Information divergence in high dimensions

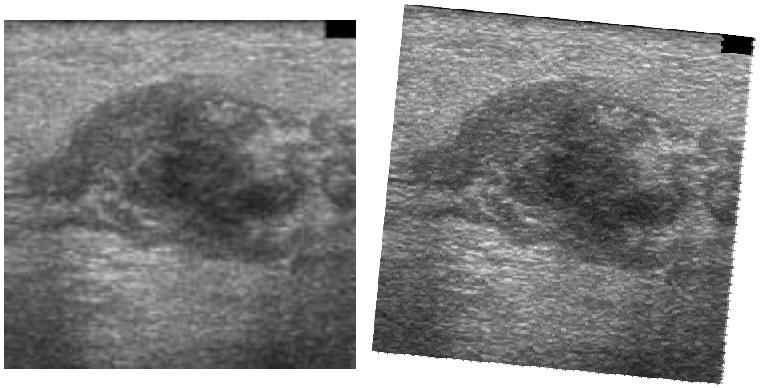
Alfred O. Hero
University of Michigan - Ann Arbor
hero@eecs.umich.edu
http://www.eecs.umich.edu/~hero

Oct 2004

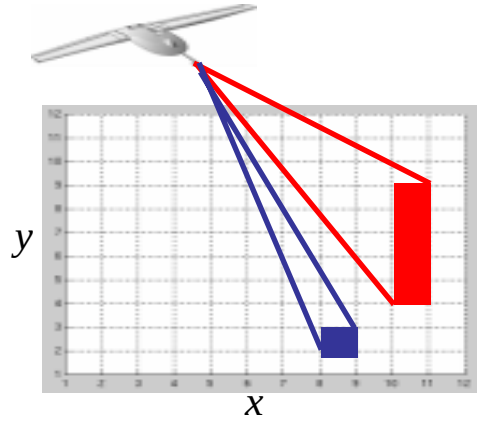
1. **Motivating applications**
2. **Entropic graph estimates**
3. **Dimension reduction and manifold learning**

Collaborators: Jose Costa and Huzefa Neemuchwala

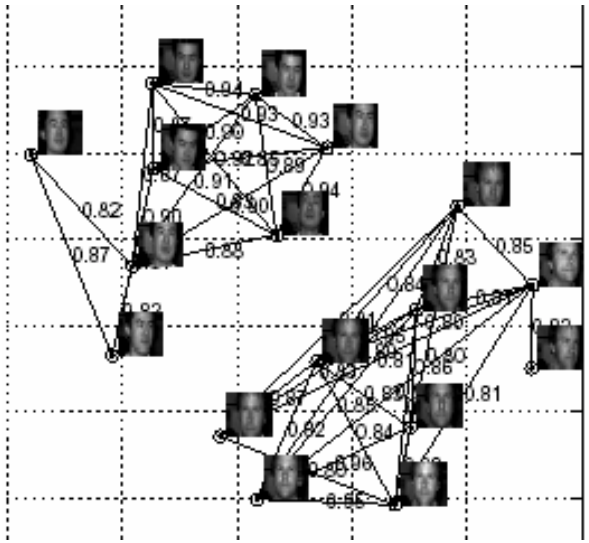
1. Motivating applications



Ultrasound Breast Registration



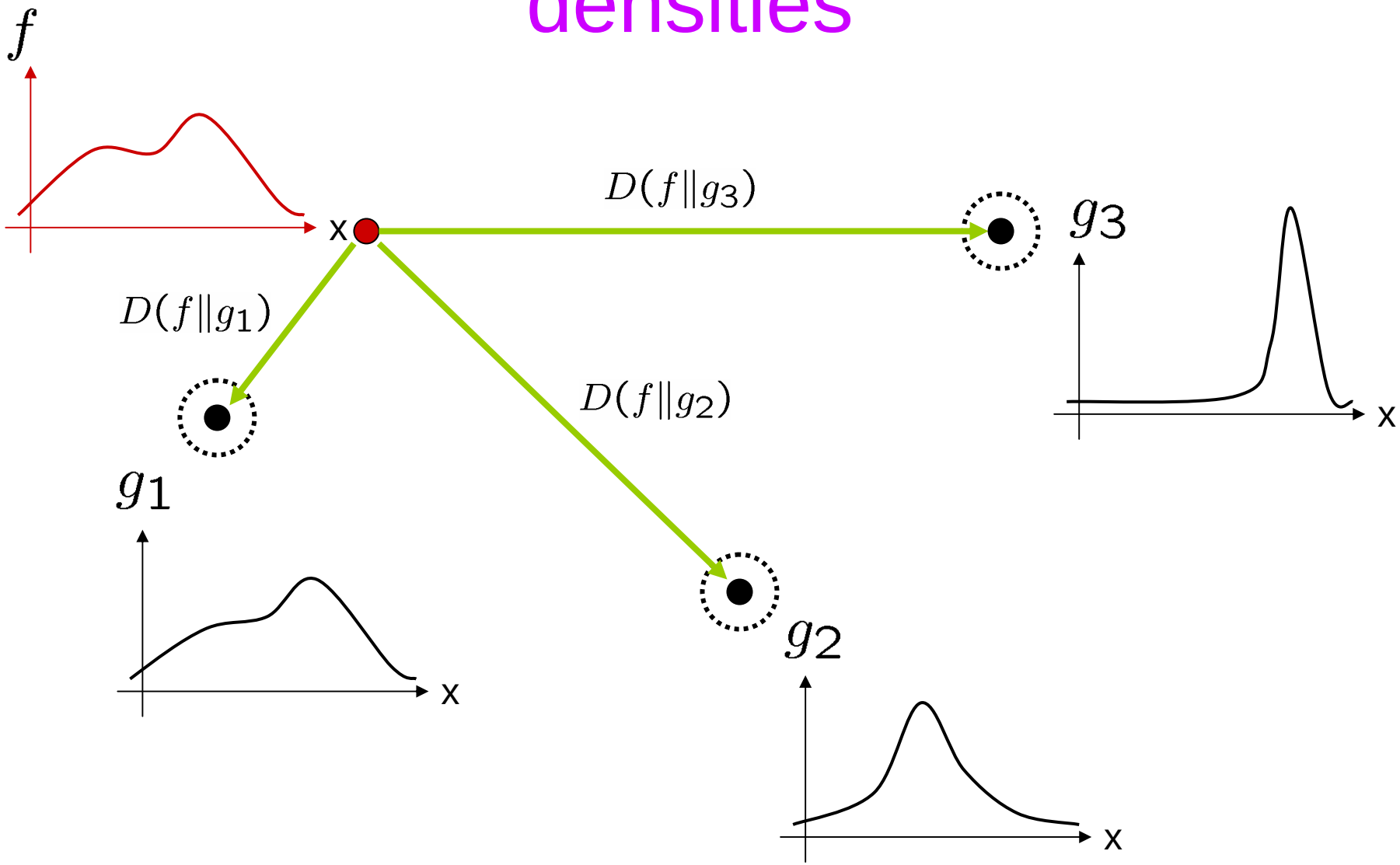
Adaptive scheduling of measurements



Clustering and classification

Characterizing image manifolds

Divergence between feature densities



Divergence functions

- A good measure of dissimilarity between densities f,g of random vector X might have the following properties:

1. $D(f\|g) \geq 0$, with = iff f=g
2. Convexity in f,g
3. Induce a Riemannian metric on smooth family $\mathcal{F}_\Theta$ of densities
4. Be invariant to reparameterization of data X
5. Be closely related to classification performance

Note: standard Euclidean l2 metric does not satisfy 4 or 5

$$D(f\|g) = \int (f(x) - g(x))^2 dx$$

Alpha-Divergence

- α–divergence between densities f,g

$$D_\alpha(f\|g) = \frac{1}{\alpha - 1} \ln \int f^\alpha g^{1-\alpha}$$

- Special cases:
  - the *Kullback-Liebler* and *Hellinger-type*

$$\lim_{\alpha \to 1} D_\alpha(f\|g) = \int f \ln\left(\frac{f}{g}\right), \quad D_{1/2}(f\|g) = -2 \ln \int \sqrt{fg},$$

  - the α-entropy

$$H_\alpha(f) = D_\alpha(f\|U) = \frac{1}{\alpha - 1} \ln \int f^\alpha$$

Relation to Classification

- Given i.i.d. sample $X_1, \ldots, X_n$, test

$$H_0 : X_i \sim g$$
$$H_1 : X_i \sim h$$

- Sanov-Chernoff bound

$$-\ln P(\text{decide } h|g) \leq \max_{\alpha}(1-\alpha)D_\alpha(h\|g)$$

- "=" attained by LRT and large n

Ref: Dembo&Zeitouni (1993)

Divergence Applications

- Finding “distance preserving projections” for dimension reduction with classification constraints (Random projections)
- Finding divergence (MI) maximizing transformations (Image registration)
- Finding divergence (MI) minimizing transformations (ICA)
- Estimating dimension of support set M of density f(x) (Manifold learning)

Entropy and dimension

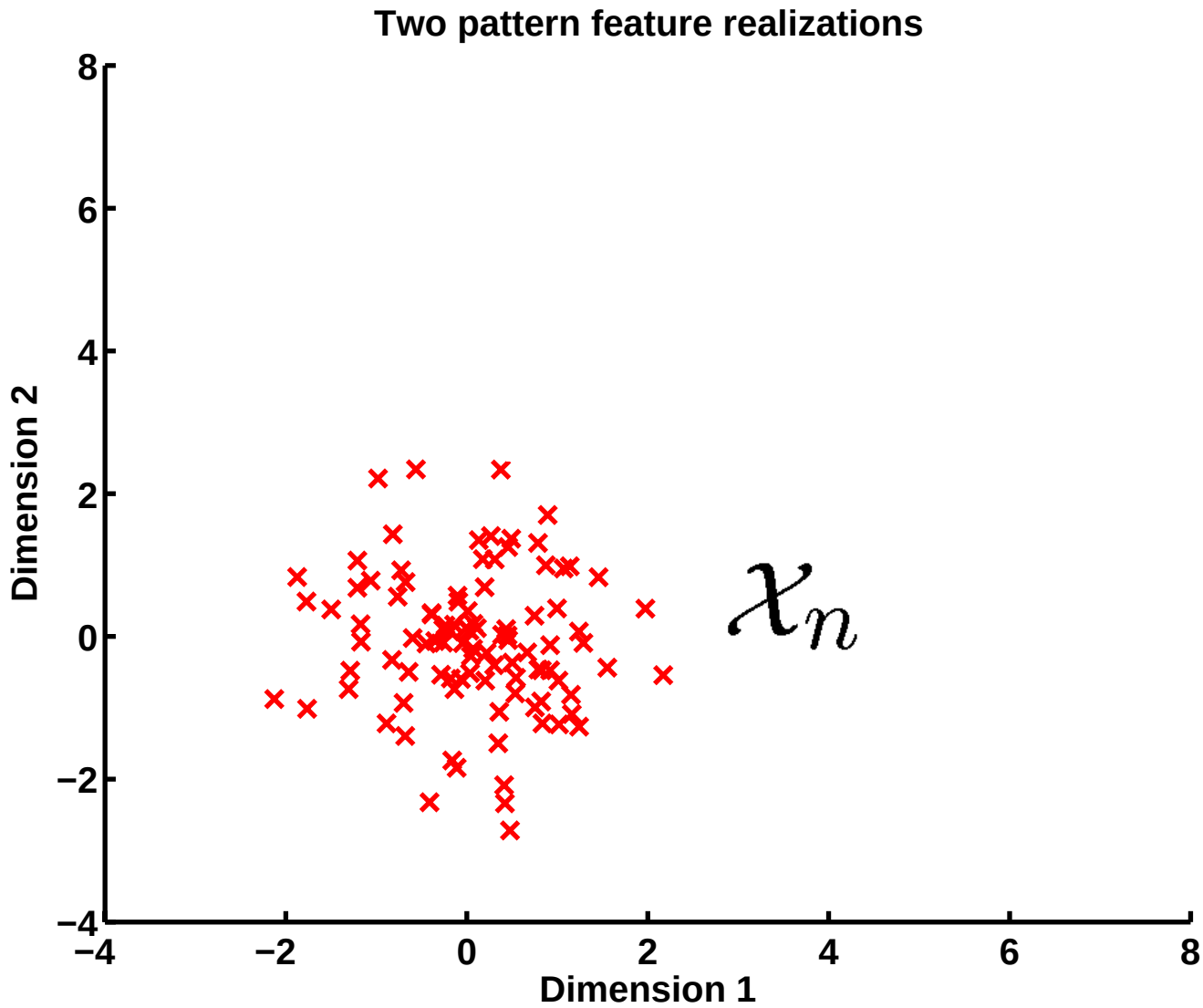
- Consider the entropy integral of $f(x),\ x \in \mathcal{M} \subset \mathcal{R}^D$

$$I(S) = \int_{\mathcal{S}} f^{\alpha}(x) dx = D_{\alpha}(f \| U_{\mathcal{S}})$$

  - Extrinsic entropy when: $\mathcal{S} = \mathcal{R}^D$
  - Intrinsic entropy when: $\mathcal{S} = \mathcal{M}$

- Key observation:
  - if $\dim \mathcal{S} > \dim \mathcal{M}$ then $I(S) = 0$

But, only realizations of X available



Key Issue: Divergence Estimation

Random sample: $\mathcal{X}_n = \{X_1, \ldots, X_n\}$

- Indirect estimators
  - Parametric estimation-substitution methods

$$\widehat{D}_\alpha(f\|g) = D_\alpha(f_{\widehat{\theta}}\|g)$$

  - Non-parametric density estimation-substitution

$$\widehat{D}_\alpha(f\|g) = D_\alpha(\widehat{f}\|g)$$

- Direct estimators
  - Spacing and hyper-spacing estimators (Vasicek:76, Miller:03)
  - Entropic graph estimators (Hero&etal:SPM02)

$$\widehat{D}_\alpha(f\|g) = L(T[\mathcal{X}_n])/n^\alpha$$

2. Minimal graphs

- $\mathcal{Y}_n = \{Y_1, \ldots, Y_n\}$ data in d-dimensional Euclidean space
- Euclidean MST with edge power weighting gamma:

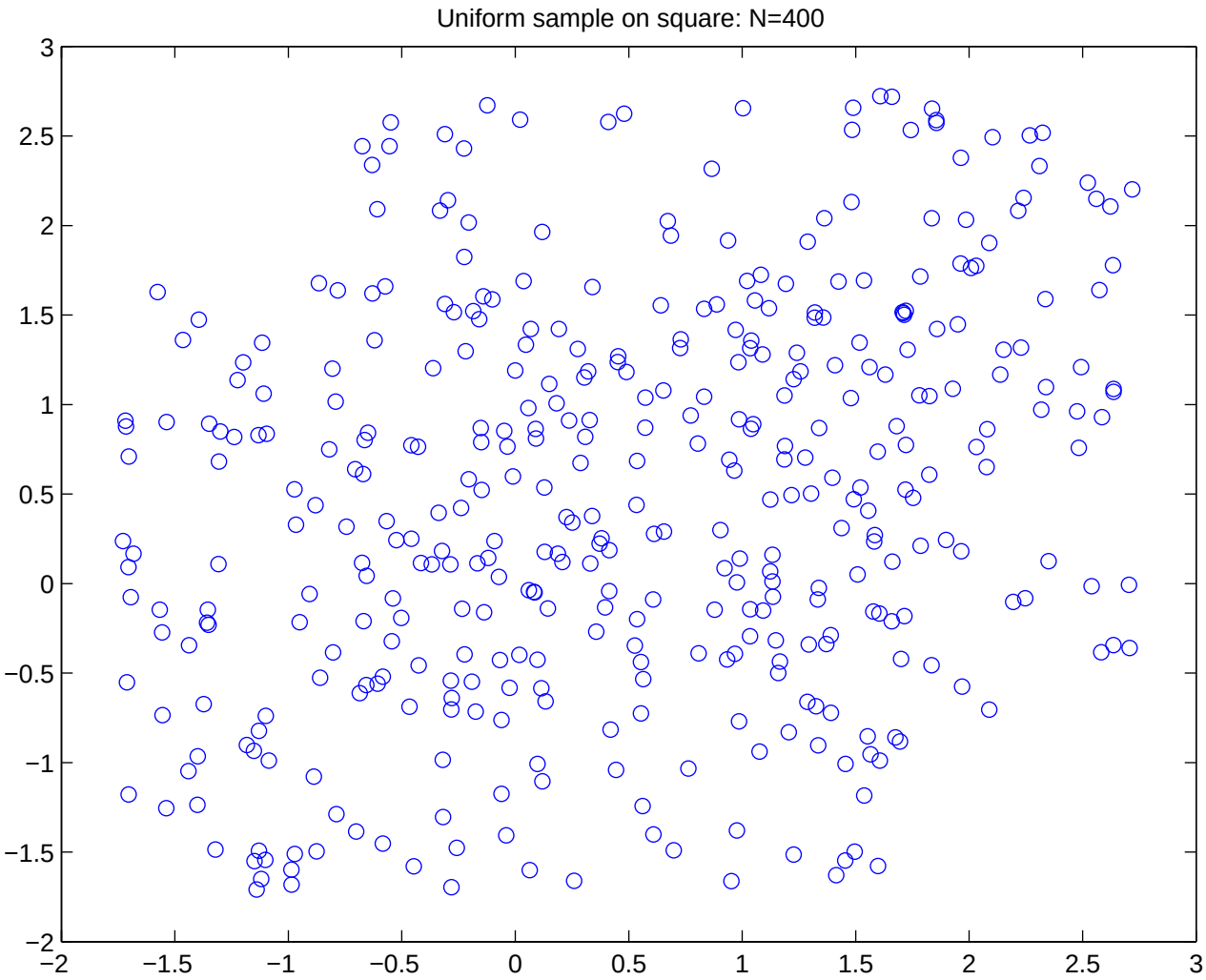
$$L_\gamma^{\mathbf{R}^m}(\mathcal{Y}_n) = \min_{E \in \mathcal{E}} \sum_{|e| \in E} |e|^\gamma$$

- $\mathcal{E}$ pairwise distance matrix over $\mathcal{Y}_n$
- E edge length matrix of spanning trees over $\mathcal{Y}_n$

- Euclidean k-NNG with edge power weighting gamma:

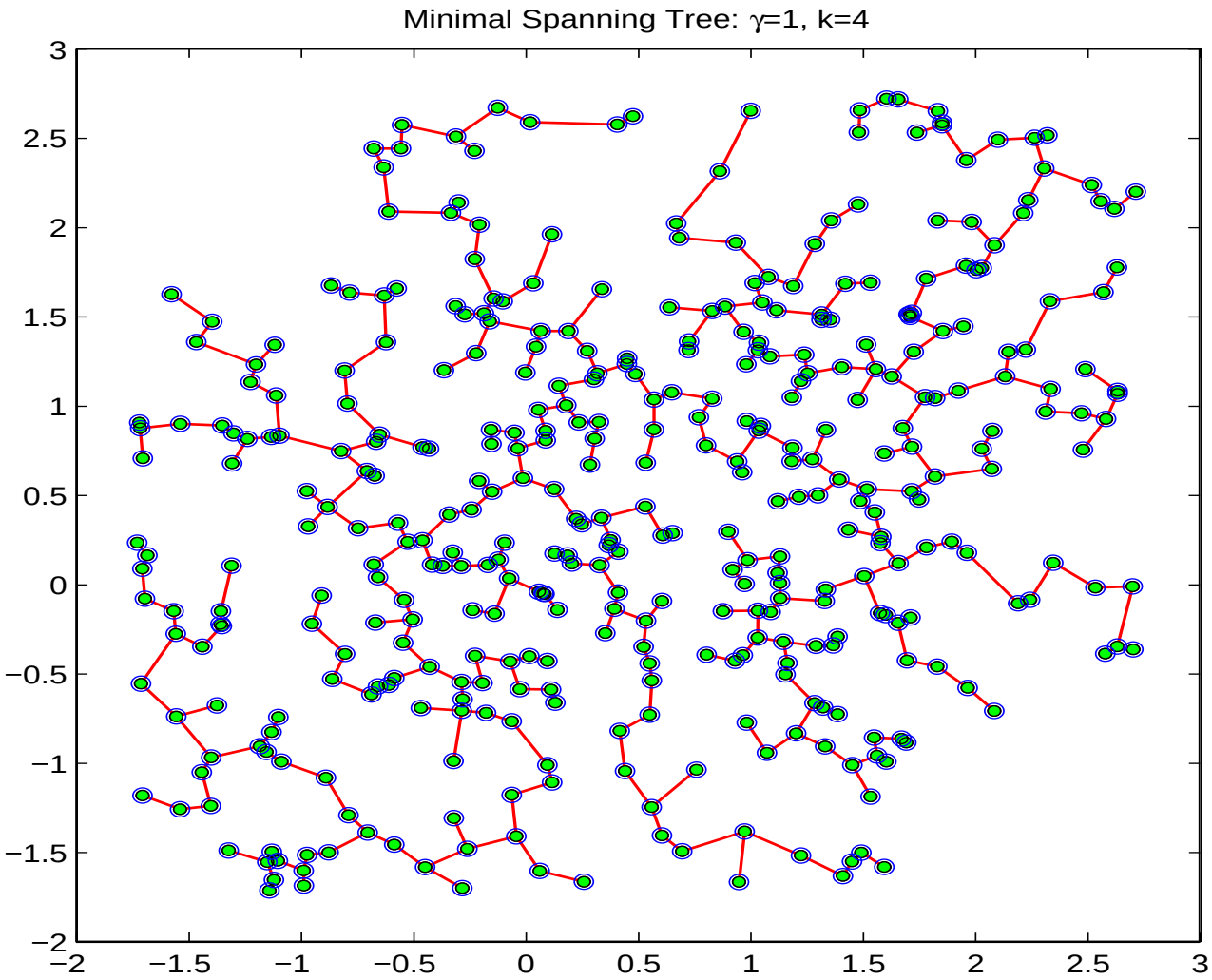
$$\mathcal{L}_{k,\gamma}^{\mathbf{R}^m}(\mathcal{Y}_n) = \sum_{i=1}^{n} \sum_{|e| \in E_k(Y_i)} |e|^\gamma$$

Example: Uniform Planar Sample



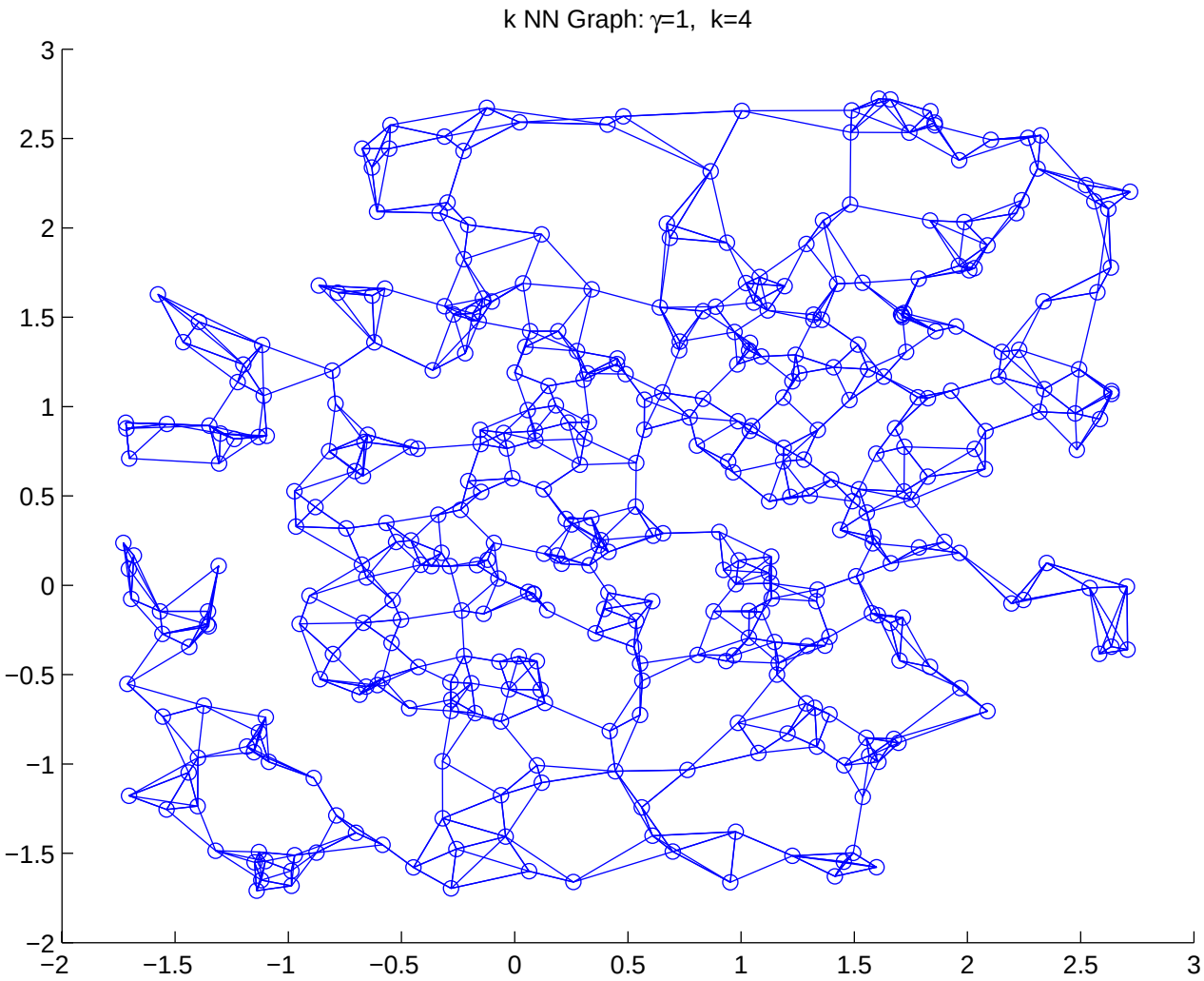
Feature dimension $d = 2$

Example: MST on Planar Sample



$$d = 2, \quad \gamma = 1$$

Example: k-NNG on Planar Sample



$$d = 2, \quad \gamma = 1, k = 4$$

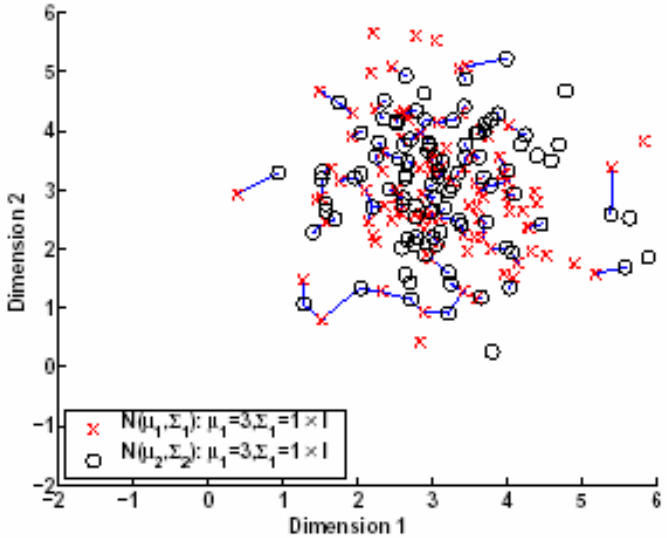
Graphs and divergence

- Simple functionals of minimal graphs can sometimes be found that estimate divergences
- When this is possible we call these “entropic graph estimators”
- Example: Henze-Penrose divergence between two densities f_X, f_Y

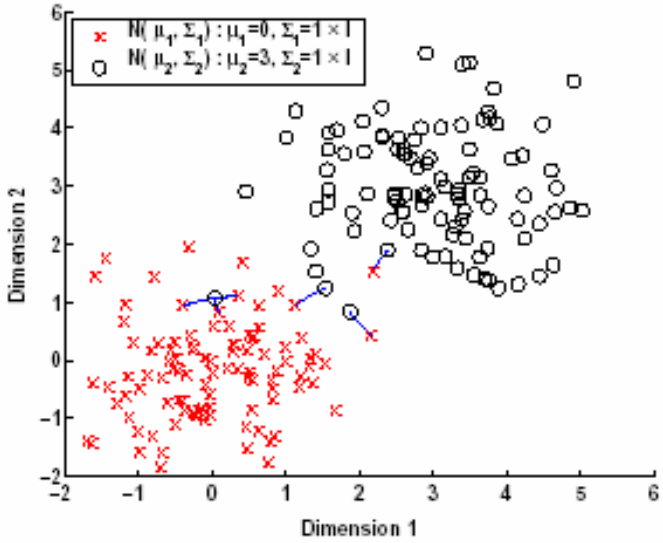
$$D^{hp}(f\|g) = 1 - 2\epsilon(1-\epsilon)\int \frac{f_X f_Y}{\epsilon f_X + (1-\epsilon) f_Y}$$

  - Friedman-Rafsky’s MST “coincident edge counter” is a direct entropic graph estimator

Friedman-Rafsky Entropic Graph



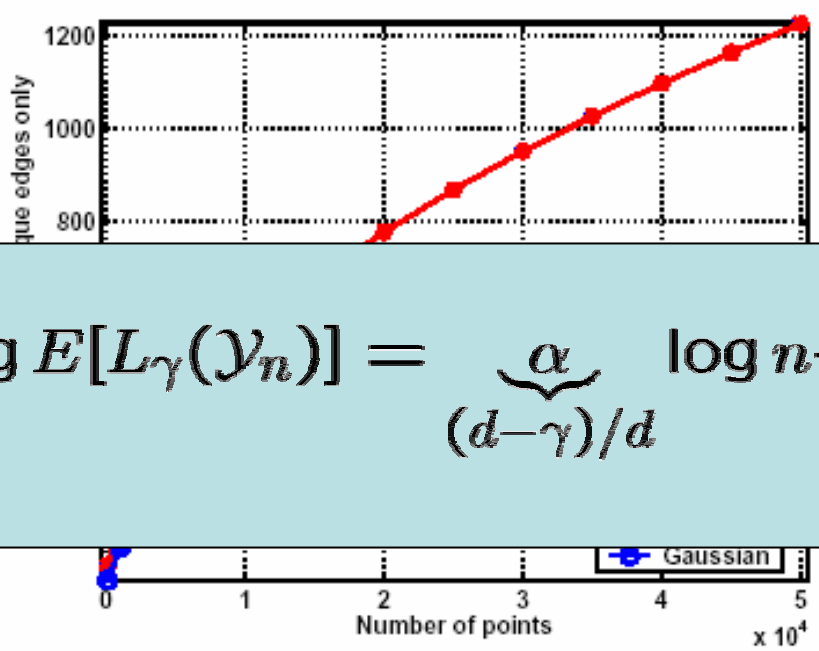
MST

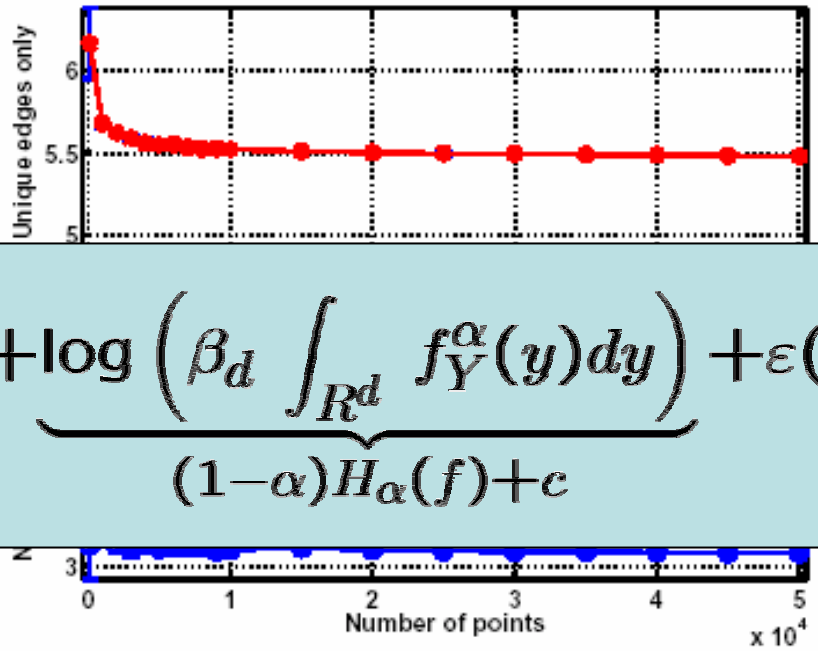


$$\Delta L_1(\mathcal{X}_m \cup \mathcal{Y}_n)/(m+n) \to c \int \frac{f_X f_Y}{\epsilon f_X + (1-\epsilon) f_Y}$$

Henze&Penrose:99

Convergence of minimal graph length functional





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

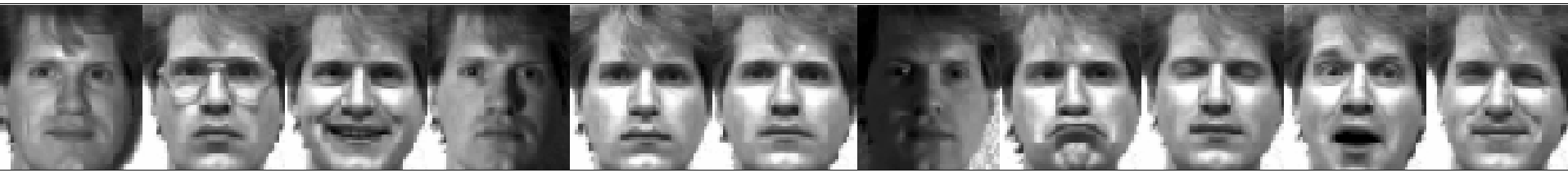
Beardwood, Halton, Hammersley Theorem (BHH:1959):

$$L_\gamma(\mathcal{Y}_n)/n^\alpha \rightarrow \beta_d \int_{R^d} f_Y^\alpha(y)dy$$

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

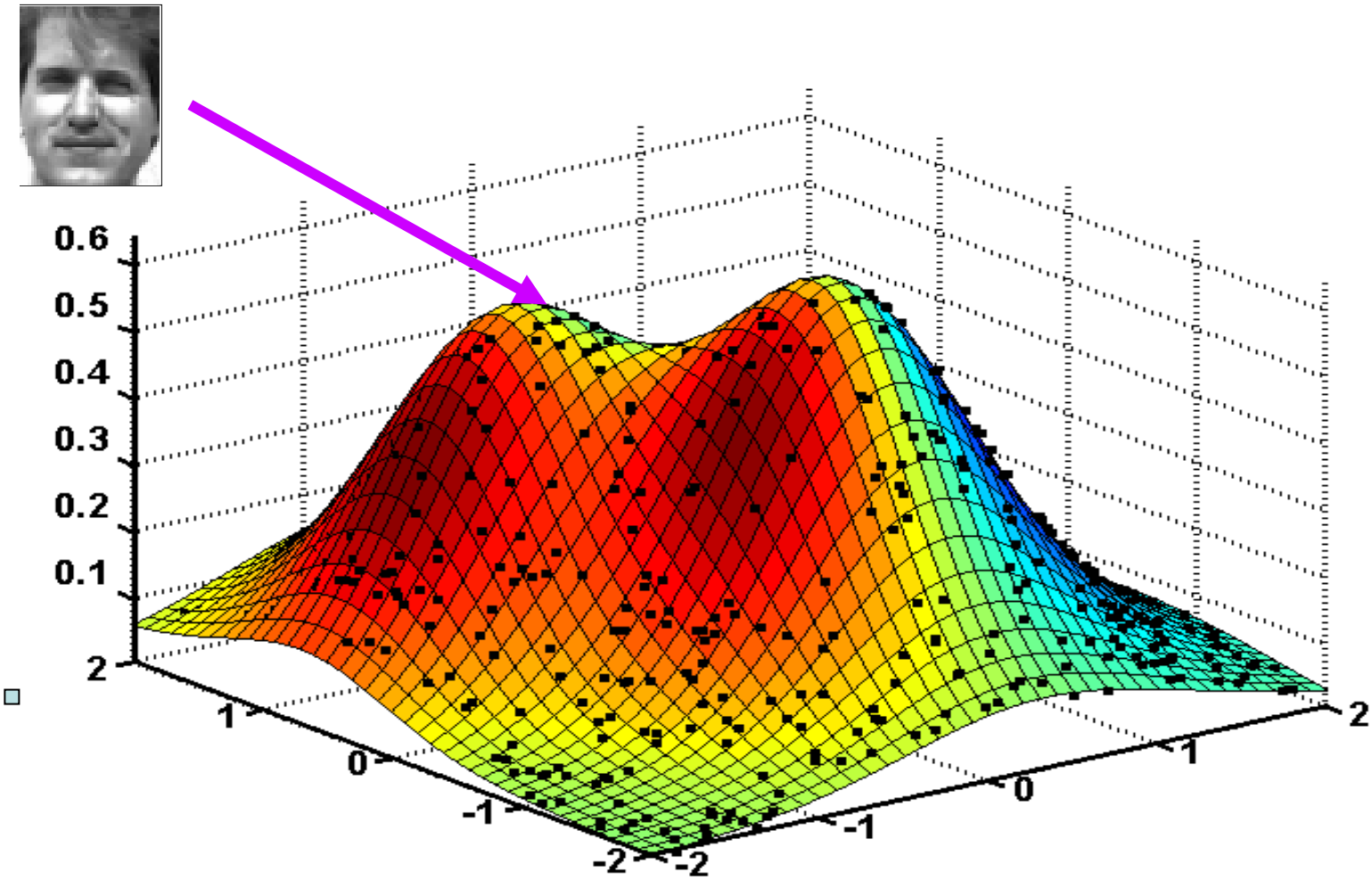
How to apply BHH Thm to non-linear dimension reduction?

- 128x128 images of faces
- Different poses, illuminations, facial expressions

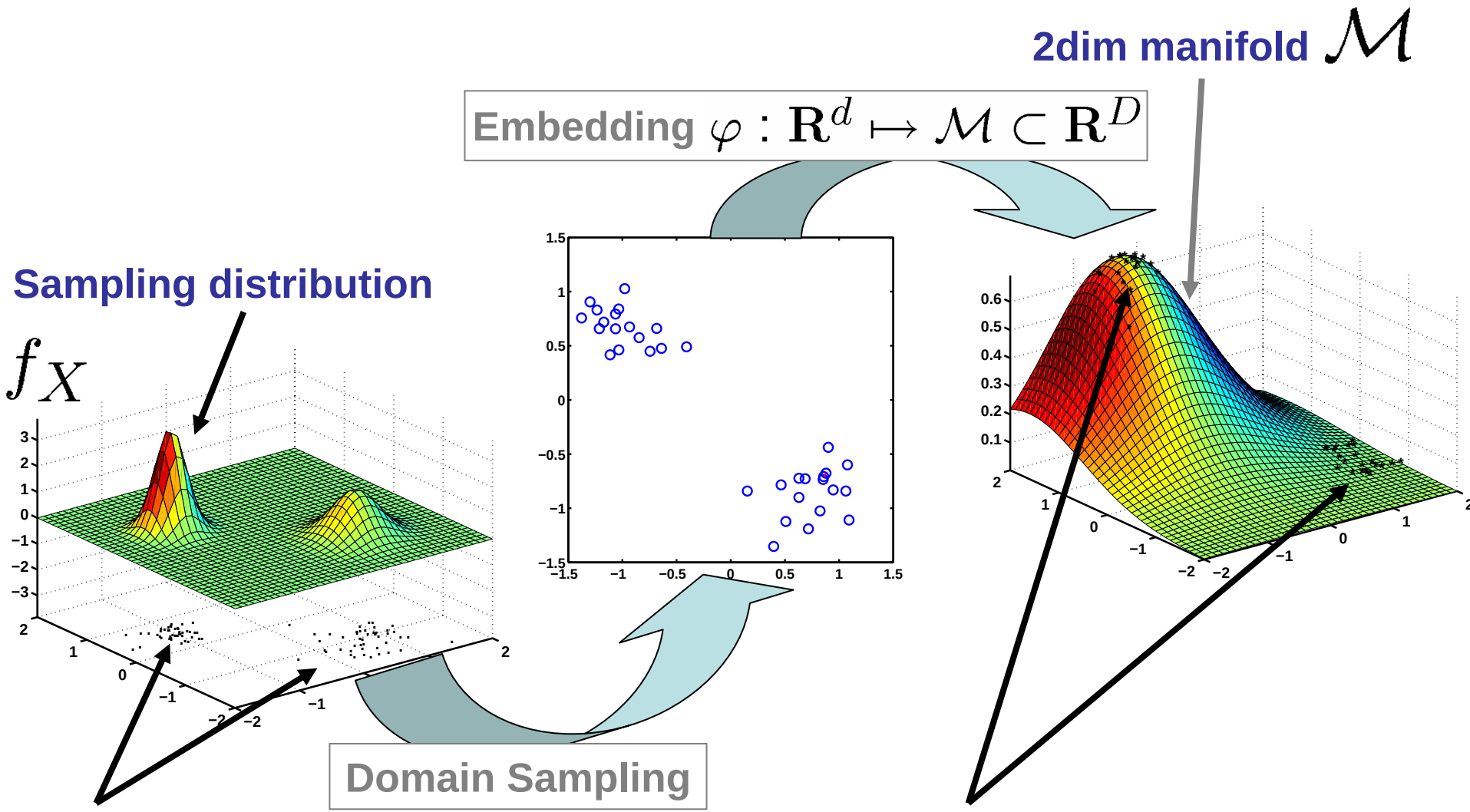


- The set of all face images evolve on a lower dimensional imbedded manifold in R^(16384)

Embedded Manifold



Generative model



A statistical sample

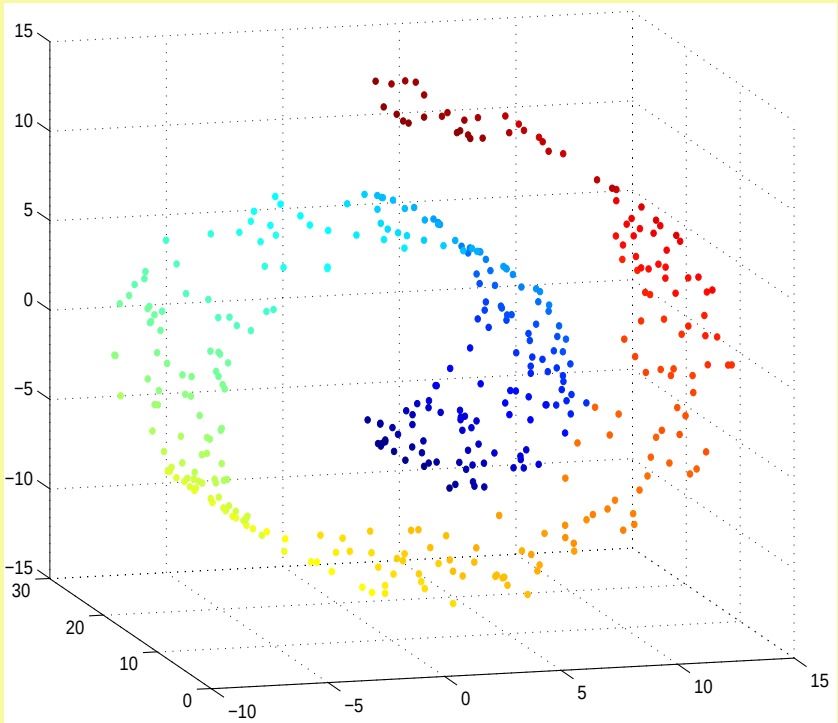
$\mathcal{X}_n = \{X_1, \ldots, X_n\}$

Observed sample

$\mathcal{Y}_n = \{Y_1, \ldots, Y_n\}$

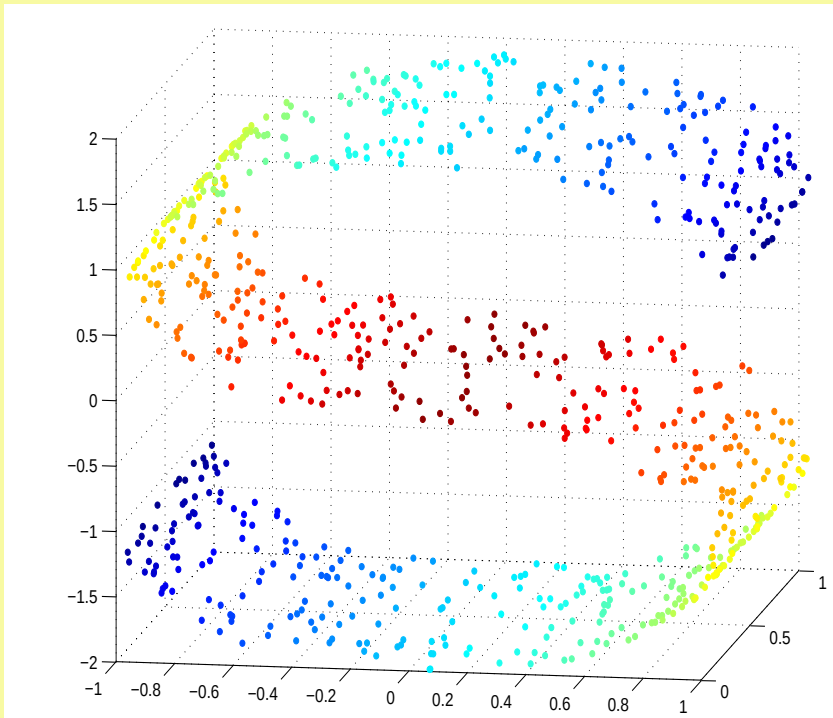
2D Manifolds in 3D

Ref: Tenenbaum&etal (2000)



Swiss Roll N=400

Ref: Roweiss&etal (2000)

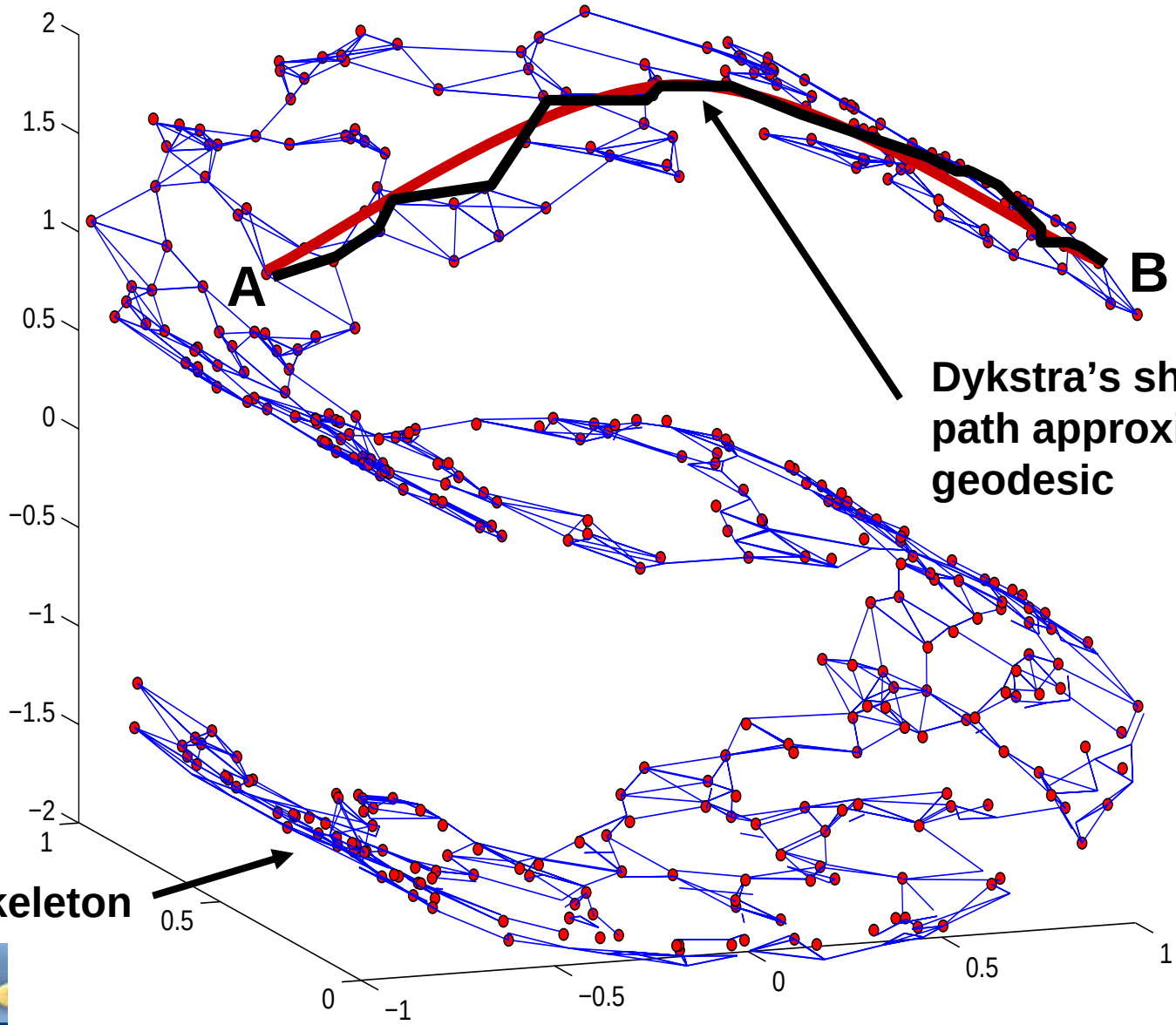


S-Curve N=800

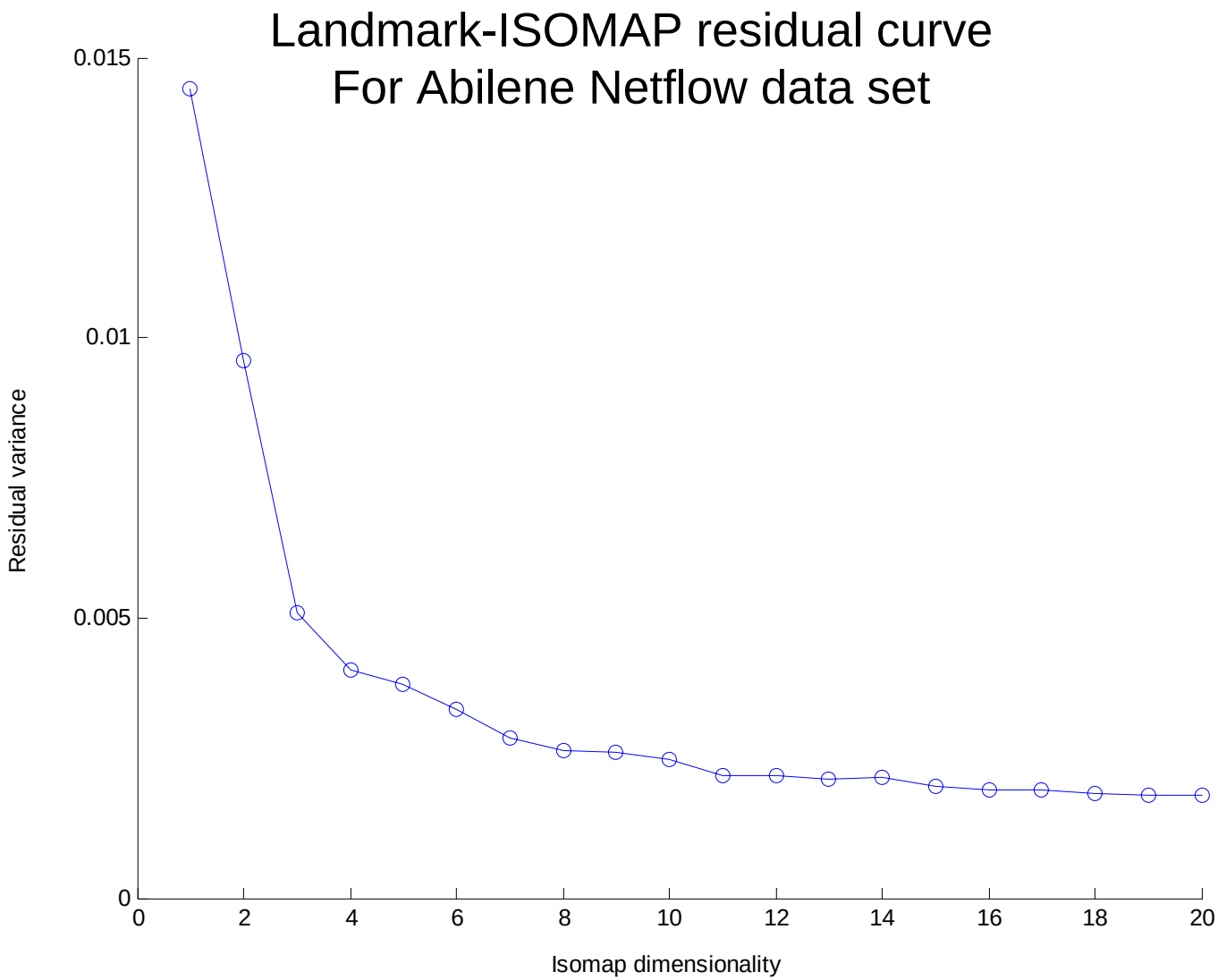
$$D = 3,\ d = 2$$

- Sampling density fy = Uniform on manifold

ISOMAP Geodesic Approximation



How to Estimate d?



Geodesic MST

- Construct MST using geodesic distance estimates $\mathcal{E}^G$ and obtain length functional

$$L_\gamma^{\mathcal{M}}(\mathcal{Y}_n)$$

- What is limiting behavior

$$L_\gamma^{\mathcal{M}}(\mathcal{Y}_n)/n^\alpha \rightarrow ?$$

- where $\alpha = (d' - \gamma)/d'$?

GMST Convergence Theorem

Let $\mathcal{M}$ be a smooth d-dimensional manifold embedded in $\mathbf{R}^D$ through a smooth map $\varphi : \mathcal{D} \mapsto \mathcal{M}$ where $\mathcal{D}$ is an open convex subset of $\mathbf{R}^d$. Let $2 \leq d \leq D$ and $0 < \gamma < d$. Suppose that $Y_1, \ldots, Y_n$ are i.i.d. random vectors on $\mathcal{M}$ with common density f_Y w.r.t. Lebesgue measure $\mu_{\mathcal{M}}$ on $\mathcal{M}$. Then the total length of the GMST satisfies

$$L_\gamma^{\mathcal{M}}(\mathcal{Y}_n)/n^{(d'-\gamma)/d'} \to \begin{cases} \infty, & d' < d \\ \beta_d c^{-\gamma} \int_{\mathcal{M}} f_Y^\alpha(y)\, \mu_{\mathcal{M}}(dy), & d' = d \\ 0, & d' > d \end{cases},$$

(a.s) where $\alpha = (d-\gamma)/d$, $c = \left(\sqrt{\det\left(\boldsymbol{J}_\varphi^T \boldsymbol{J}_\varphi\right)}\right)^{1/d}$.

Costa, Hero: Trans SP (2003)

k-NNG Convergence Theorem

Let $\mathcal{M}$ be a d-dimensional manifold embedded in $\mathbf{R}^D$. Let $2 \leq d \leq D$ and $0 < \gamma < d$. Suppose that $Y_1, \ldots, Y_n$ are i.i.d. random vectors on $\mathcal{M}$ with common density f_Y w.r.t. Lebesgue measure $\mu_{\mathcal{M}}$ on $\mathcal{M}$. Then the total length of the k-NNG satisfies

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

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

Costa, Hero: EUSIPCO (2004)

Joint Estimation Algorithm

- Convergence theorem suggests log-log-linear model

$$\log E[L_\gamma^{\mathcal{M}}(\mathcal{Y}_n)] = (d-\gamma)/d \log n + b + v(n)$$

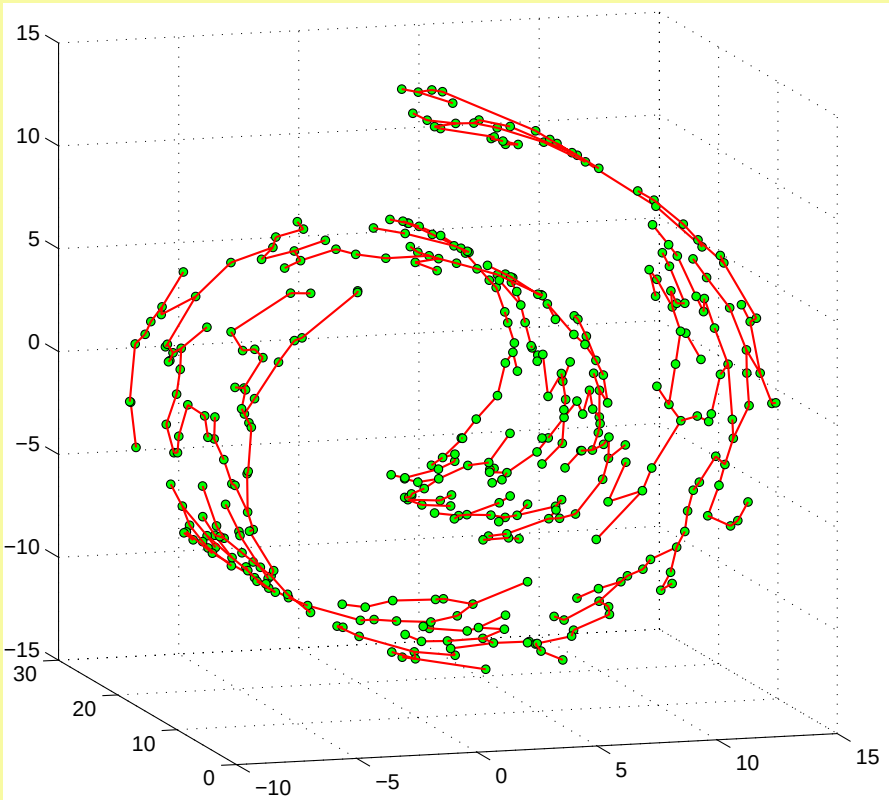
- Use bootstrap resampling to estimate mean graph length and apply LS to jointly estimate slope and intercept from sequence

$$\{\log E[L_\gamma^{\mathcal{M}}(\mathcal{Y}_n)], \log n\}_n$$

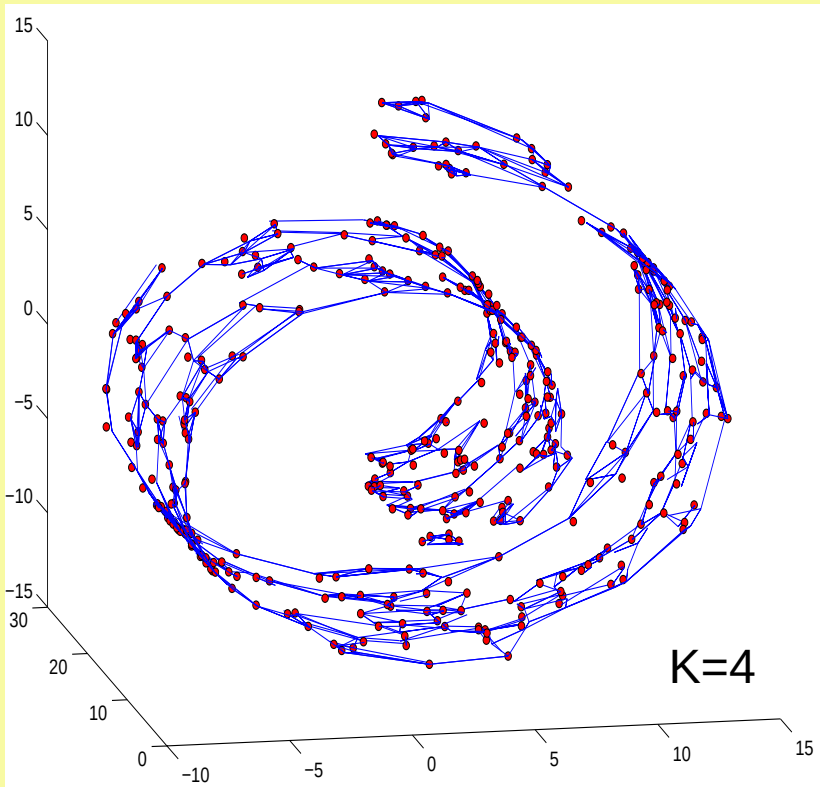
- Extract d and H from slope and intercept

Costa, Hero: Trans SP (2003)

3. Simulation Studies: Swiss Roll



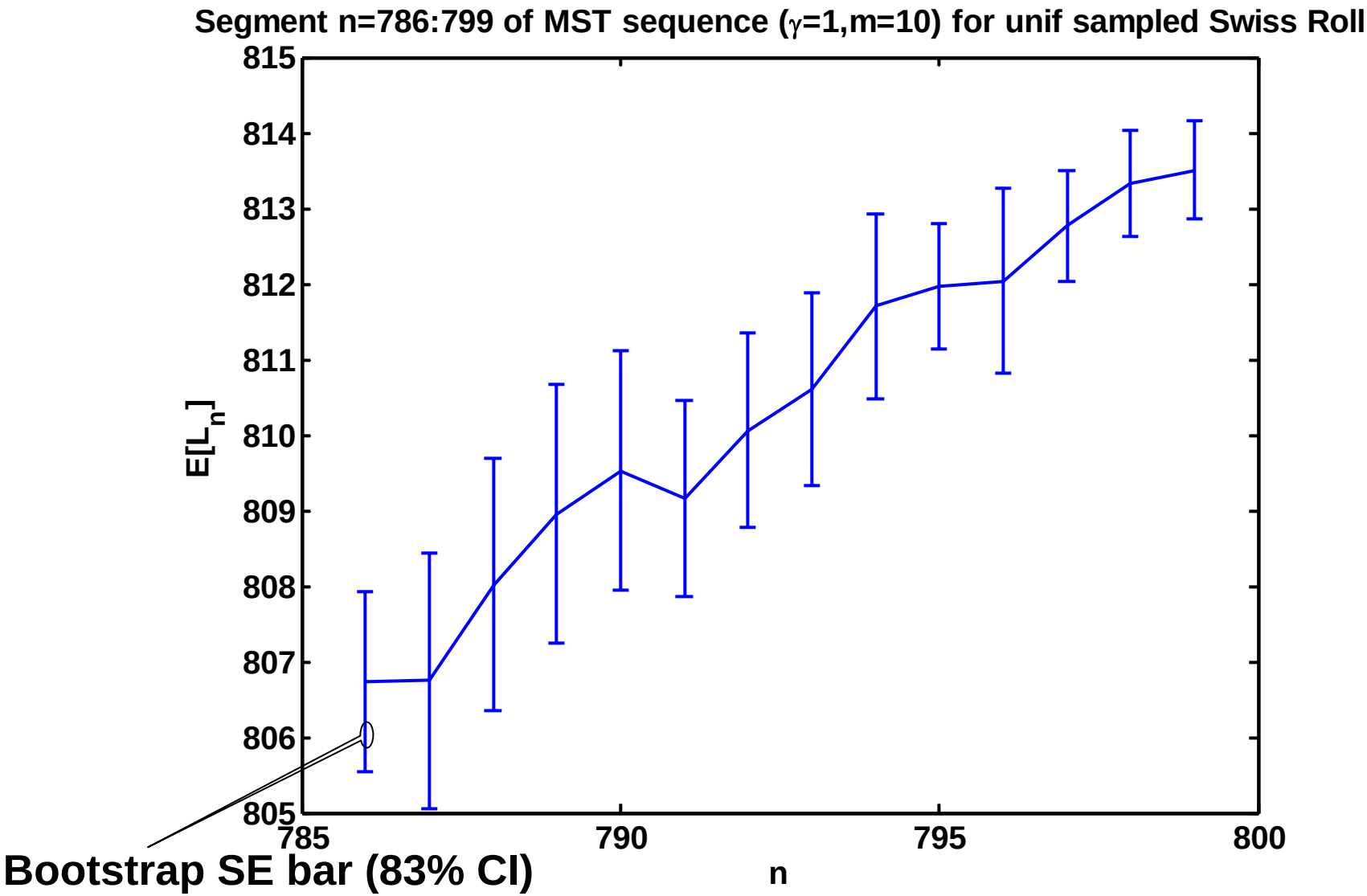
GMST



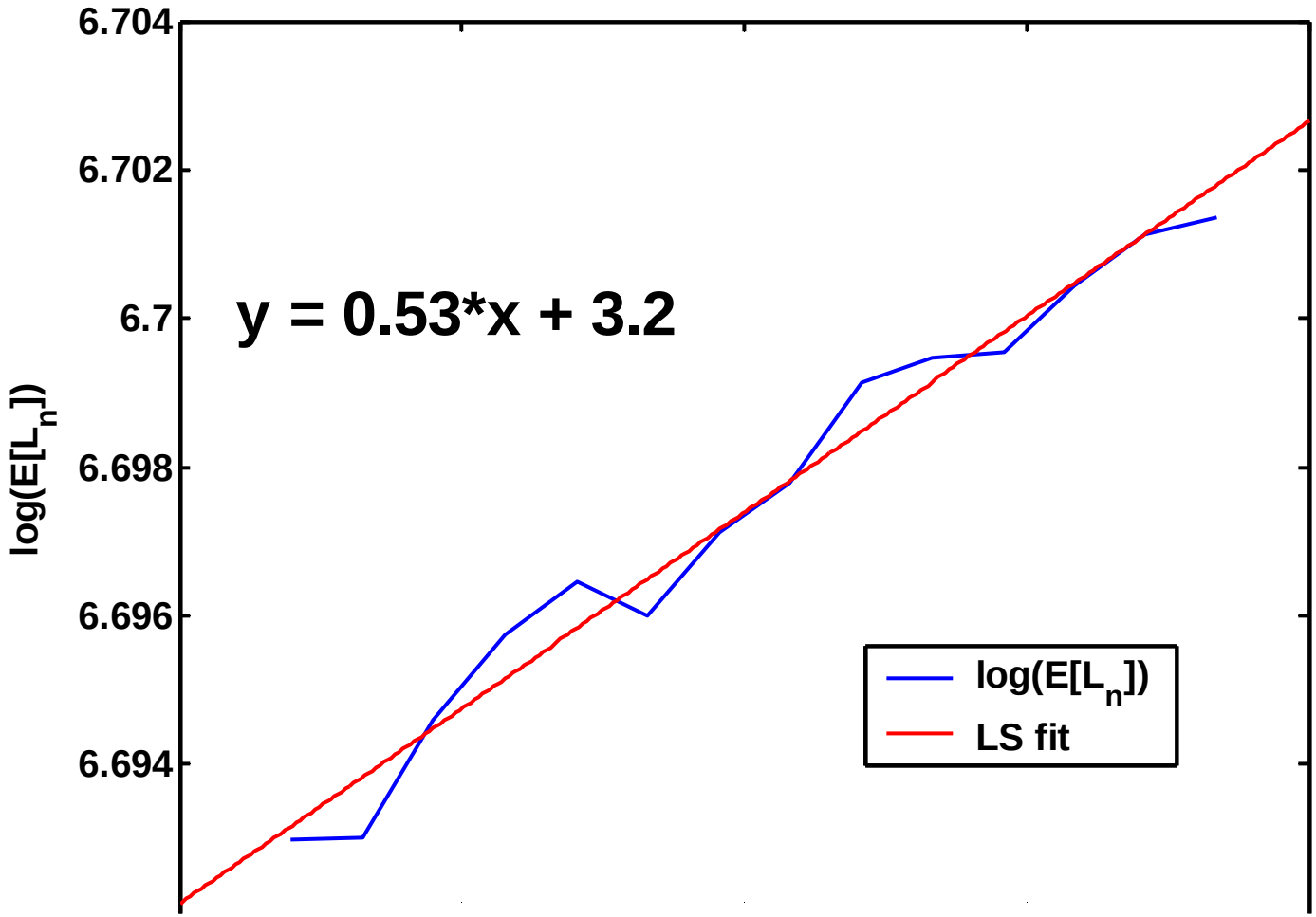
kNN

- n=400, f=Uniform on manifold

Estimates of GMST Length



loglogLinear Fit to GMST Length



GMST Dimension and Entropy Estimates

- From LS fit $y = ax + b$ find:
- Intrinsic dimension estimate

$$\hat{d} = \text{round}\underbrace{\left(\frac{\gamma}{1-a}\right)}_{2.1} = 2$$

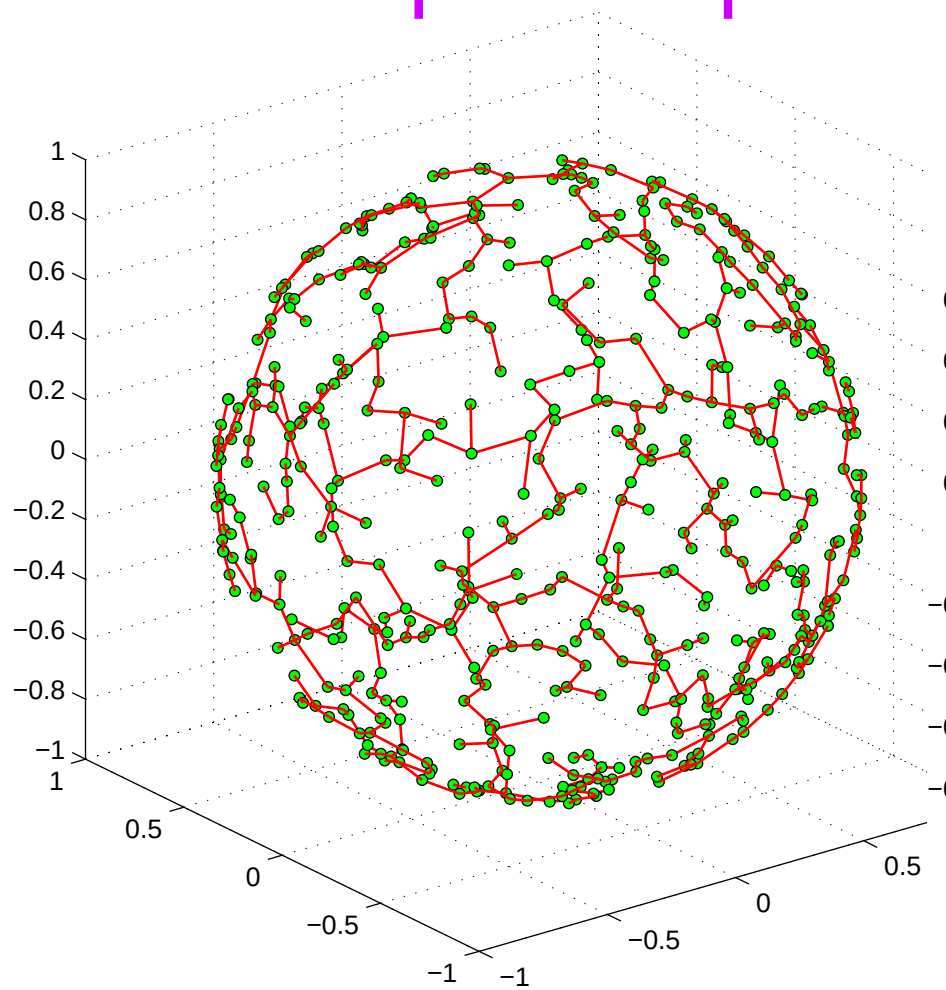
- Alpha-entropy estimate ($\alpha = a = 0.53$)

$$\hat{H}_\alpha(f_Y) = \frac{b - \gamma/2 \log \beta_{\hat{d}}}{1-a} = 7.3$$

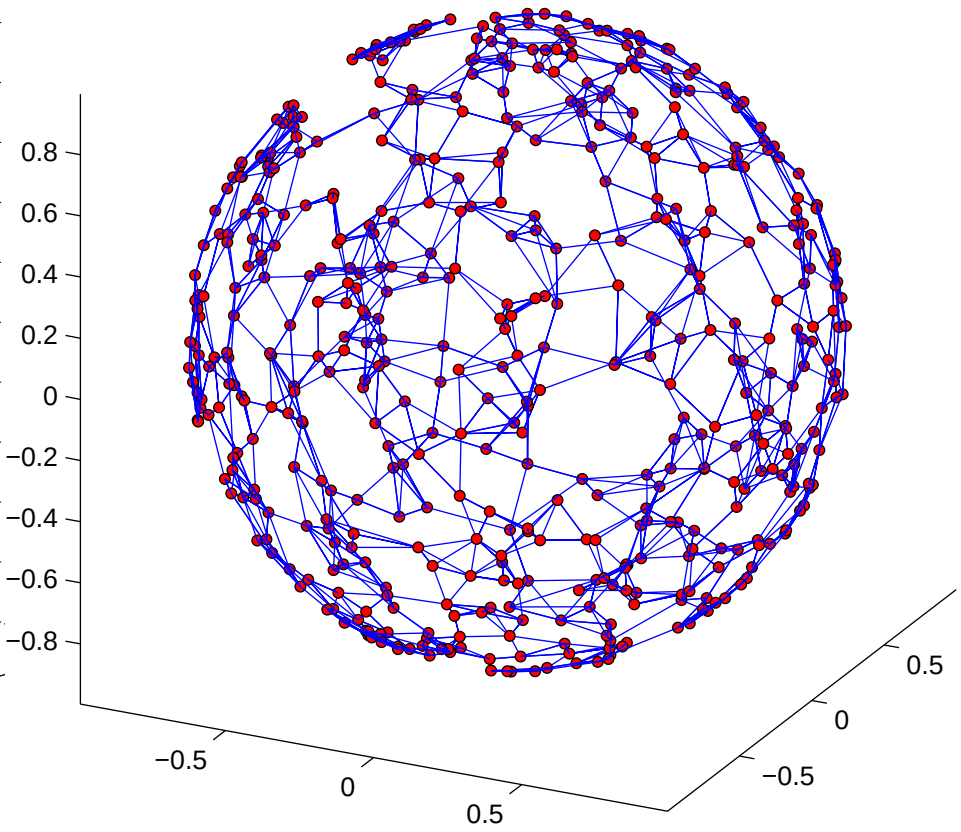
  - Ground truth:

$$H_\alpha(f_Y) = \log(1869) = 7.53$$

Entropic Graphs on S2 Sphere in 3D



GMST

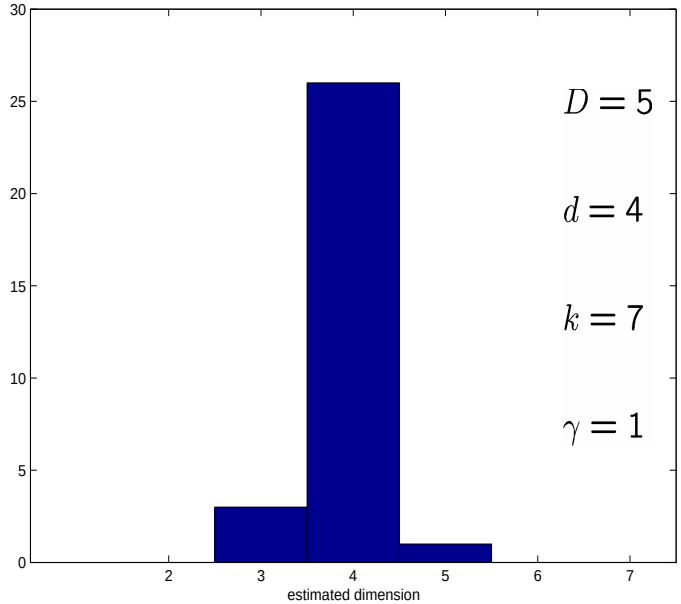


kNN

- n=500, f=Uniform on manifold

k-NNG on Sphere S4 in 5D

- **k=7 for all algorithms**
- **kNN resampled 5 times**
- **Length regressed on 10 or 20 samples at end of mean length sequence**
- **30 experiments performed**
- **ISOMAP always estimates d=5**



Histogram of resampled d-estimates of k-NNG

N=1000 points uniformly distributed on S4 (sphere) in 5D

n	600	800	1000	1200
ISOMAP	0/30	0/30	0/30	0/30
k-NNG-10	23/30	26/30	26/30	26/30
k-NNG-20	29/30	30/30	30/30	30/30

Table of relative frequencies of correct d estimate

Improve Performance by Bootstrap Resampling

- Main idea: Averaging of weak learners
  - Using fewer (N) samples per MST estimate, generate large number (M) of weak estimates of d and H

$$\{\hat{d}_N^{(j)}, \hat{H}_N^{(j)}\}_{j=1}^{M}$$

  - Reduce bias by averaging these estimates (M>>1,N=1)
  - Better than optimizing estimate of MST length (M=1,N>>1)

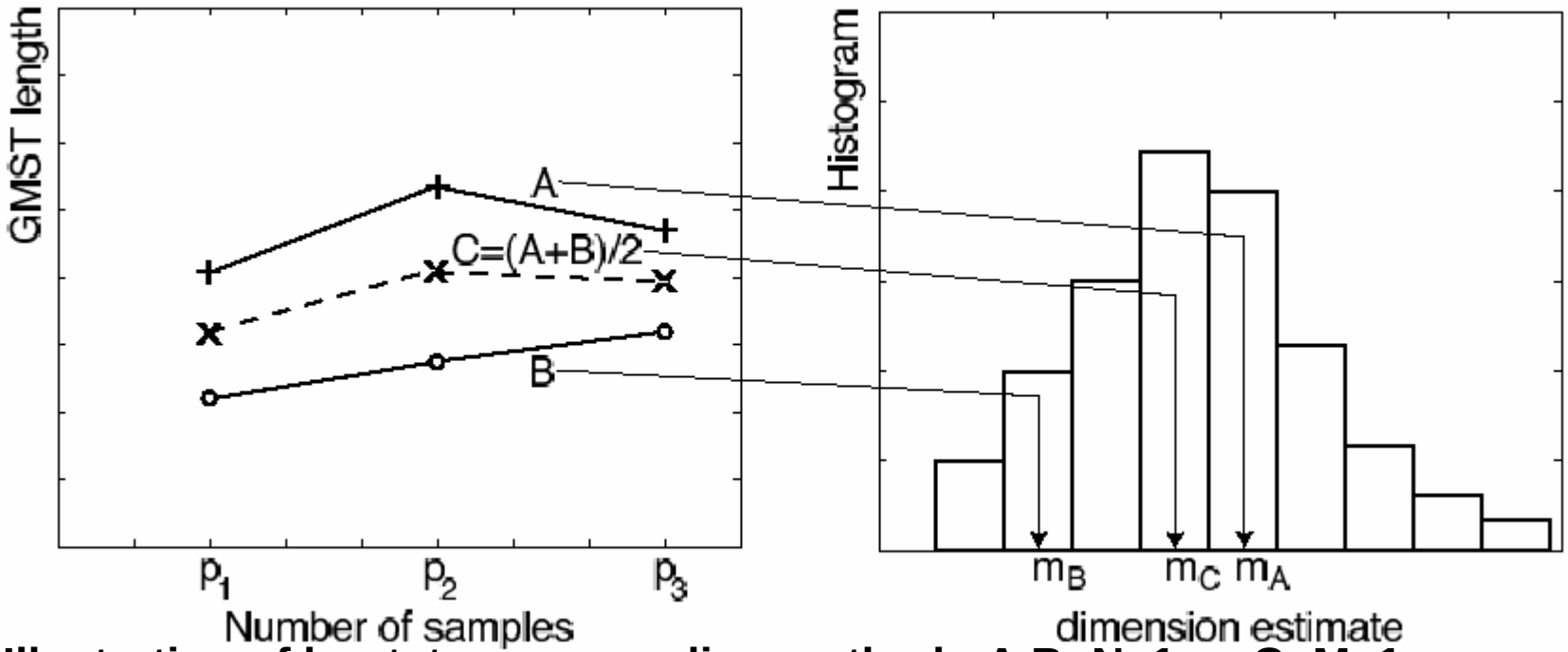


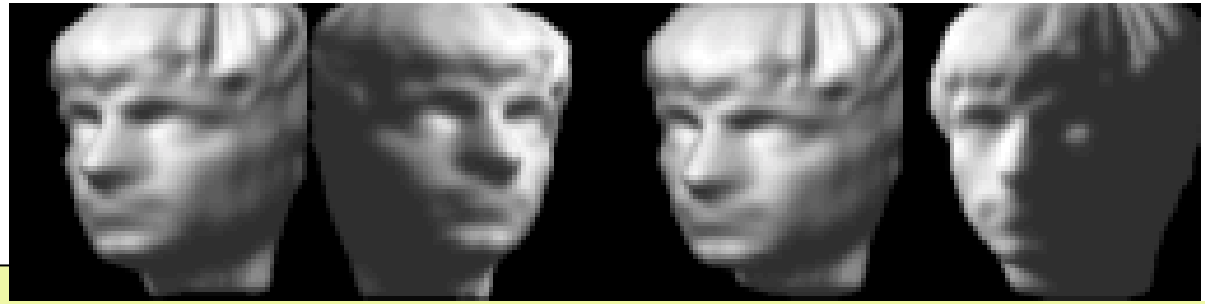
Illustration of bootstrap resampling method: A,B: N=1 vs C: M=1

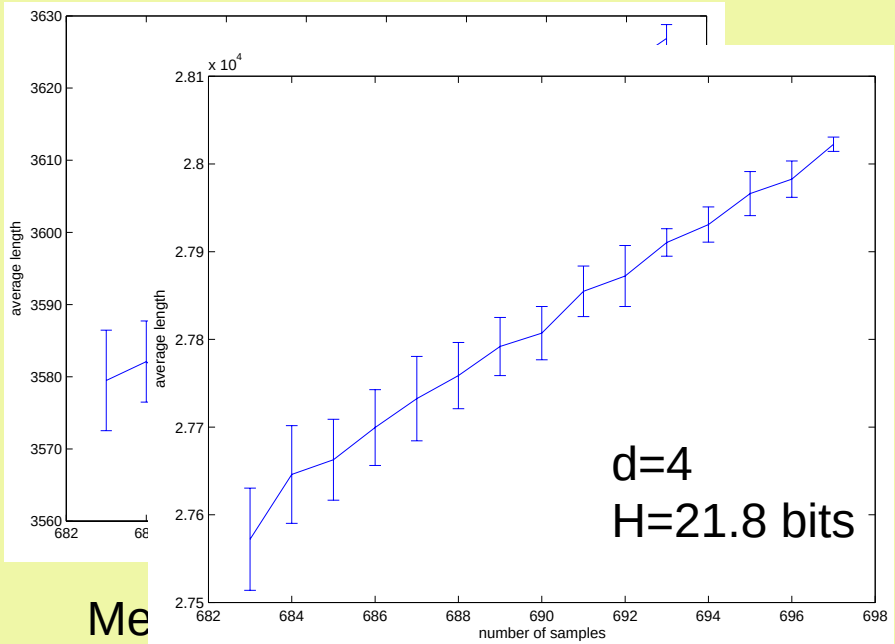
kNN/GMST Comparisons for Uniform Hyperplane

Hyper-plane dimension	Q	M	N	n		
				600	800	1000
2	10	1	5	30	30	30
		5	1	30	30	30
3	10	1	5	24	24	27
		5	1	25	26	27
	15	1	10	30	30	30
		10	1	30	30	30
4	15	1	10	24	25	26
		10	1	27	28	28
	20	1	10	25	28	29
		10	1	29	29	30

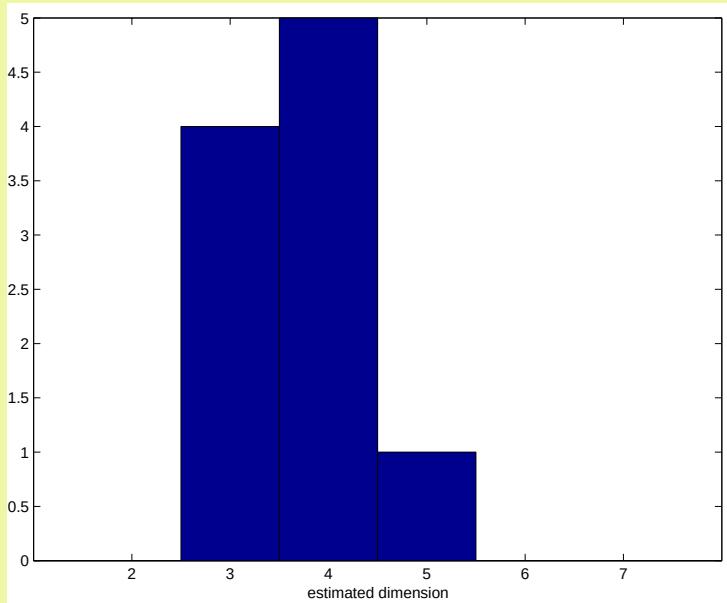
Table of relative frequencies of correct *d* estimate using the GMST, with (*N* = 1) and without (*M* = 1) bias correction.

ISOMAP Database



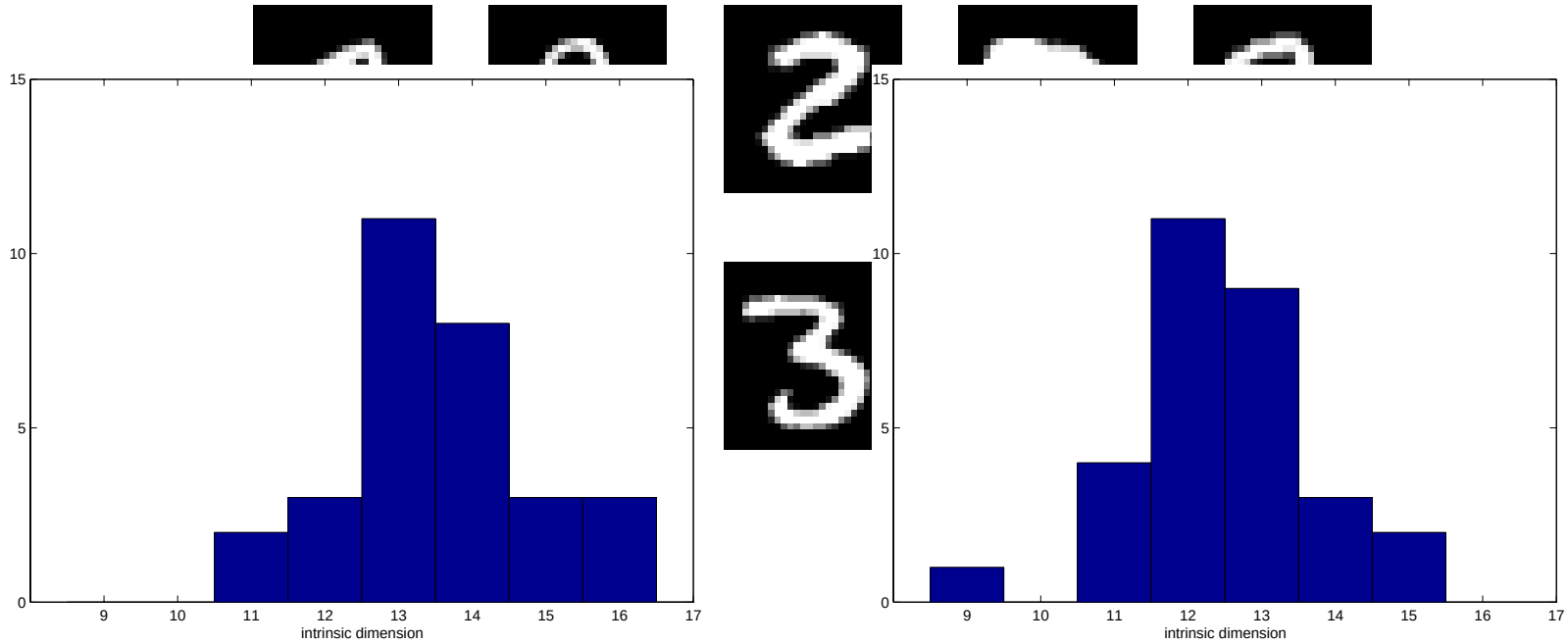


Mean kNNG (k=7) length



Resampling Histogram of d hat

MNIST Digit Database



Histogram of intrinsic dimension estimates: GMST (left) and 5-NN (right) ($M = 10$, $N = 1$, $Q = 15$).

	digit 2	digit 3	digit 2 + 3
GMST	13	12	13
5-NN	12	11	12

Samples from Face database B

- Photographic folios of many people's faces
- Each face folio contains images at 585 different illumination/pose conditions
- Subsampled to 64 by 64 pixels (4096 extrinsic dimensions)

Face 1

Face folio 1

Face 2

Face folio 2

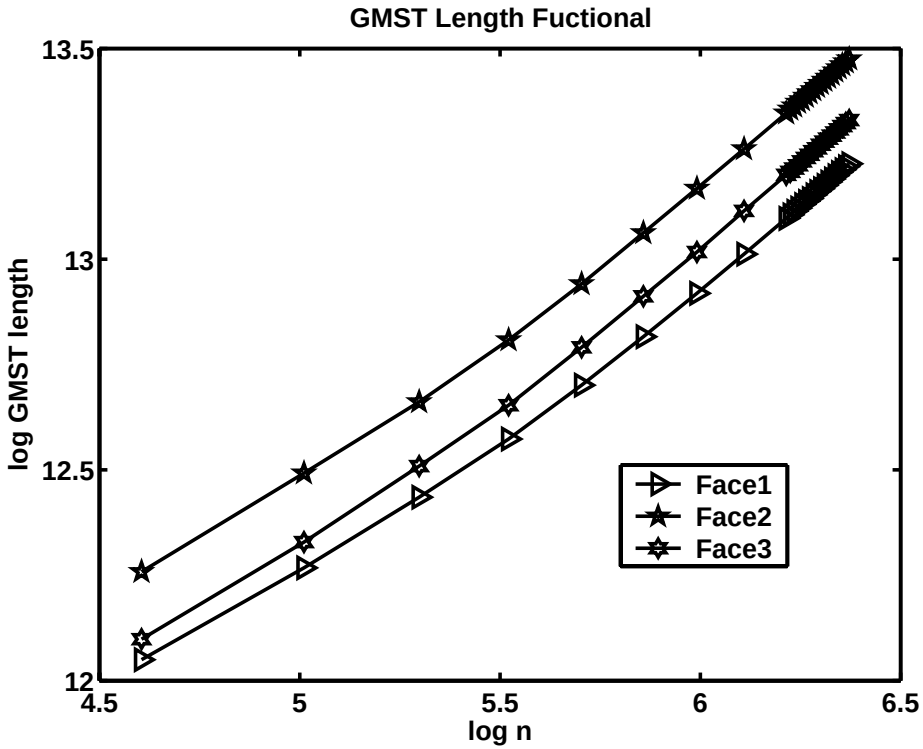
Face 3

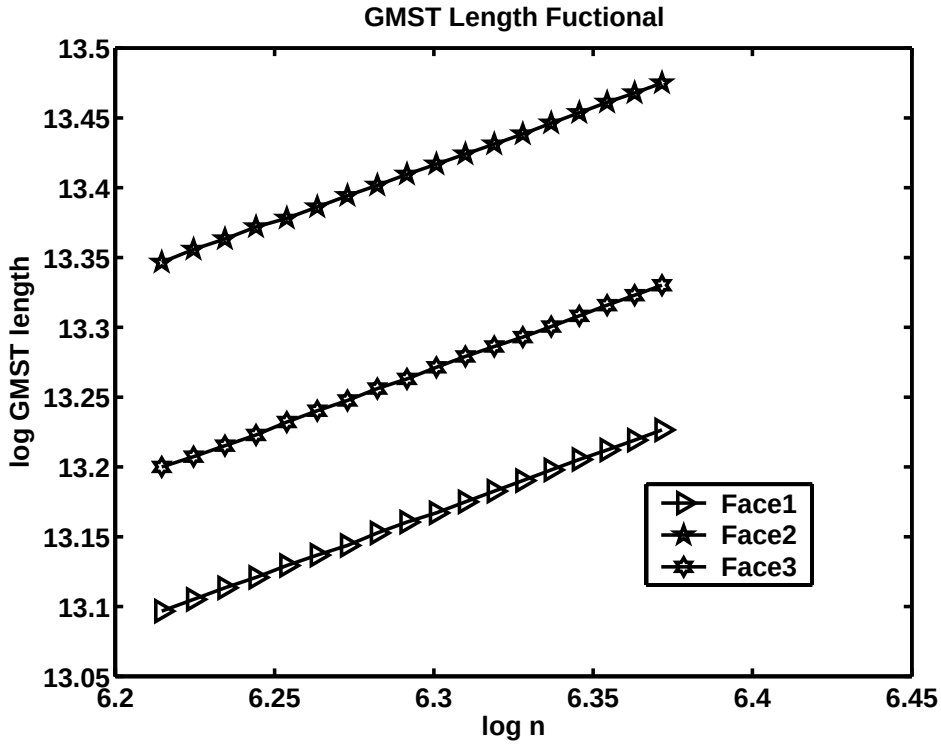
Face folio 3

Face 4

Face folio 4

GMST for 3 Face Folios





	Face 1	Face 2	Face 3	Face 4	Face 1 + 3
$\hat{m}$	5	5	6	6	6
$\hat{H}$ (bits)	20.8	22.9	20.3	24.0	21.8

Conclusions

- High-D features make divergence estimation challenging
- Entropic graph methods can be implemented for such cases, demonstrated for multi-modality image registration
- Can use entropic graph methods to obtain consistent estimators of dimension and entropy of samples on a manifold
- Manifold learning and model reduction
  - LLE, LE, HE estimate d by finding local linear representation of manifold
  - Entropic graph estimates d from global resampling
  - Initialization of ISOMAP… with entropic graph estimator
- Computational considerations
  - GMST : O(d n^2 log n)
  - GMST with greedy neighborhood search O(d n log n)
  - kNN with kdb tree partitioning: O(d n log n)