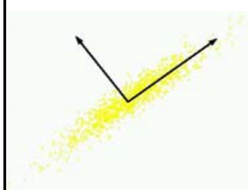
Advanced Machine Learning

Data visualization and dimensionality reduction

Eric Xing

Lecture 3, August 10, 2009



Text document retrieval/labelling

- Represent each document by a high-dimensional vector in the space of words

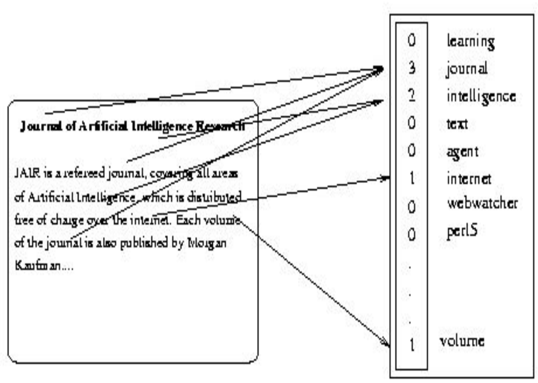


Image retrieval/labelling

img1.jpg

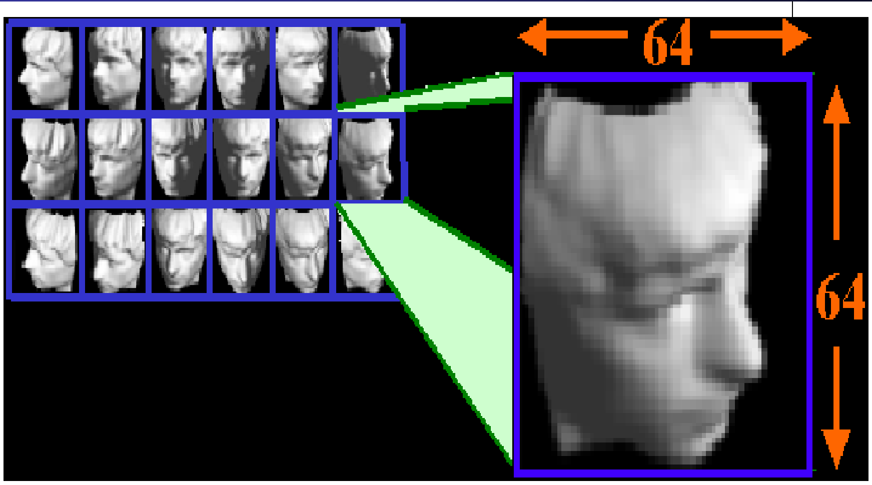
$$x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

Dimensionality Bottlenecks

- Data dimension
  - Sensor response variables Y:
    - 1,000,000 samples of an EM/Acoustic field on each of N sensors
    - 1024^2 pixels of a projected image on a IR camera sensor
    - N^2 expansion factor to account for all pairwise correlations
- Information dimension
  - Number of free parameters describing probability densities f(Y) or f(S|Y)
    - For known statistical model: info dim = model dim
    - For unknown model: info dim = dim of density approximation
- Parametric-model driven dimension reduction
  - DR by sufficiency, DR by maximum likelihood, DR by ancillarity
- Data-driven dimension reduction
  - Manifold learning, structure discovery

Intuition: how does your brain store these pictures?

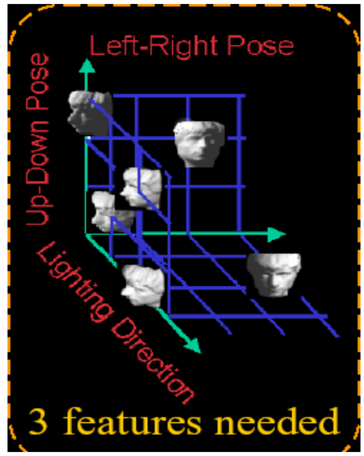
Brain Representation



Brain Representation

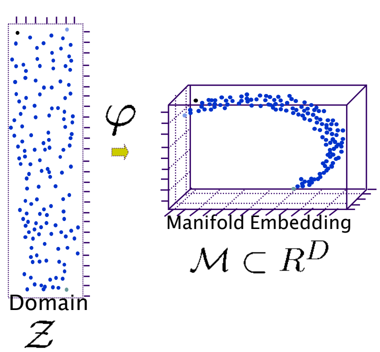
- Every pixel?
- Or perceptually meaningful structure?
  - Up-down pose
  - Left-right pose
  - Lighting direction

  So, your brain successfully reduced the high-dimensional inputs to an intrinsically 3-dimensional manifold!



Two Geometries to Consider

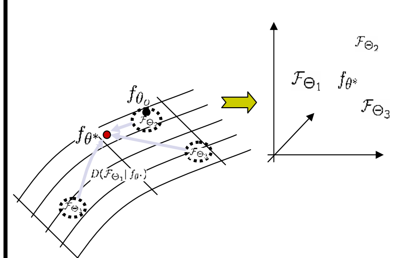
(Metric) data geometry

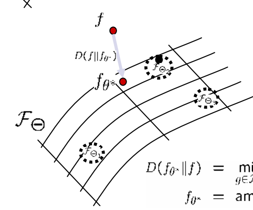


$\{X_i \in \mathcal{M}\}_i$ are i.i.d. samples fr$f_X(x)$

$$P(X \in B) = \int_{B \cap \mathcal{M}} f_X(x)dx = \int_{\varphi^{-1}(B \cap \mathcal{M})} f_Z(z)dz$$

(Non-metric) information geometry

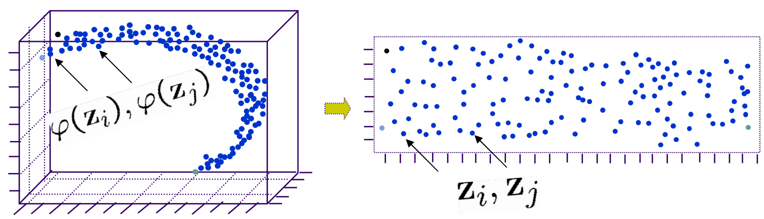




$$D(f_{\theta^*} \| f) = \min_{g \in \mathcal{F}_\Theta} D(g \| f)$$
$$f_{\theta^*} = \text{amin}_{g \in \mathcal{F}_\Theta} D(g \| f)$$

Data-driven DR

- Data-driven projection to lower dimensional subsapce
- Extract low-dim structure from high-dim data
- Data may lie on curved (but locally linear) subspace



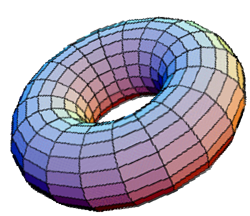
[1] Josh .B. Tenenbaum, Vin de Silva, and John C. Langford "A Global Geometric Framework for Nonlinear Dimensionality Reduction" *Science*, 22 Dec 2000.

[2] Jose Costa, Neal Patwari and Alfred O. Hero, "Distributed Weighted Multidimensional Scaling for Node Localization in Sensor Networks", *IEEE/ACM Trans. Sensor Networks*, to appear 2005.

[3] Misha Belkin and Partha Niyogi, "Laplacian eigenmaps for dimensionality reduction and data representation," *Neural Computation*, 2003.

What is a Manifold?

- A manifold is a topological space which is locally Euclidean.
- Represents a very useful and challenging unsupervised learning problem.
- In general, any object which is nearly "flat" on small scales is a manifold.



Manifold Learning

- Discover low dimensional structures (smooth manifold) for data in high dimension.
- Linear Approaches
  - Principal component analysis.
  - Multi dimensional scaling.
- Non Linear Approaches
  - Local Linear Embedding
  - ISOMAP
  - Laplacian Eigenmap.

Principal component analysis

- Areas of variance in data are where items can be best discriminated and key underlying phenomena observed
- If two items or dimensions are highly correlated or dependent
  - They are likely to represent highly related phenomena
  - We want to combine related variables, and focus on uncorrelated or independent ones, especially those along which the observations have high variance
- We look for the phenomena underlying the observed covariance/co-dependence in a set of variables
- These phenomena are called “factors” or “principal components” or “independent components,” depending on the methods used
  - Factor analysis: based on variance/covariance/correlation
  - Independent Component Analysis: based on independence

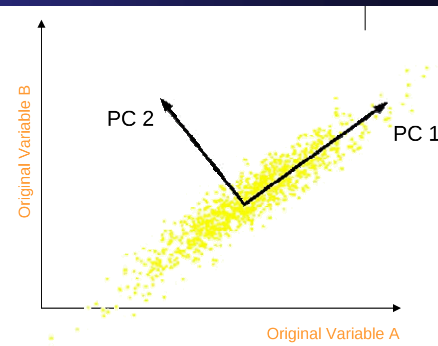
An example:

Principal Component Analysis

- The new variables/dimensions
  - Are linear combinations of the original ones
  - Are uncorrelated with one another
    - Orthogonal in original dimension space
  - Capture as much of the original variance in the data as possible
  - Are called Principal Components

- Orthogonal directions of greatest variance in data

- Projections along PC1 discriminate the data most along any one axis



- First principal component is the direction of greatest variability (covariance) in the data
- Second is the next orthogonal (uncorrelated) direction of greatest variability
  - So first remove all the variability along the first component, and then find the next direction of greatest variability
- And so on …

Computing the Components

- Projection of vector $\mathbf{x}$ onto an axis (dimension) $\mathbf{u}$ is $\mathbf{u}^T\mathbf{x}$
- Direction of greatest variability is that in which the average square of the projection is greatest:

Maximise $\quad \mathbf{u}^T\mathbf{X}\mathbf{X}^T\mathbf{u}$

s.t $\quad \mathbf{u}^T\mathbf{u} = 1$

Construct Langrangian $\mathbf{u}^T\mathbf{X}\mathbf{X}^T\mathbf{u} - \lambda\mathbf{u}^T\mathbf{u}$

Vector of partial derivatives set to zero

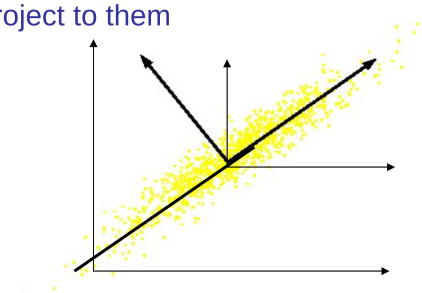
$\mathbf{x}\mathbf{x}^T\mathbf{u} - \lambda\mathbf{u} = (\mathbf{x}\mathbf{x}^T - \lambda\mathbf{I})\ \mathbf{u} = 0$

As $\mathbf{u} \neq \mathbf{0}$ then $\mathbf{u}$ must be an eigenvector of $\mathbf{X}\mathbf{X}^T$ with eigenvalue λ

- λ is the principal eigenvalue of the correlation matrix C= $\mathbf{X}\mathbf{X}^T$
- The eigenvalue denotes the amount of variability captured along that dimension

Computing the Components

- Similarly for the next axis, etc.
- So, the new axes are the eigenvectors of the matrix of correlations of the original variables, which captures the similarities of the original variables based on how data samples project to them



- Geometrically: centering followed by rotation
  - Linear transformation

Eigenvalues & Eigenvectors

- For symmetric matrices, eigenvectors for distinct eigenvalues are **orthogonal**

$$Sv_{\{1,2\}} = \lambda_{\{1,2\}} v_{\{1,2\}}, \text{ and } \lambda_1 \neq \lambda_2 \Rightarrow v_1 \bullet v_2 = 0$$

- All eigenvalues of a real symmetric matrix are **real**.

$$\text{if } |S - \lambda I| = 0 \text{ and } S = S^T \Rightarrow \lambda \in \Re$$

- All eigenvalues of a positive semidefinite matrix are **non-negative**

$$\forall w \in \Re^n, w^T S w \geq 0, \text{ then if } Sv = \lambda v \Rightarrow \lambda \geq 0$$

Eigen/diagonal Decomposition

- Let $\mathbf{S} \in \mathbb{R}^{m \times m}$ be a **square** matrix with ***m*** **linearly independent eigenvectors** (a "non-defective" matrix)

- **Theorem**: Exists an **eigen decomposition**

$$\mathbf{S} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^{-1}$$

diagonal

Unique for distinct eigen-values

(cf. matrix diagonalization theorem)

- Columns of ***U*** are **eigenvectors** of ***S***
- Diagonal elements of $\mathbf{\Lambda}$ are **eigenvalues** of $\mathbf{S}$

$$\mathbf{\Lambda} = \text{diag}(\lambda_1, \ldots, \lambda_m), \quad \lambda_i \geq \lambda_{i+1}$$

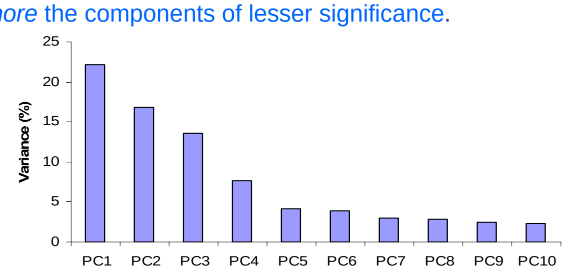
PCs, Variance and Least-Squares

- The first PC retains the greatest amount of variation in the sample
- The k^{th} PC retains the kth greatest fraction of the variation in the sample
- The k^{th} largest eigenvalue of the correlation matrix C is the variance in the sample along the k^{th} PC
- The least-squares view: PCs are a series of linear least squares fits to a sample, each orthogonal to all previous ones

How Many PCs?

- For n original dimensions, sample covariance matrix is nxn, and has up to n eigenvectors. So n PCs.
- Where does dimensionality reduction come from?

  Can *ignore* the components of lesser significance.



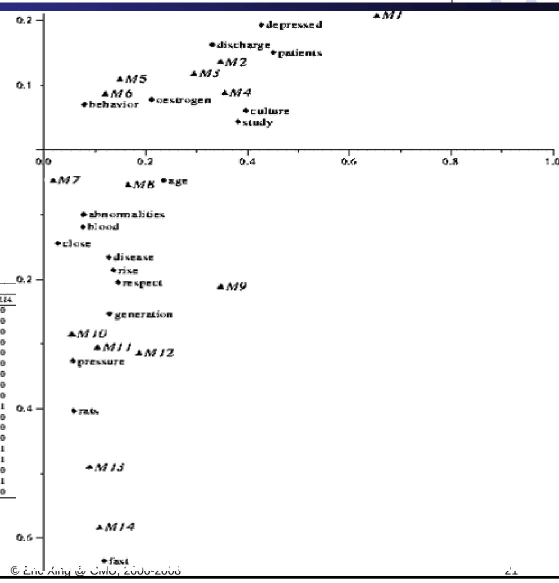
You do lose some information, but if the eigenvalues are small, you don't lose much

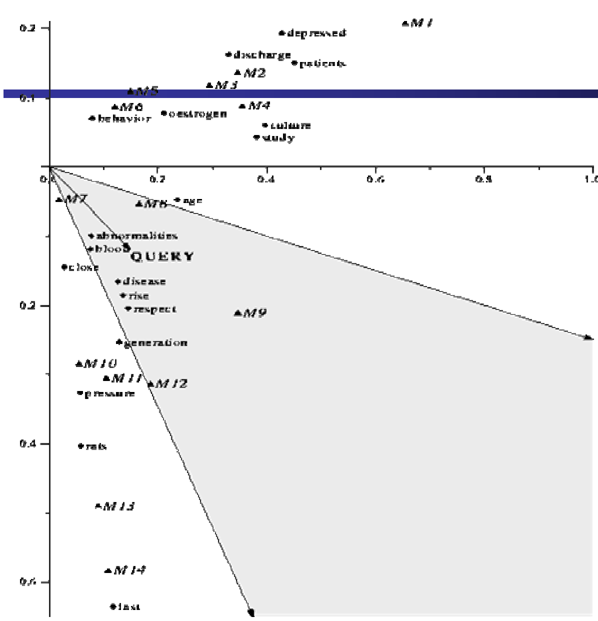
- n dimensions in original data
- calculate n eigenvectors and eigenvalues
- choose only the first p eigenvectors, based on their eigenvalues
- final data set has only p dimensions

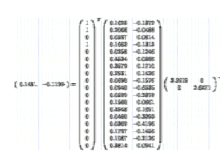
Application: querying text doc.

Label	Medical Topic
M1	study of depressed patients after discharge with regard to age of onset and culture
M2	culture of pleuropneumonia like organisms found in vaginal discharge of patients
M3	study showed oestrogen production is depressed by ovarian irradiation
M4	cortisone rapidly depressed the secondary rise in oestrogen output of patients
M5	boys tend to react to death anxiety by acting out behavior while girls tended to become depressed
M6	changes in children's behavior following hospitalization studied a week after discharge
M7	surgical technique to close ventricular septal defects
M8	chromosomal abnormalities in blood cultures and bone marrow from leukemic patients
M9	study of christmas disease with respect to generation and culture
M10	insulin not responsible for metabolic abnormalities accompanying a prolonged fast
M11	close relationship between high blood pressure and vascular disease
M12	mouse kidneys show a decline with respect to age in the ability to concentrate the urine during a water fast
M13	fast cell generation in the eye lens epithelium of rats
M14	fast rise of cerebral oxygen pressure in rats

Terms	Documents													
	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11	M12	M13	M14
abnormalities	0	0	0	0	0	0	0	1	0	1	0	0	0	0
age	1	0	0	0	0	0	0	0	0	0	0	1	0	0
behavior	0	0	0	0	1	1	0	0	0	0	0	0	0	0
blood	0	0	0	0	0	0	0	1	0	0	1	0	0	0
close	0	0	0	0	0	0	1	0	0	0	1	0	0	0
culture	1	1	0	0	0	0	0	1	1	0	0	0	0	0
depressed	1	0	1	1	1	0	0	0	0	0	0	0	0	0
discharge	1	1	0	0	0	1	0	0	0	0	0	0	0	0
disease	0	0	0	0	0	0	0	0	1	0	1	0	0	0
fast	0	0	0	0	0	0	0	0	0	1	0	1	1	1
generation	0	0	0	0	0	0	0	0	1	0	0	0	1	0
oestrogen	0	0	1	1	0	0	0	0	0	0	0	0	0	0
patients	1	1	0	1	0	0	0	1	0	0	0	0	0	0
pressure	0	0	0	0	0	0	0	0	0	0	1	0	0	1
rats	0	0	0	0	0	0	0	0	0	0	0	0	1	1
respect	0	0	0	0	0	0	0	1	0	0	0	1	0	0
rise	0	0	0	1	0	0	0	0	0	0	0	0	0	1
study	1	0	1	0	0	0	0	0	1	0	0	0	0	0







Number of Factors					
$k=2$		$k=4$		$k=8$	
M 9	1.00	M 8	0.92	M 8	0.67
M12	0.88	M 9	0.89	M12	0.55
M 8	0.85	M 2	0.64	M10	0.54
M11	0.82	M10	0.48		
M10	0.79	M12	0.46		
M 7	0.74	M11	0.40		
M14	0.72				
M13	0.71				
M 4	0.67				
M 1	0.56				
M 2	0.42				

Within .40 threshold

K is the number of singular values used

Summary:

- Principle
  - Linear projection method to reduce the number of parameters
  - Transfer a set of correlated variables into a new set of uncorrelated variables
  - Map the data into a space of lower dimensionality
  - Form of unsupervised learning
- Properties
  - It can be viewed as a rotation of the existing axes to new positions in the space defined by original variables
  - New axes are orthogonal and represent the directions with maximum variability
- Application: In many settings in pattern recognition and retrieval, we have a feature-object matrix.
  - For text, the terms are features and the docs are objects.
  - Could be opinions and users …
  - This matrix may be redundant in dimensionality.
  - Can work with low-rank approximation.
  - If entries are missing (e.g., users' opinions), can recover if dimensionality is low.

Going beyond

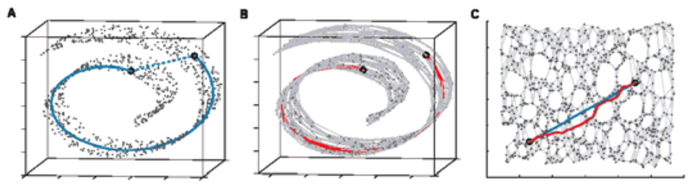
- The is the essence of the C matrix?

$$C = E[XX^T] = \frac{1}{n}\mathbf{X}\mathbf{X}^T$$

- The elements in C captures some kind of affinity between a pair of data points in the semantic space
- We can replace it with any reasonable affinity measure
  - E.g., $D = \left(\|x_i - x_j\|^2\right)_{ij}$: distance matrix MDS
  - E.g., the geodistance ISOMAP

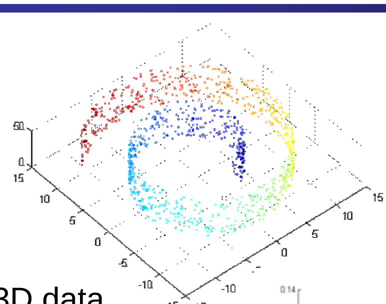
Nonlinear DR – Isomap

[Josh. Tenenbaum, Vin de Silva, John langford 2000]

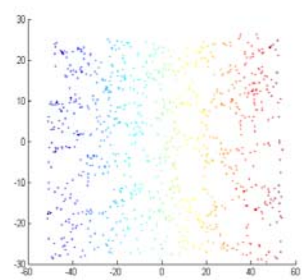


- Constructing neighbourhood graph G
- For each pair of points in G, Computing shortest path distances ---- geodesic distances.
  - Use Dijkstra's or Floyd's algorithm
- Apply kernel PCA for C given by the centred matrix of squared geodesic distances.
- Project test points onto principal components as in kernel PCA.

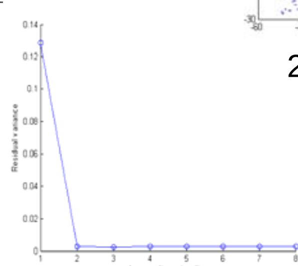
"Swiss Roll" dataset



3D data



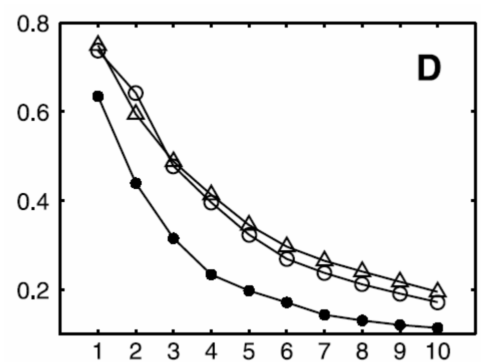
2D coord chart

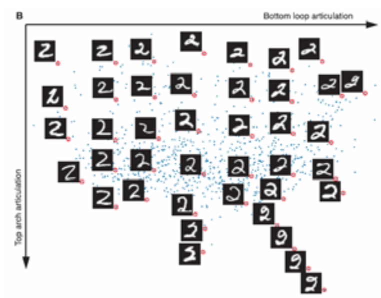


Error vs. dimensionality of coordinate chart

PCA, MD vs ISOMAP

- The residual variance of PCA (open triangles), MDS (open circles), and Isomap





ISOMAP algorithm Pros/Cons

Advantages:

- Nonlinear
- Globally optimal
- Guarantee asymptotically to recover the true dimensionality

Drawback:

- May not be stable, dependent on topology of data
- As N increases, pair wise distances provide better approximations to geodesics, but cost more computation

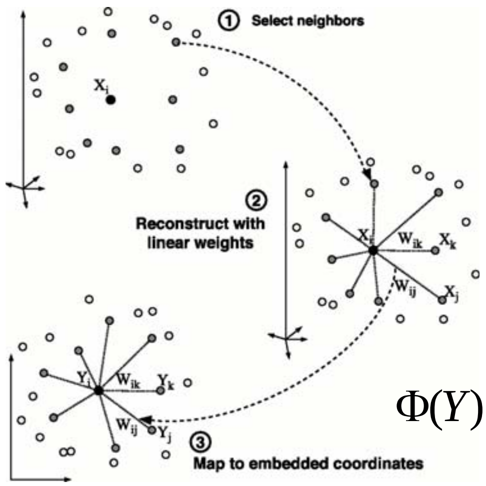
Local Linear Embedding (a.k.a LLE)

- LLE is based on simple geometric intuitions.
- Suppose the data consist of N real-valued vectors X_i, each of dimensionality D.
- Each data point and its neighbors expected to lie on or close to a locally linear patch of the manifold.

Steps in LLE algorithm

- Assign neighbors to each data point $\vec{X}_i$
- Compute the weights W_{ij} that best linearly reconstruct the data point from its neighbors, solving the constrained least-squares problem.
- Compute the low-dimensional embedding vectors $\vec{Y}_i$ best reconstructed by W_{ij}.

Fit locally, Think Globally



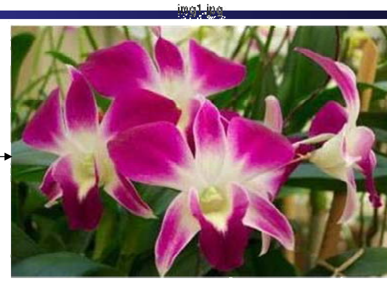
From Nonlinear Dimensionality Reduction by Locally Linear Embedding

Sam T. Roweis and Lawrence K. Saul

$$\Phi(Y) = \sum_i |\vec{Y} - \sum_j W_{ij}\vec{Y}_j|^2$$

Super-Resolution Through Neighbor Embedding [Yeung et al CVPR 2004]

Training Xs^i



Training Ys^i

Testing Xt

?

Testing Yt

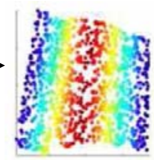
Intuition

- Patches of the image lie on a manifold

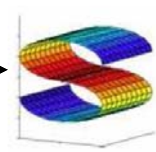
Training Xs^i

img1.jpg

Training Ys^i



Low dimensional Manifold



High dimensional Manifold

Algorithm

1. Get feature vectors for each low resolution training patch.
2. For each test patch feature vector find K nearest neighboring feature vectors of training patches.
3. Find optimum weights to express each test patch vector as a weighted sum of its K nearest neighbor vectors.
4. Use these weights for reconstruction of that test patch in high resolution.

Results

Training Xs^i

Training Ys^i

Testing Xt

Testing Yt

Summary:

- Principle
  - Linear and nonlinear projection method to reduce the number of parameters
  - Transfer a set of correlated variables into a new set of uncorrelated variables
  - Map the data into a space of lower dimensionality
  - Form of unsupervised learning
- Applications
  - PCA and Latent semantic indexing for text mining
  - Isomap and Nonparametric Models of Image Deformation
  - LLE and Isomap Analysis of Spectra and Colour Images
  - Image Spaces and Video Trajectories: Using Isomap to Explore Video Sequences
  - Mining the structural knowledge of high-dimensional medical data using isomap

Isomap Webpage: http://isomap.stanford.edu/