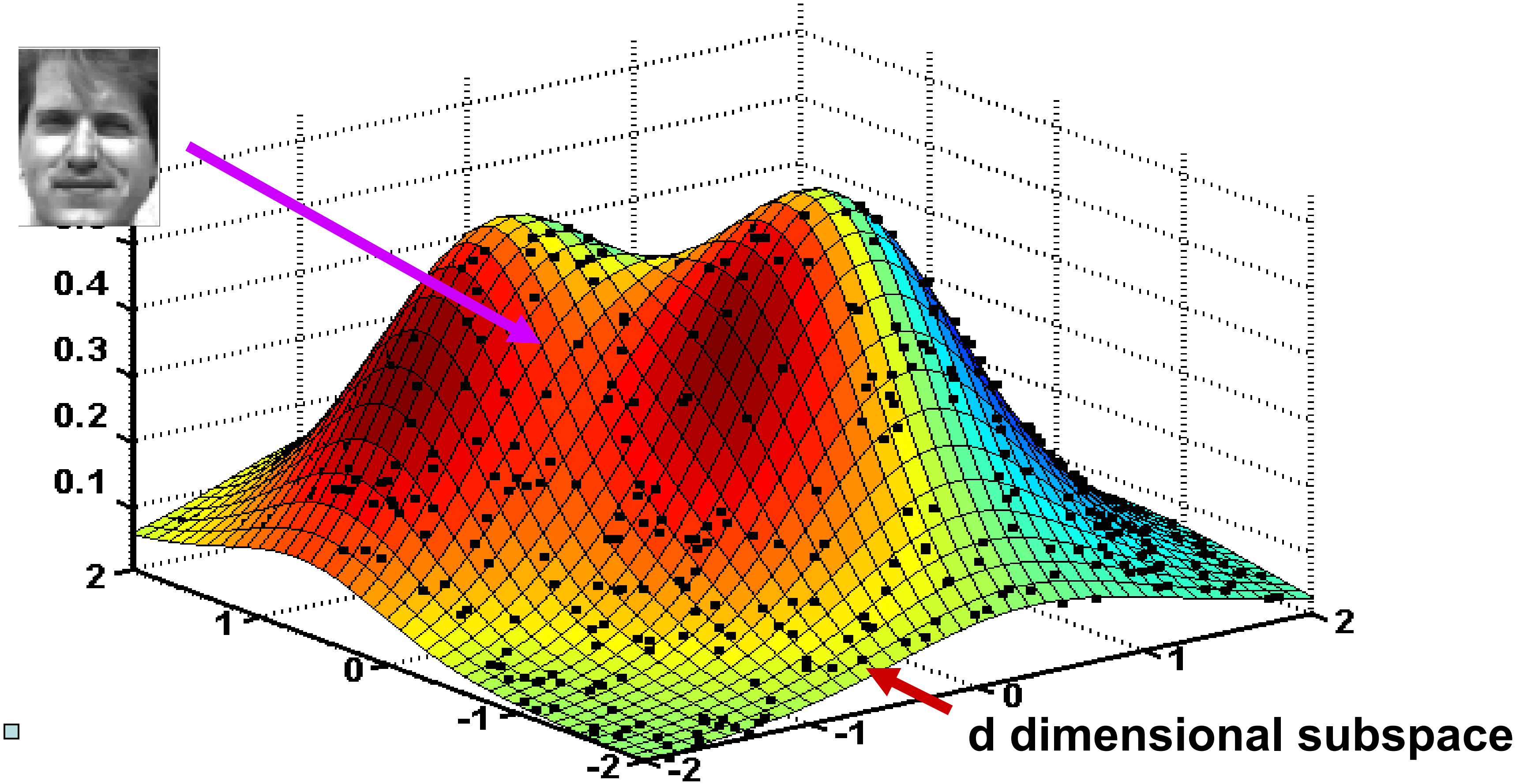
Dimension Reduction for Classification

Alfred O. Hero
Dept. EECS, Dept BME, Dept. Statistics
University of Michigan - Ann Arbor
hero@eecs.umich.edu
http://www.eecs.umich.edu/~hero

BIRS, July. 2005

1. **The manifold supporting data sample**
2. **Classification constrained dimension reduction**
3. **Dimension estimation on smooth manifolds**
4. **Applications**
5. **Conclusions**

1. Manifold supporting data sample



- Yale Face Database B: each 128x128 image lies in R^(16384)

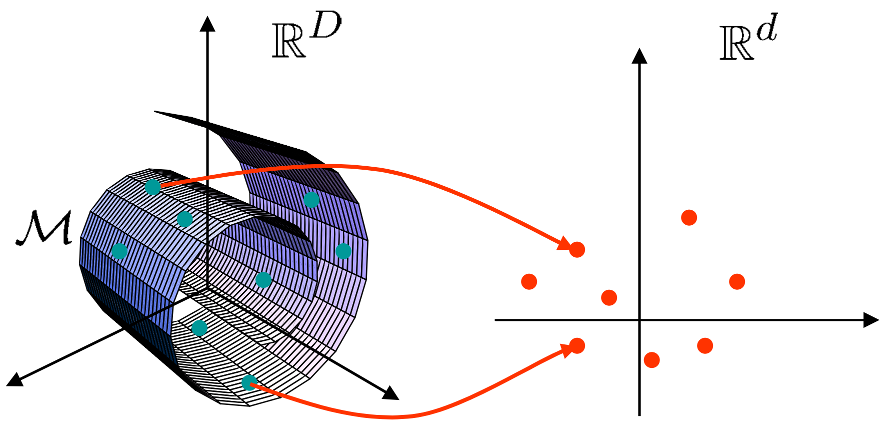
Data-driven dimensionality reduction

- Data-driven dimensionality reduction consists of:
  - Estimation of intrinsic dimension d:
    - Direct intrinsic dimension estimation
  - Reconstruction of data samples in the manifold domain:
    - Manifold learning
- Classifiers on intrinsic data dimension
  - Estimated dimension as a discriminant
  - Label-constrained manifold learning

Manifold Learning

Manifold learning problem setup:

Given a finite sampling $\mathcal{Y}_n = \{\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n\} \subset \mathbb{R}^D$ of a *d*-dimensional manifold $\mathcal{M} \subset \mathbb{R}^D$, find an embedding of $\mathcal{Y}_n$ into a subset $\mathcal{X}_n = \{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n\} \subset \mathbb{R}^d$ (usually $d \ll D$) without any prior knowledge about $\mathcal{M}$ or d.



Manifold learning background

Reconstructing the mapping and attributes of the manifold from a finite dataset falls into the general manifold learning problem.

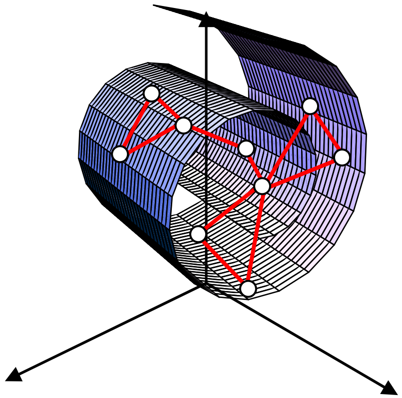
Manifold reconstruction for fixed d:

1. *ISOMAP*, Tenenbaum, de Silva, Langford (2000);
2. *Locally Linear Embeddings* (LLE), Roweiss, Saul (2000);
3. *Laplacian Eigenmaps*, Belkin, Niyogi (2002);
4. *Hessian Eigenmaps* (HLLE), Grimes, Donoho (2003);
5. *Local Space Tangent Alignment* (LTSA), Zhang, Za (2003);
6. *SemiDefinite Embedding* (SDE), Weinberger, Saul (2004).

Laplacian Eigenmaps

Laplacian Eigenmaps: preserving local information
(Belkin & Niyogi 2002)

1. Constructing an Adjacency Graph:



   a. compute a k-NN graph on the dataset;

   b. compute a similarity/weight matrix W between data points, that encodes neighborhood information (e.g., heat kernel):

$$w_{ij} = \begin{cases} e^{-\frac{|\boldsymbol{y}_i - \boldsymbol{y}_j|^2}{\epsilon}} & \text{, if } \boldsymbol{y}_i, \boldsymbol{y}_j \text{ are } k\text{-NN} \\ 0 & \text{, o.w.} \end{cases}$$

Laplacian Eigenmaps

2. Manifold learning as an optimization problem:

   a. objective function:

$$E(\mathcal{X}_n) = \sum_{ij} w_{ij} \, |\boldsymbol{x}_i - \boldsymbol{x}_j|^2 = 2 \, \mathrm{tr}\left(X \, L \, X^T\right) \; ,$$

   where $L = D - W$ $\left(D = \mathrm{diag}\left\{\sum_j W_{ji}\right\}\right)$

   is the *Graph Laplacian*.

   b. embedding is solution of

$$\min_{\substack{X D \mathbf{1} = \mathbf{0} \\ X D X^T = I}} \mathrm{tr}\left(X \, L \, X^T\right) \qquad (\star)$$

Laplacian Eigenmaps

3. Eigenmaps:

   a. solution to (✶) is given by the *d* generalized eigenvectors associated with the *d* smallest generalized eigenvalues that solve:

   $$L\boldsymbol{v} = \lambda D \boldsymbol{v}$$

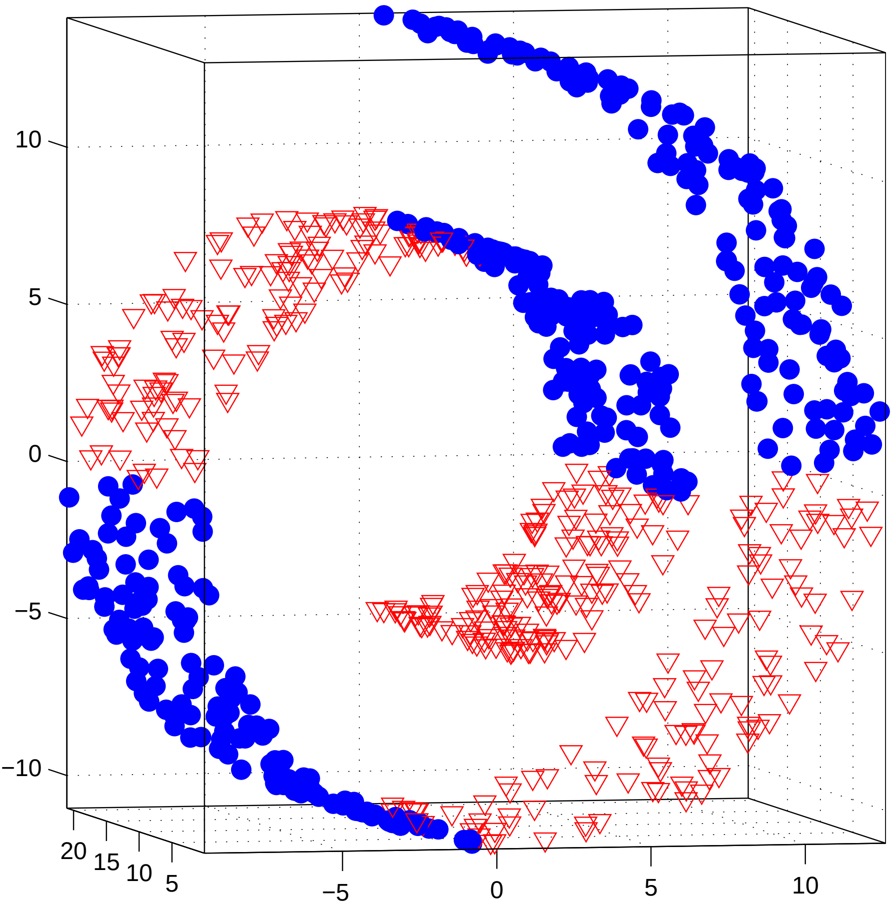
   equivalently, eigenvectors of the *normalized Graph Laplacian*

   $$\tilde{L} = D^{-1/2} L D^{-1/2}$$

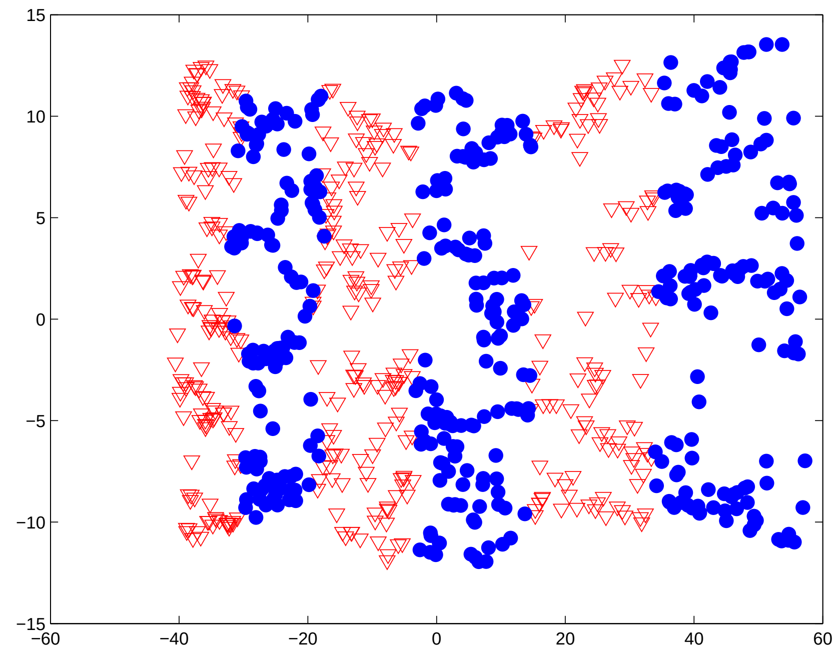
   b. if $V = [\boldsymbol{v}_1 \ldots \boldsymbol{v}_m]$ is the collection of such eigenvectors, then the embedded points are given by

   $$\boldsymbol{x}_i = (v_{i1}, \ldots, v_{im})^T, \; 1 \leq i \leq n$$

Dimension Reduction for Labeled Data



Original Data Y

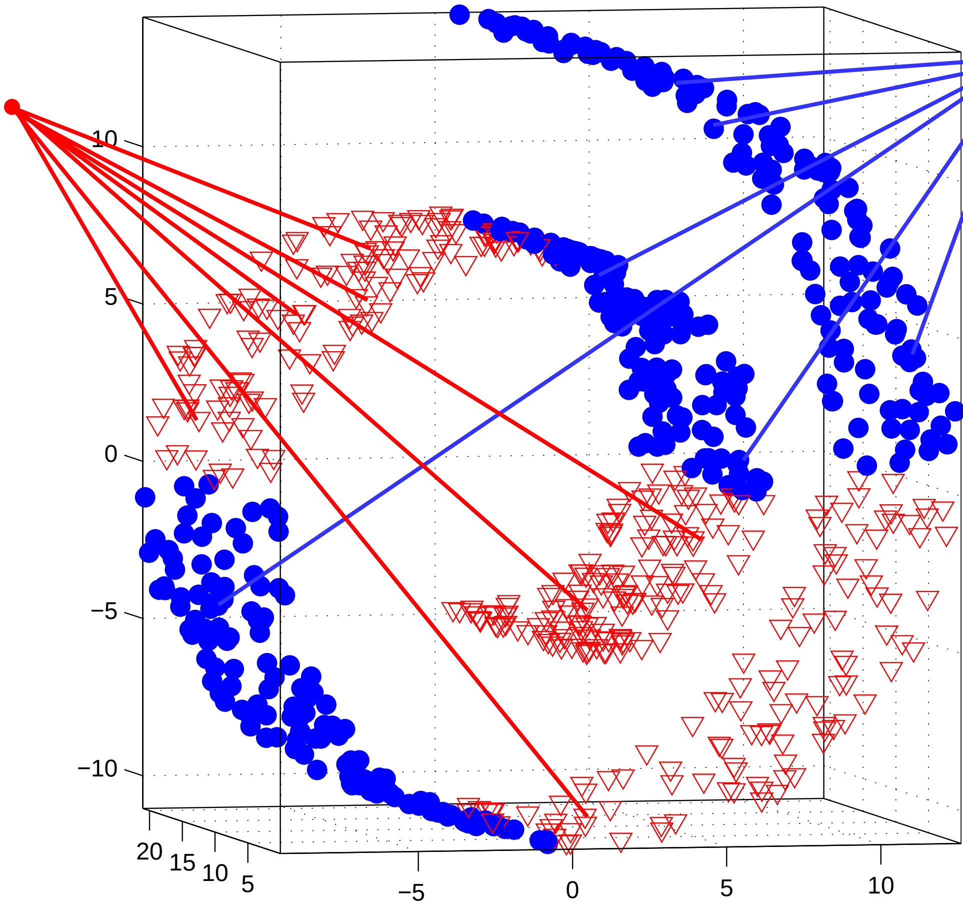


Dimension reduced Data X

800 points uniform on Swiss roll, 400 each class

2. Classification constrained dimensionality reduction

Adding class dependent constraints – “virtual” class vertices.



Label-penalized Laplacian Eigenmaps

1. If C is the class membership matrix (i.e., c_ij = 1 if point j is from class i), define the objective function:

$$E(\mathcal{Z}_n) = \sum_{ki} c_{ki} \, |\boldsymbol{z}_k - \boldsymbol{x}_i|^2 + \beta \sum_{ij} w_{ij} \, |\boldsymbol{x}_i - \boldsymbol{x}_j|^2 \ ,$$

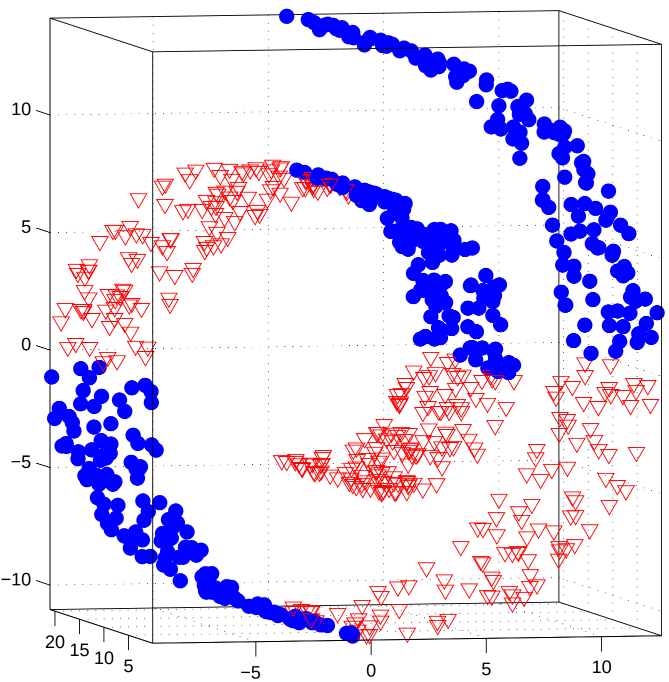
where $\mathcal{Z}_n = \{\boldsymbol{z}_1, \boldsymbol{z}_2, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$, $\boldsymbol{z}_1, \boldsymbol{z}_2$ are the “virtual” class centers and $\beta \geq 0$ is a regularization parameter.

2. Embedding is solution of

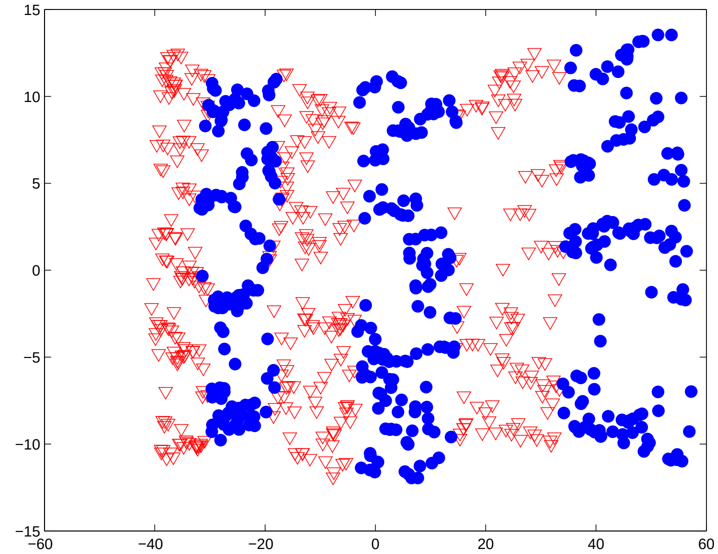
$$\min_{\substack{Z\,D\,\mathbf{1} = \mathbf{0} \\ Z\,D\,Z^T = I}} \operatorname{tr}\left(Z\,L\,Z^T\right) \ ,$$

where *L* is Laplacian of augmented weight matrix $K = \begin{bmatrix} I & C \\ C^T & \beta\,W \end{bmatrix}$

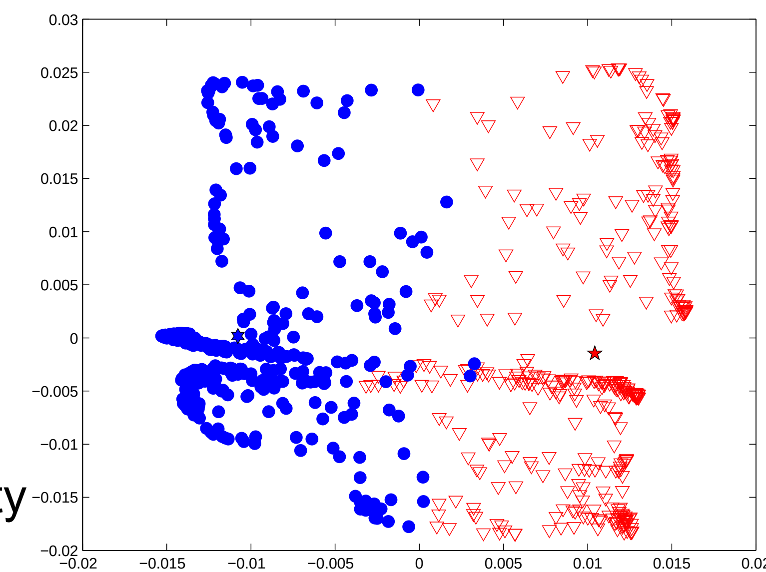
University of MICHIGAN



Unconstrained Dimensionality Reduction

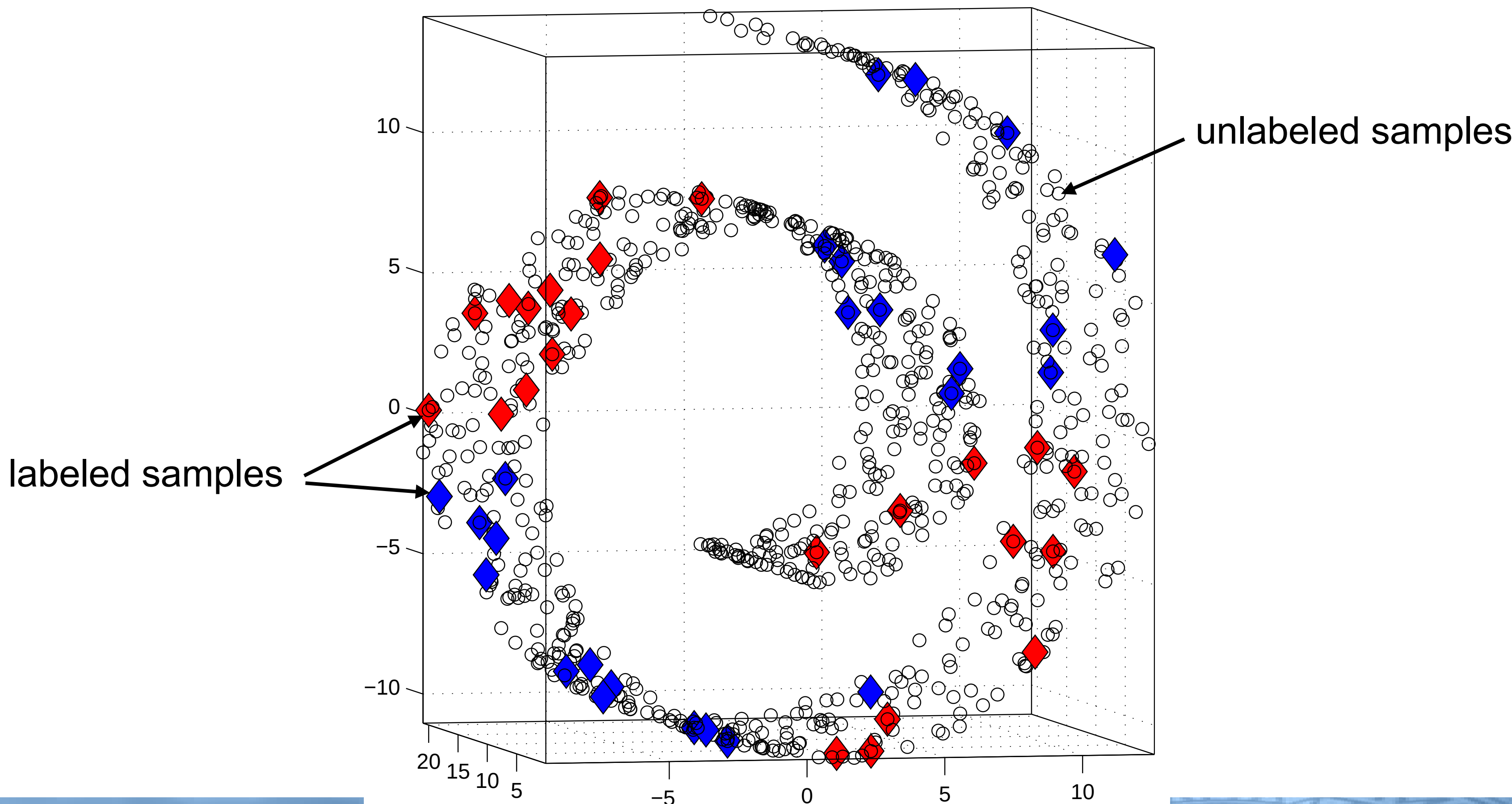


Classification Constrained Dimensionality Reduction



Partially Labeled Data

Semi-Supervised Learning on Manifolds



Semisupervised extension

Algorithm:

1. Compute the constrained embedding of the entire data set, inserting a zero column in *C* for each unlabeled sample.

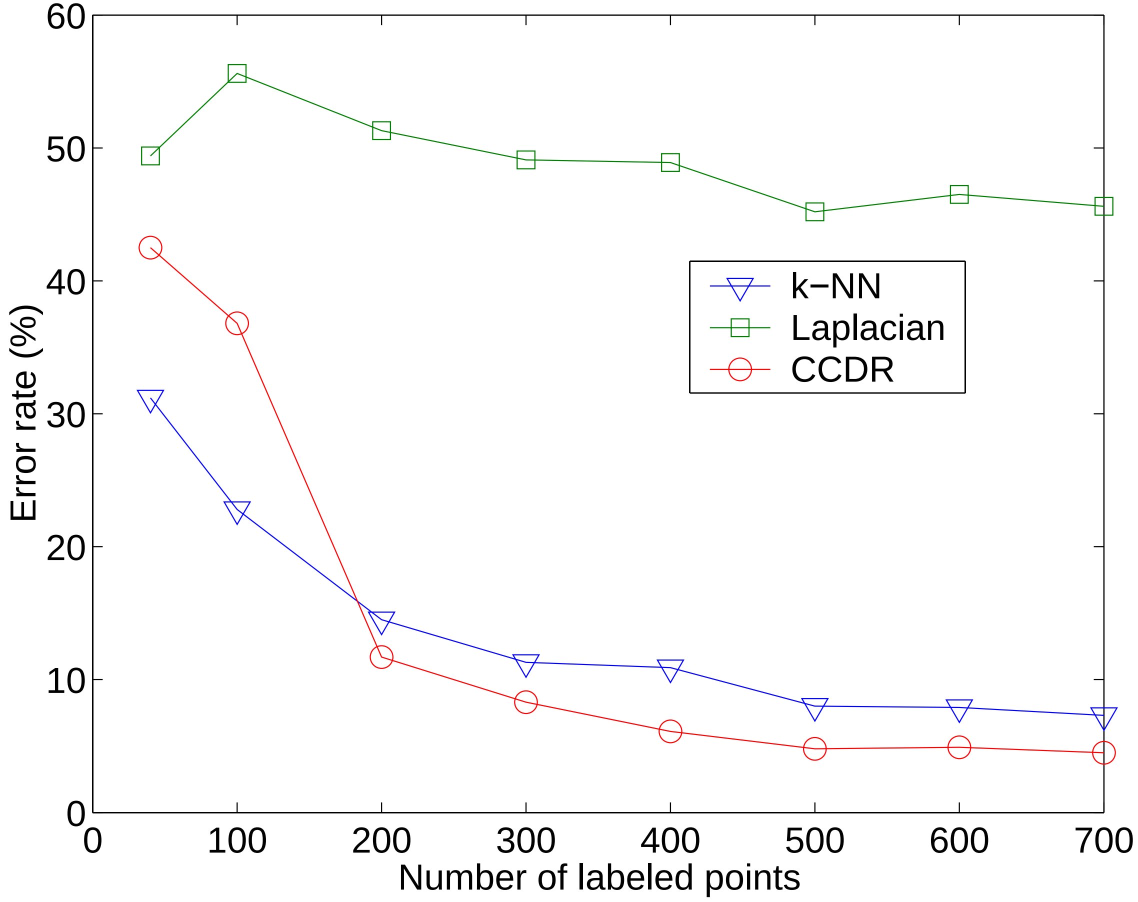
2. Fit a (e.g., linear) classifier to the labeled embedded points by minimizing the quadratic error loss:

$$\ell(\boldsymbol{a}) = \sum_{\substack{i\,:\,\boldsymbol{y}_i \text{ is} \\ \text{labeled}}} \left(c_i - \boldsymbol{a}^T \boldsymbol{x}_i\right)^2$$

3. For an unlabeled point , label it using the fitted (linear) classifier:

$$c_j = \begin{cases} 1 & \text{if } \boldsymbol{a}^T \boldsymbol{x}_j \geq 0 \\ -1 & \text{if } \boldsymbol{a}^T \boldsymbol{x}_j < 0 \end{cases}$$

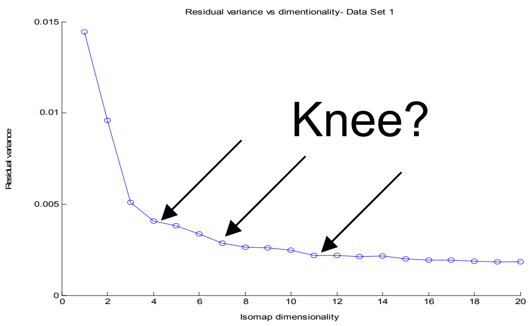
Classification Error Rates



Percentage of errors for labeling unlabeled samples as a function of the number of labeled points, out of a total of 1000 points on the Swiss roll.

3. Methods of dimension estimation

- Scree plots $g(u) = u, f(x) = c$
  - Plot residual fitting errors of SVD, Isomap, LE, LLE



ISOMAP residual curve

- Kolmogorov/Entropy/Correlation dimension $g(u) = u, f(x) = c$
  - Box counting, sphere packing (Liebovitch and Toth)

$$d = \lim_{r \to 0} \frac{\log N(r)}{\log(1/r)}$$

- Maximum likelihood $g(u) = u, f(x) = c$
  - Poisson approximation to Binomial (Levina&Bickel:2004)

$$\frac{k}{n} \approx f(x_o) V(d) \|x_o - x_{(k)}^{x_o}\|^d$$

- Entropic graphs $g(u) = u^\alpha$
  - Spanner-graph length approximation to entropy functional (Costa&Hero:2003)

$$L_n(\mathcal{X}_n)/n^{(d-1)/d} \to \beta_d \int_{\mathcal{M}} f^\alpha(x) dx$$

Euclidean Random Graphs

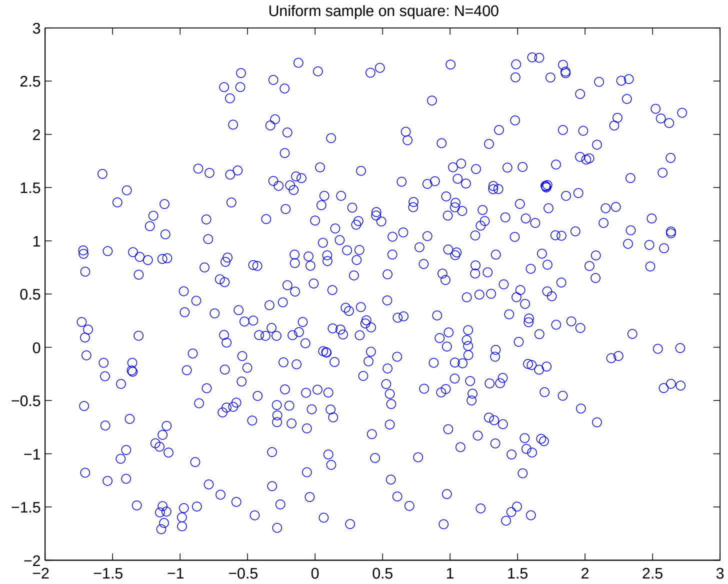
- $\mathcal{X}_n = \{X_1, \ldots, X_n\}$ data in D-dimensional Euclidean space
- Euclidean MST with edge power weighting gamma:

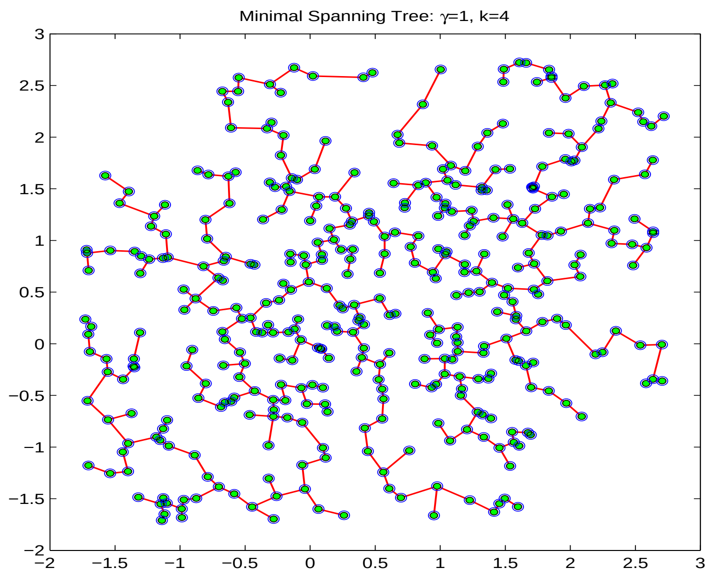
$$L_\gamma(\mathcal{X}_n) = \min_{E \in \mathcal{E}} \sum_{|e| \in E} |e|^\gamma$$

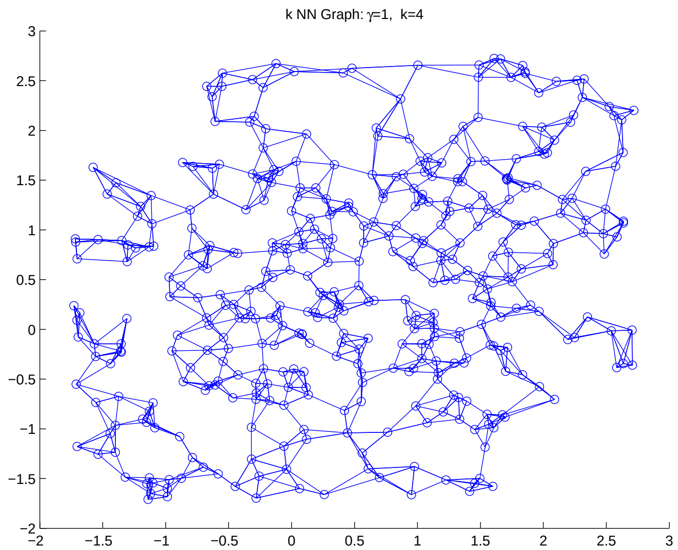
- $\mathcal{E}$ pairwise distance matrix over $\mathcal{X}_n$
- E edge length matrix of spanning trees over $\mathcal{X}_n$
- Euclidean k-NNG with edge power weighting gamma:

$$\mathcal{L}_{k,\gamma}(\mathcal{X}_n) = \sum_{i=1}^{n} \sum_{|e| \in E_k(X_i)} |e|^\gamma$$

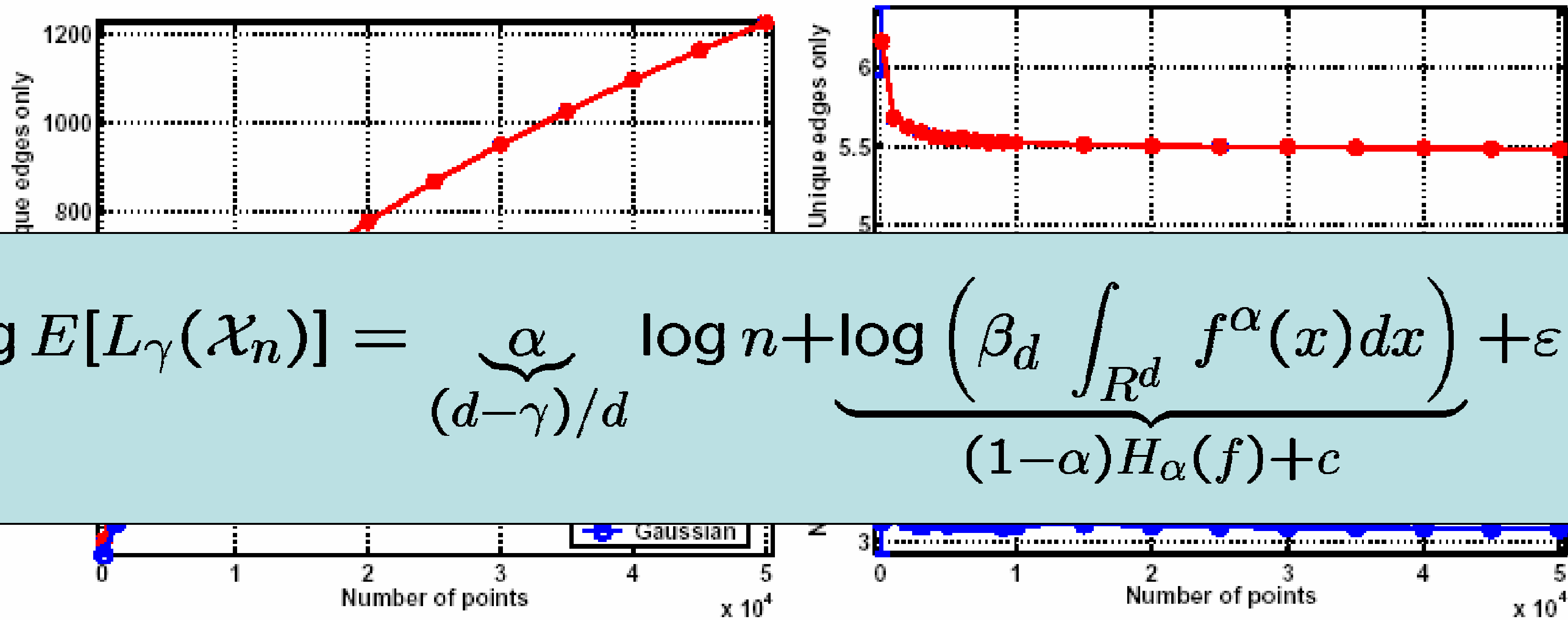
Example: Uniform Planar Sample







Convergence of Euclidean CQF's



$$\log E[L_\gamma(\mathcal{X}_n)] = \underbrace{\alpha}_{(d-\gamma)/d} \log n + \underbrace{\log\left(\beta_d \int_{R^d} f^\alpha(x)dx\right)}_{(1-\alpha)H_\alpha(f)+c} + \varepsilon(n)$$

Beardwood, Halton, Hammersley Theorem (BHH:1959):

$$L_\gamma(\mathcal{X}_n)/n^\alpha \to \beta_d \int_{R^d} f^\alpha(x)dx$$

$$\alpha = (d-\gamma)/d$$

k-NNG Convergence Theorem in Non-Euclidean Spaces

Let $\mathcal{M}$ be a compact smooth Riemann d-dimensional manifold embedded in $\mathbf{R}^D$. Let $2 \leq d \leq D$ and $0 < \gamma < d$. Suppose that $X_1, \ldots, X_n$ are i.i.d. random vectors on $\mathcal{M}$ with common bounded density f relative to $\mu_{\mathcal{M}}$. Then the total length of the k-NNG satisfies

$$L_\gamma(\mathcal{X}_n)/n^{(d'-\gamma)/d'} \to \begin{cases} \infty, & d' < d \\ \beta_d \int_{\mathcal{M}} f^\alpha(x)\, \mu_{\mathcal{M}}(dx), & d' = d \\ 0, & d' > d \end{cases},$$

(a.s) where $\alpha = (d-\gamma)/d$.

Costa, Hero: TSP(2004), Birkhauser(2005)

Number of correct dimension estimates over 30 trials as a function of the number of samples for the torus ($N = 5$, Q=10).

n	200	400	600
GMST	29	30	30
5-NN	29	30	30

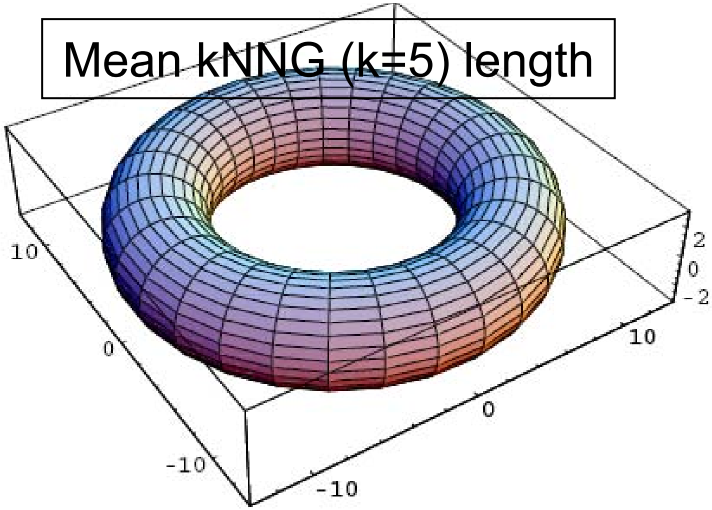
TABLE II

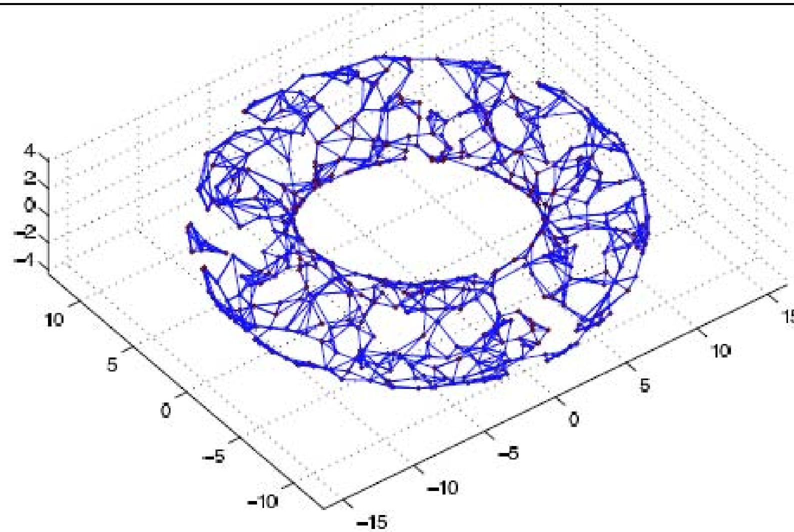
Entropy estimates for the torus ($n = 600$, $N = 5$, Q=10).

	emp. mean	std. deviation
GMST	10.0	0.55
5-NN	9.6	0.93

2.8
2
2.7
2.7
2.7
2.7
2.7
average length
number of samples

Mean kNNG (k=5) length





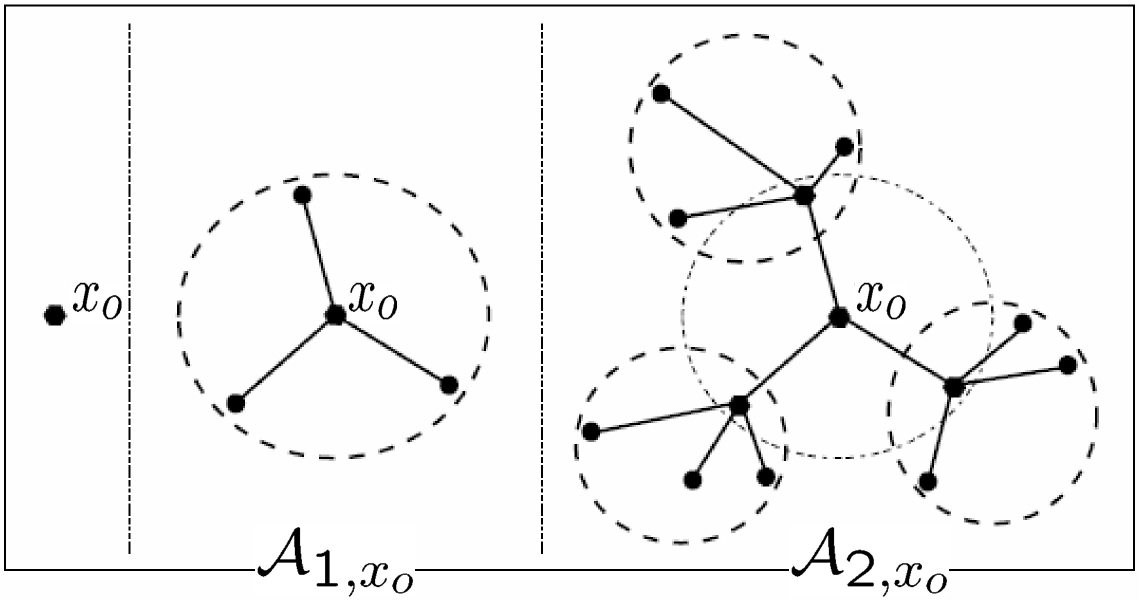
Local Extension via kNNG

– Initialize: $x_o \in \mathcal{X}_n, \quad \mathcal{A}_{1,x_o} = \mathcal{N}_{k,x_o}$

– For i=1,2,…p

- $\forall x_j \in \mathcal{A}_{i,x_o}$ compute $\mathcal{N}_{k,x_j}$ and set

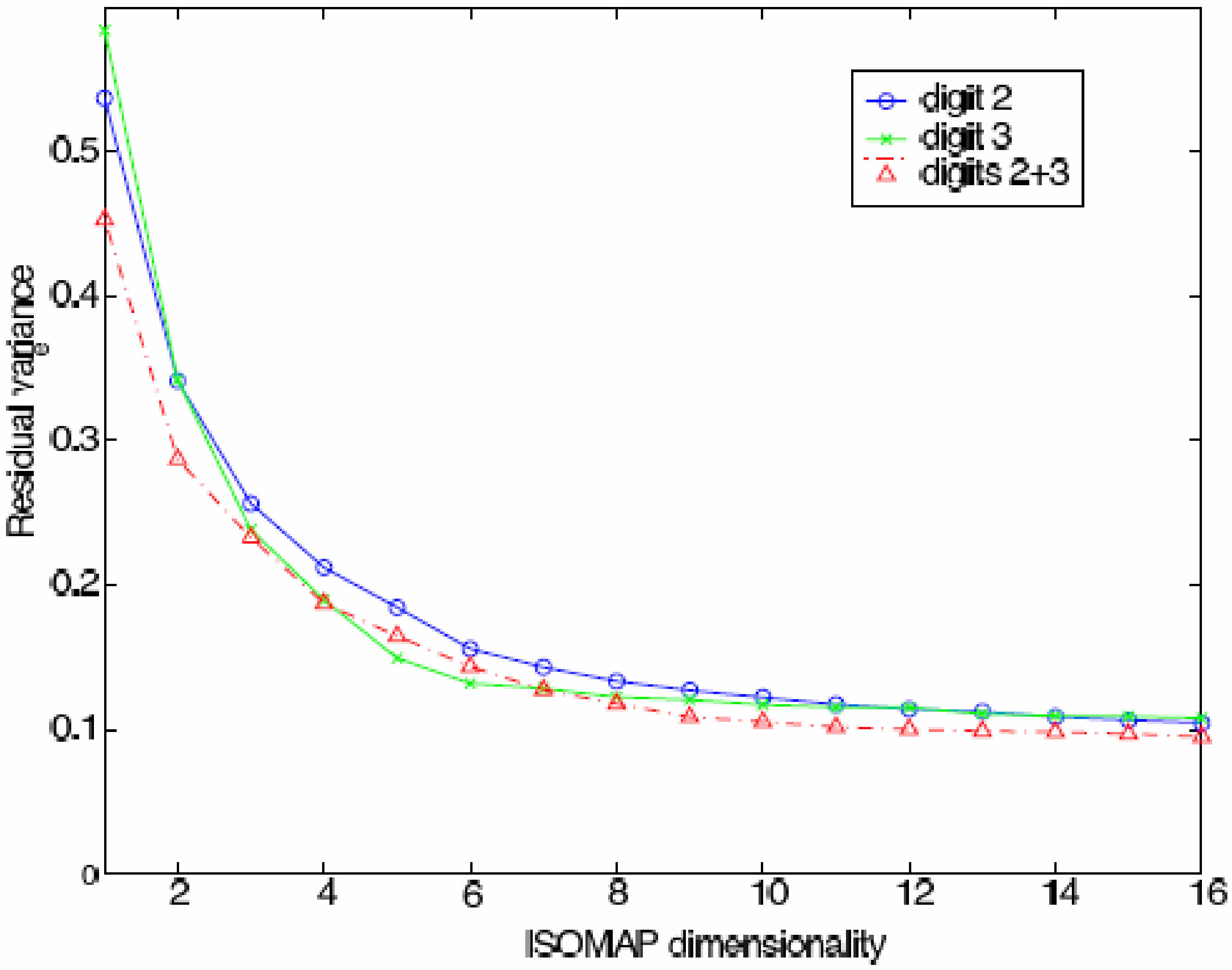
$$\mathcal{A}_{i+1,x_o} = \cup_{x_j} \mathcal{N}_{k,x_j} \cup \mathcal{A}_{i,x_o}$$



4. Application to MNIST Digits

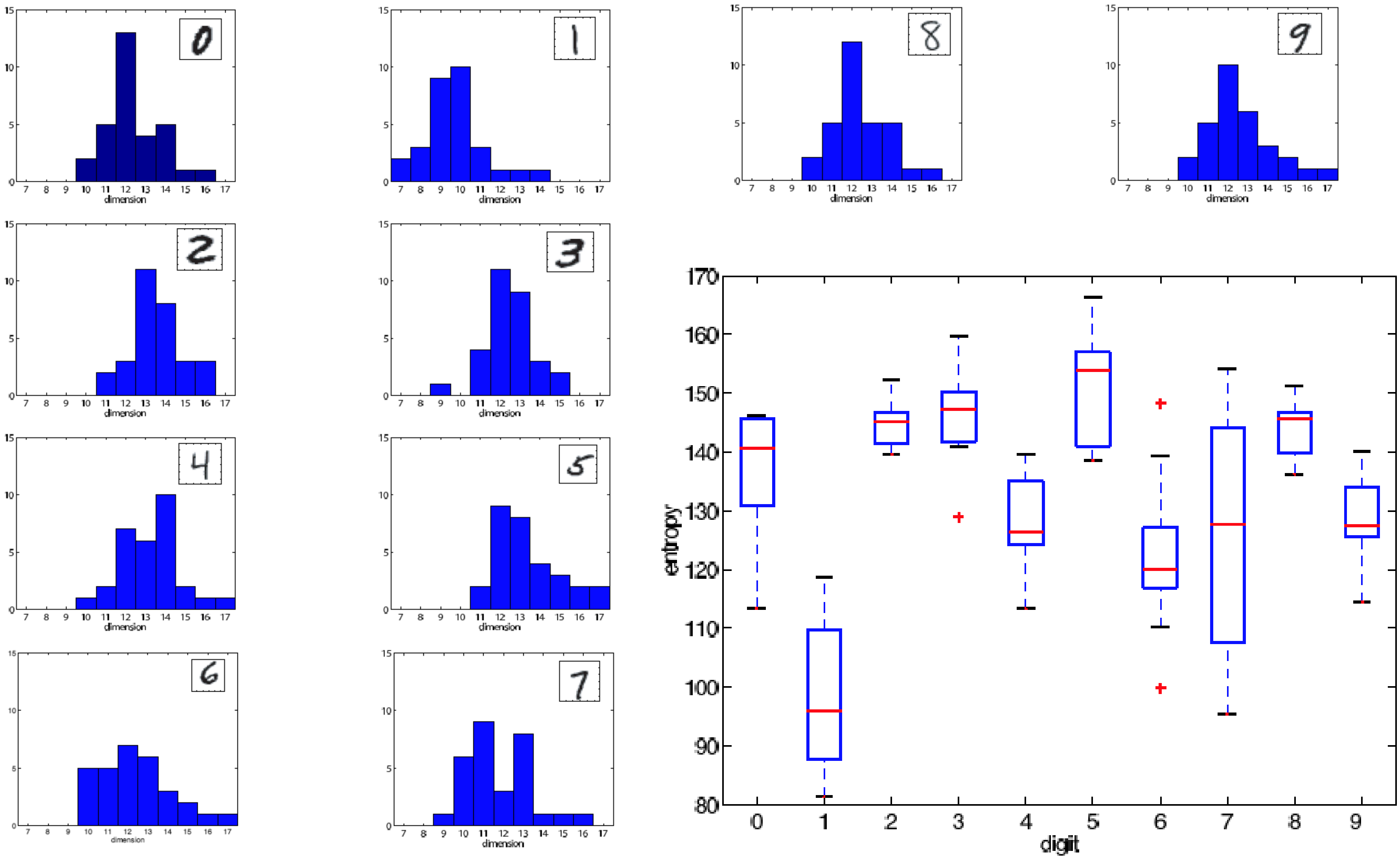
- Large database of 8 bit images of digits 0-9.
- 28x28 pixels for each image
- First 1000 images in training set used here
- Non-adaptive: digit labels are known

Scree Plot



Costa&Hero:Birkhauser05

Local Dimension/Entropy Statistics

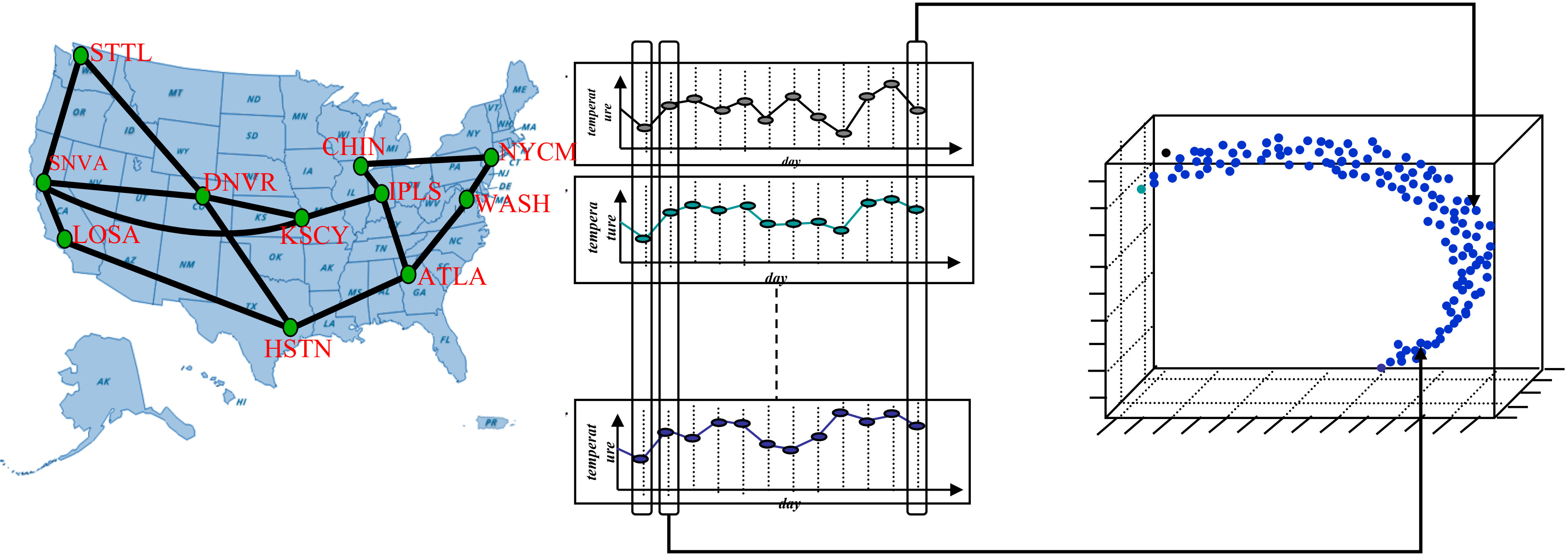


Costa&Hero:Birkhauser05

Adaptive Anomaly Detection

- Spatio-temporal measurement vector:

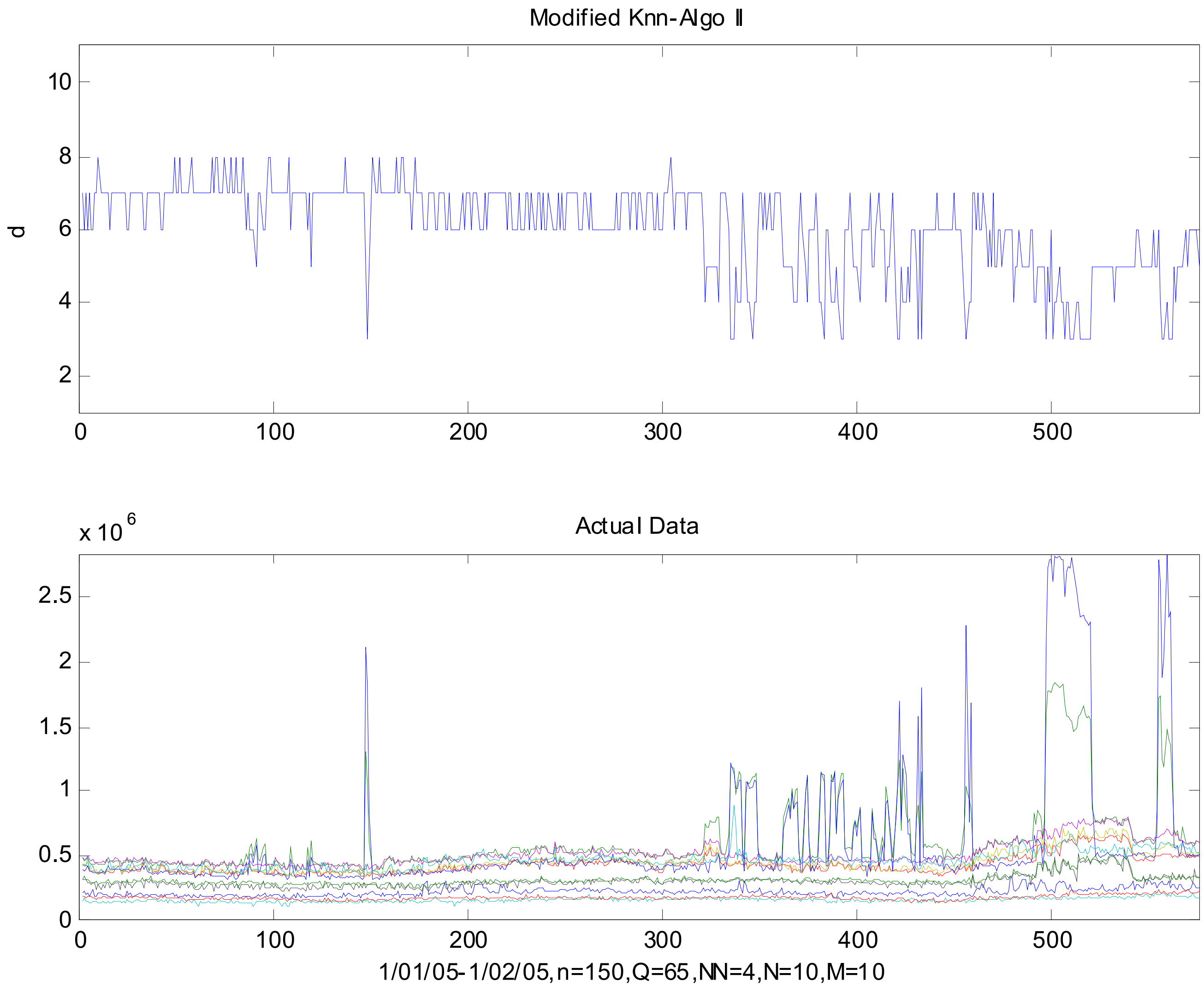
$$\mathbf{x}(t) = [\mathbf{x}_1(t), \ldots, \mathbf{x}_N(t)] \quad \forall t = 1 \ldots \tau$$



Data Observed from Abilene Network

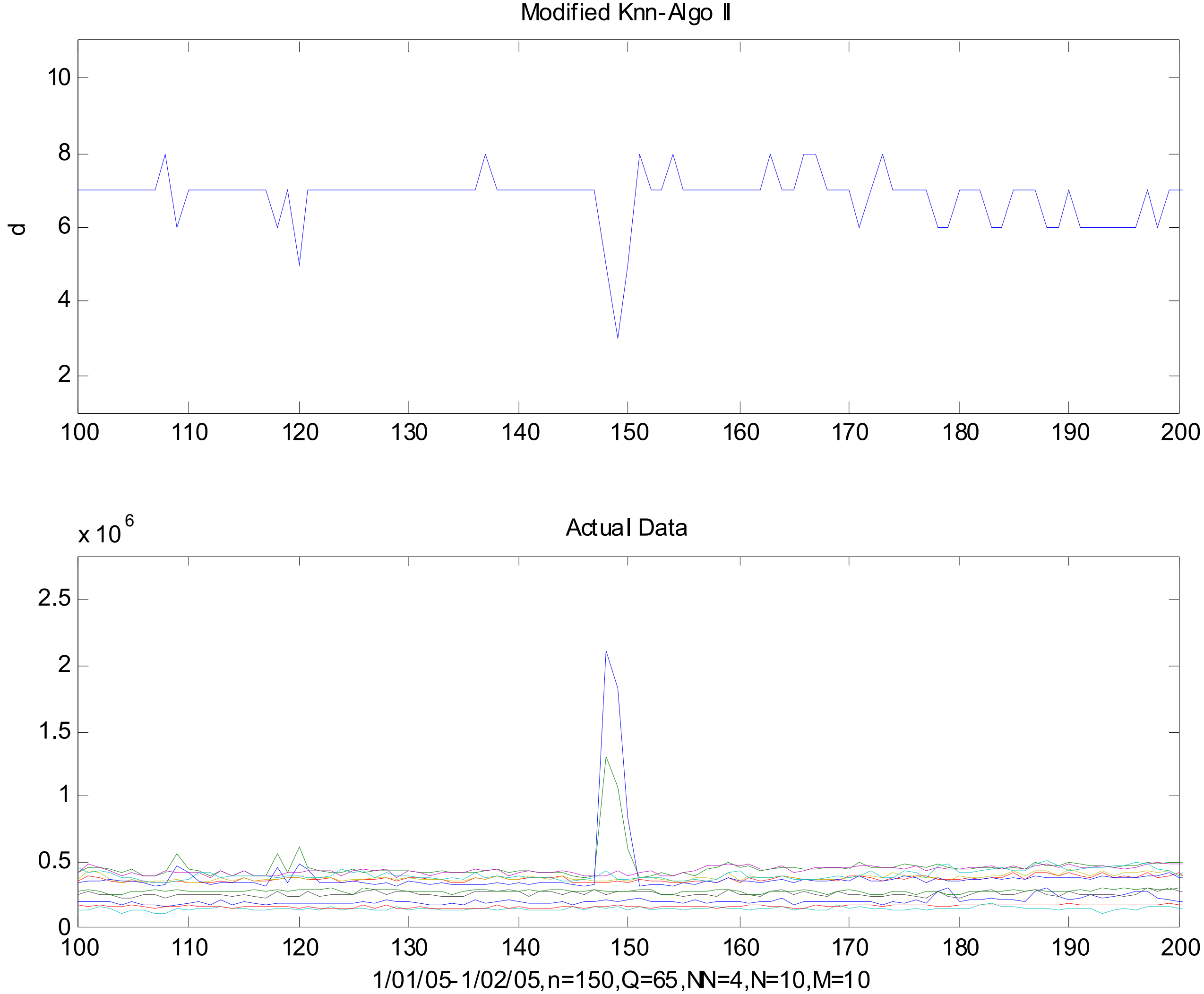
- Objective: detect changes in network traffic via local intrinsic dimension
- Hypotheses:
  - High traffic from few sources lowers the local dimension of the network traffic
  - Changes in distribution of dimension estimate can be used as a marker for more subtle changes in traffic
- Data collection period: 1/1/05-1/2/05
- Data sampling: packet flow sampled every 5 minutes from all 11 routers on Abilene Network
- Data fields: aggregate of all flows to/from all ports

KNN Algorithm (Costa)



Example – 1/1/05, 12:20 pm

- Large data transfers from IPs 145.146.96 and 192.31.120 drastically increase flows through Chicago and NYC.



5. Conclusions

- Classification constraints can be included in manifold learning dimension reduction algorithms
- kNNG jointly estimate dimension and entropy of high dimensional data
- Dimension can be used as a discriminant in anomaly detection
- Can be used as precursor to model reduction and database compression
- Methods only suffer from curse of *intrinsic* dimensionality

References

- J. Costa, N. Patwari and A. O. Hero, "Distributed multidimensional scaling with adaptive weighting for node localization in sensor networks", (http://www.eecs.umich.edu/~hero/Preprints/wmds_v9.pdf), ACM Journal on Networking To appear 2005.
- J. Costa and A. O. Hero, "Geodesic entropic graphs for dimension and entropy estimation in manifold learning," (http://www.eecs.umich.edu/~hero/Preprints/sp_mlsi_final_twocolumn.pdf) , *IEEE Trans. on Signal Process.*, Vol. 52, No. 8, pp. 2210-2221, Aug. 2004.
- J.A. Costa, A. Girotra and A.O. Hero, "Estimating Local Intrinsic Dimension with k-Nearest Neighbor Graphs," IEEE Workshop on Statistical Signal Processing (SSP), Bordeaux, July 2005. (http://www.eecs.umich.edu/~hero/Preprints/ssp_2005_final_1.pdf)
- J. Costa and A. O. Hero, "Classification constrained dimensionality reduction," (http://www.eecs.umich.edu/~hero/Preprints/costa_icassp2005.pdf), *Proc. of ICASSP* , Philadelphia, March, 2005.