Metric Learning

Technical University of Munich
Department of Informatics
Computer Vision Group

November 22, 2018

Outline

Outline

Motivation

- How do we measure similarity?
- What is a metric?
- Why to learn a metric?
- How to learn a metric?

How do we measure similarity?

- Most algorithms that intend to extract knowledge from data, have to investigate relationships between objects.
- This often reduces to computing *distances* between data points.
- Thus, such an algorithm's performance critically depends on its *notion of similarity* between objects.



What is a metric?

A **metric** or **distance function** is a function that defines a distance between each pair of elements of a set.

Formally, it is a mapping $\mathcal{D} : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}^+$ over a vector space $\mathcal{X}$, where the following conditions are satisfied $\forall x_i, x_j, x_k \in \mathcal{X}$:

1.	$\mathcal{D}(x_i, x_j) \geq 0$	**Non-negativity**
2.	$\mathcal{D}(x_i, x_j) = \mathcal{D}(x_j, x_i)$	**Symmetry**
3.	$\mathcal{D}(x_i, x_j) \leq \mathcal{D}(x_i, x_k) + \mathcal{D}(x_k, x_j)$	**Triangle inequality**
4.	$\mathcal{D}(x_i, x_j) = 0 \Leftrightarrow x_i = x_j$	**Identity of indiscernibles**

If condition 4 is not met, we are referring to a **pseudo-metric**. Usually we do not distinguish between metrics and pseudo-metrics.

Why learn a metric?

"The greatest thing by far is to be a master of metaphor; it is the one thing that cannot be learned from others; and it is also a sign of genius, since a good metaphor implies an intuitive perception of the similarity of the dissimilar."

Aristotle

Why learn a metric?

- Sometimes, the problem implicitly defines a suitable similarity measure, e.g. Hamming distance to compare binary vectors: $d_H(\text{“1001”}, \text{“1010”}) = 2$.
- In many interesting problems however, a proper similarity measure is not easy to define. It is preferable then to learn a metric directly from data.

A family of metrics

A family of metrics over $\mathcal{X}$ is defined by computing Euclidean distances after applying a linear transformation $\mathbf{L}$ such that $x \rightarrow \mathbf{L}x$. These metrics compute **squared** distances as

$$\mathcal{D}_{\mathbf{L}}(x_i, x_j) = ||\mathbf{L}x_i - \mathbf{L}x_j||_2^2 \tag{1}$$

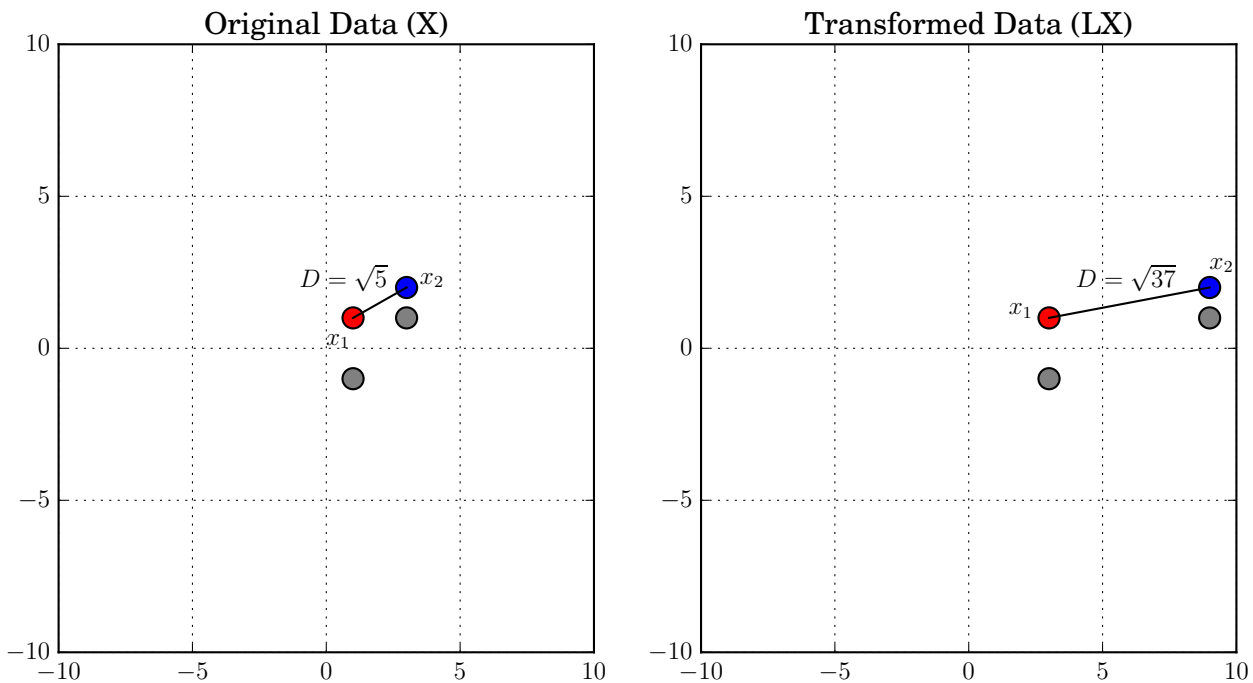
Equation (1) defines a valid metric if $\mathbf{L}$ is full rank and a valid pseudo-metric otherwise.

Intuitively, we want to

- *stretch* the dimensions that contain more information and
- *contract* the dimensions that explain less of the data.

A family of metrics - An example

Consider two data points $x_1 = (1, 1)$ and $x_2 = (3, 2)$ that are known to be dissimilar. The transformation $\mathbf{L} = \begin{pmatrix} 3 & 0 \\ 0 & 1 \end{pmatrix}$ maps the points to $x_1' = (3, 1)$ and $x_2' = (9, 2)$ as it *weights* distances along the first axis 3 times more than the second. The squared distance of the points changed from $(3-1)^2 + (2-1)^2 = 5$ to $(9-3)^2 + (2-1)^2 = 37$.



Another view: Mahalanobis metrics

Expanding the squared distances equation:

$$\mathcal{D}_{\mathbf{L}}(x_i, x_j) = ||\mathbf{L}x_i - \mathbf{L}x_j||_2^2 = (x_i - x_j)^T \mathbf{L}^T \mathbf{L}(x_i - x_j)$$

This allows us to express squared distances in terms of the square matrix $\mathbf{M} = \mathbf{L}^T\mathbf{L}$ which is guaranteed to be *positive semidefinite*. In terms of $\mathbf{M}$ we denote squared distances as

$$\mathcal{D}_{\mathbf{M}}(x_i, x_j) = (x_i - x_j)^T \mathbf{M}(x_i - x_j)$$

We refer to pseudo-metrics of this form as **Mahalanobis** metrics.

It is easy to see that by setting $\mathbf{M}$ equal to the identity matrix, we fall back to common *squared* Euclidean distances.

To learn **L** or **M**

Thus, we have two options on what to learn, which gives rise to two approaches in Metric Learning:

- Learn a linear transformation **L** of the data
  - $\mathbf{M} = \mathbf{L}^T\mathbf{L}$ is then uniquely defined
  - The optimization of **L** is unconstrained
- Learn a Mahalanobis metric **M**
  - **M** defines **L** up to rotation (does not influence distances)
  - Constraint: **M** must be positive semidefinite
  - But has certain advantages:
    - The dimensions of **M** are fixed a-priori ($d \times d$)
    - Interpretation: Similar to inverse covariance matrix
    - Objectives are linear in **M** (gradient is independent of **M**)

Outline

Principal Component Analysis [Pearson, 1901]

The main goal of PCA is to find the linear transformation **L** that projects the data to a subspace that **maximizes the variance**.

The variance is expressed with the covariance matrix

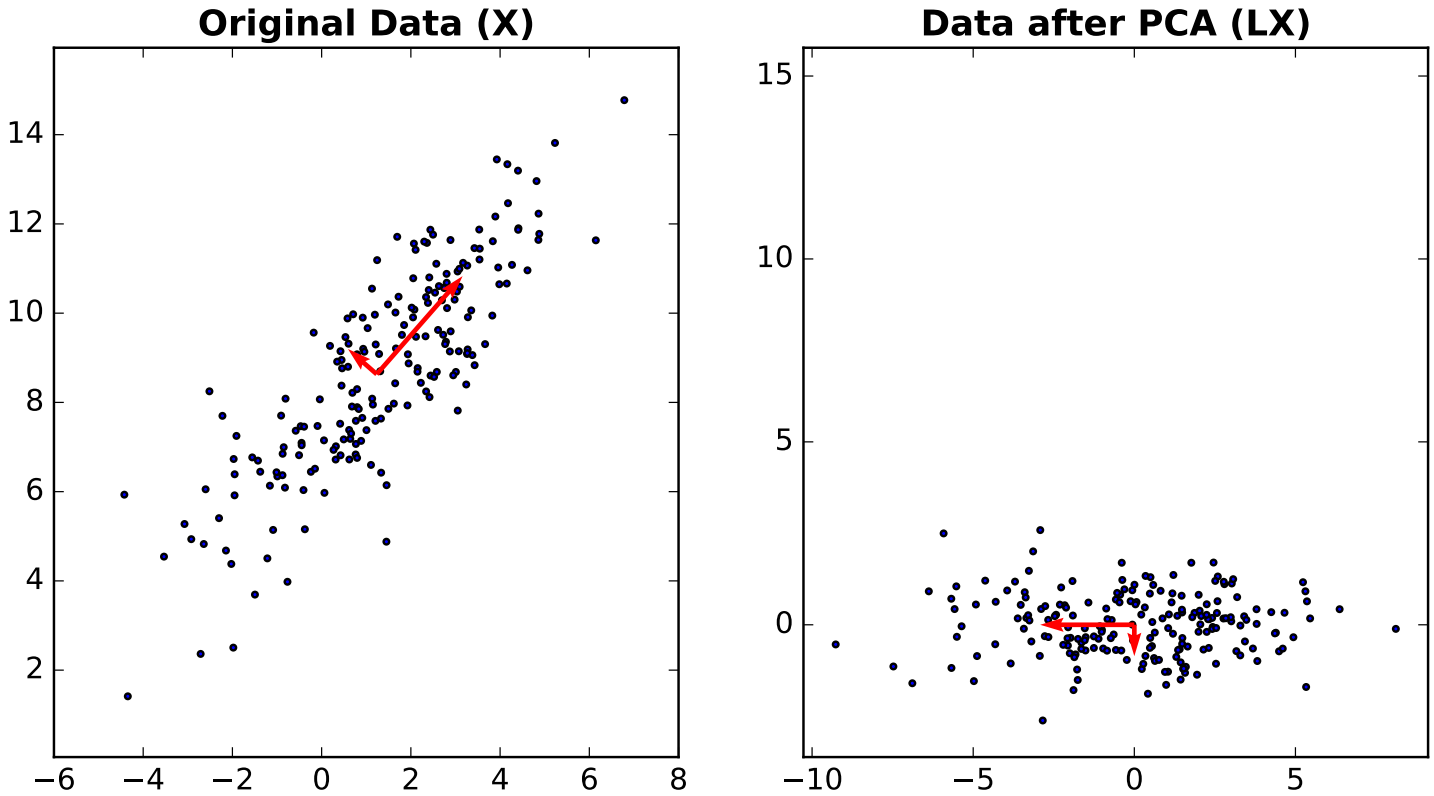
$$\mathbf{C} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)(x_i - \mu)^T$$

where $\mu = \frac{1}{n}\sum_{i=1}^{n} x_i$ is the sample mean.

It turns out that $\mathbf{C} = \frac{1}{n}XX^T$ (assuming zero-mean $X \in \mathbb{R}^{d \times n}$).
The covariance of the projected inputs is then

$$\mathbf{C}' = \frac{1}{n}(\mathbf{L}X)(\mathbf{L}X)^T = \frac{1}{n}\mathbf{L}XX^T\mathbf{L}^T = \mathbf{L}\mathbf{C}\mathbf{L}^T$$

Principal Component Analysis - Illustration



In red: The first two eigenvectors of the covariance matrix, scaled by the square roots of the two largest eigenvalues respectively.

Principal Component Analysis (cont'd)

We can formulate PCA as an optimization problem:

$$\max_{\mathbf{L}} \operatorname{Tr}(\mathbf{L}\mathbf{C}\mathbf{L}^T) \quad \text{subject to} \quad \mathbf{L}\mathbf{L}^T = \mathbf{I}$$

Closed-form solution: Rows of $\mathbf{L}$ are the eigenvectors of $\mathbf{C}$.
Eigen-decomposing $\mathbf{C}$ is equivalent to computing the SVD of $\mathbf{X}$.

Remarks around PCA

- Is an unsupervised method (does **not** use data labels)
- Is widely used for dimensionality reduction: $\mathbf{L} \in \mathbb{R}^{p \times d}, p < d$
- Can be used for:
  - *De-noising*: By removing the bottom eigenvectors
  - Speeding up search of nearest neighbors.

Outline

Linear Discriminant Analysis [Fisher, 1936]

Unlike PCA, LDA is supervised: it uses labels of the inputs.
Goal: Find the **L** that **maximizes the between-class variance w.r.t. the within-class variance**.

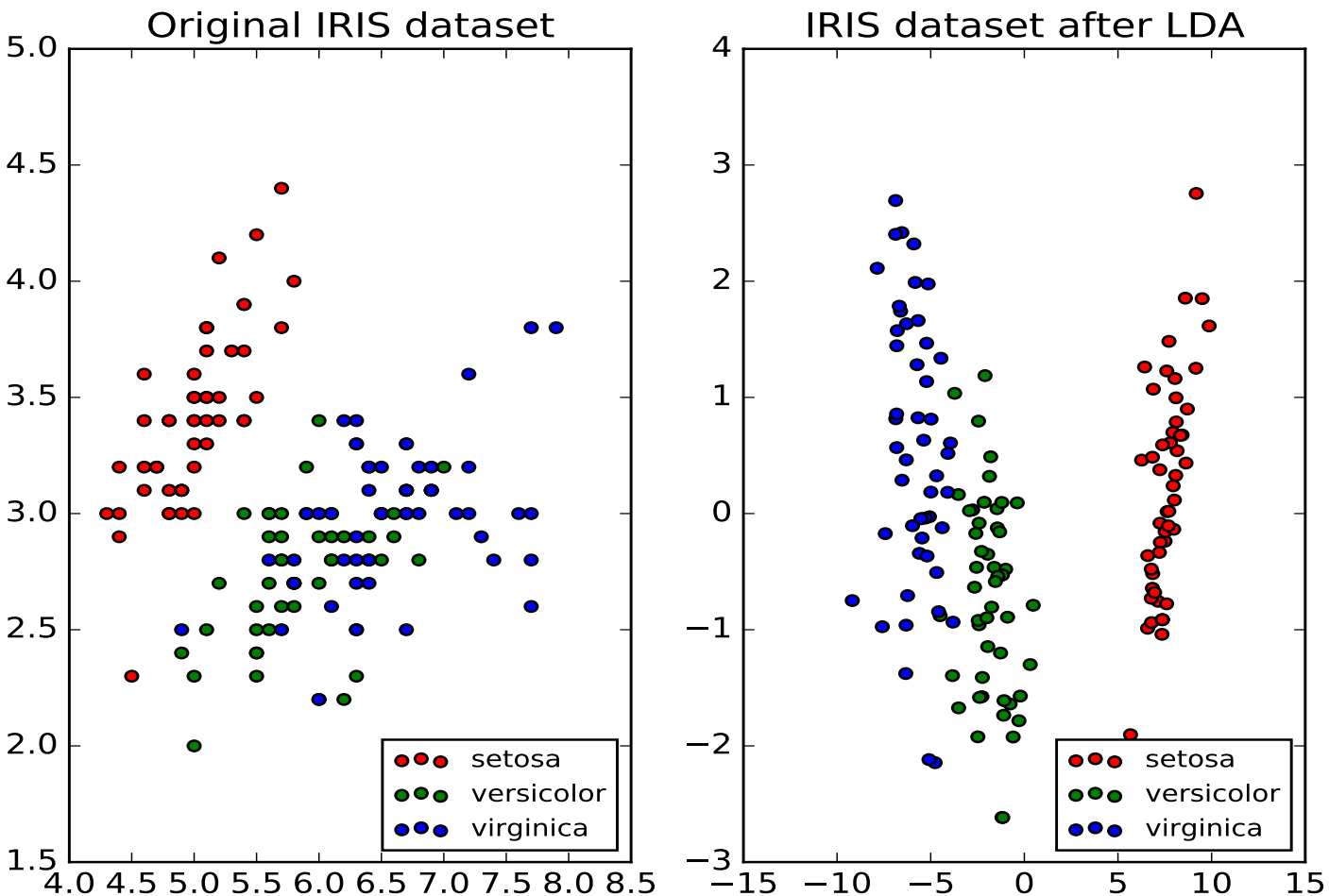
Assuming we have m classes, the covariance matrices are

$$\mathbf{C}_b = \frac{1}{m} \sum_{c=1}^{m} \mu_c \mu_c^T$$

$$\mathbf{C}_w = \frac{1}{n} \sum_{c=1}^{m} \sum_{i \in \Omega_c} (x_i - \mu_c)(x_i - \mu_c)^T,$$

where Ω_c is the set of indices of inputs that belong to class c, μ_c is the sample mean of class c.
We assume that the data are globally centered.

Linear Discriminant Analysis - Illustration



Linear Discriminant Analysis (cont'd)

Corresponding optimization problem:

$$\max_{\mathbf{L}} \operatorname{Tr}\left(\frac{\mathbf{L}\mathbf{C}_b\mathbf{L}^T}{\mathbf{L}\mathbf{C}_w\mathbf{L}^T}\right) \quad \text{subject to} \quad \mathbf{L}\mathbf{L}^T = \mathbf{I}$$

Closed form solution: Rows of $\mathbf{L}$ are the eigenvectors of $\mathbf{C}_w^{-1}\,\mathbf{C}_b$.

Remarks around LDA

- Is a supervised method (makes use of label information)
- Is widely used as a preprocessing step for pattern classification
- Works well when class distributions are Gaussians

Neighborhood Component Analysis [Goldberger et al., 2004]

Idea: Learn a Mahalanobis metric explicitly to improve ***k-nn*** classification.
Goal: Estimate the **L** that minimizes the expected LOO error.

Observations

- LOO error is highly discontinuous w.r.t. the distance metric. ☹
- In particular, an infinitesimal change in the metric can alter the neighbour graph and thus change the validation performance.
- We need a smoother (or at least continuous) function

Idea 2: Instead of picking a fixed number of k nearest neighbors, select a single neighbor **stochastically** and count the expected votes.

Neighborhood Component Analysis (cont'd)

The neighbors x_j for each point x_i are drawn from a softmax pdf:

$$p_{ij} = \begin{cases} \frac{\exp(-||(\mathbf{L}x_i - \mathbf{L}x_j||^2))}{\sum_{k \neq i} \exp(-||(\mathbf{L}x_i - \mathbf{L}x_k||^2))} & \text{if } i \neq j \\ 0 & \text{if } i = j \end{cases}$$

The fraction of the time that x_i will be correctly labeled is:

$$p_i = \sum_{j \in C_i} p_{ij} \qquad \textit{where } C_i = \{j : y_i = y_j\}$$

The expected error then is

$$\varepsilon_{NCA} = 1 - \frac{1}{n} \sum_{ij} p_{ij} y_{ij} \quad \text{where} \quad y_{ij} = \begin{cases} 1 & \text{if } y_i = y_j \\ 0 & \text{otherwise} \end{cases}$$

Neighborhood Component Analysis (cont'd)

Optimization

- $\max_{L} f(\mathbf{L}) = \max_{L} \sum_i p_i$ (equivalent to ℓ_1 error minimization)

$$\frac{\vartheta \varepsilon}{\vartheta \mathbf{L}} = 2\mathbf{L} \sum_i \left(p_i \sum_k p_{ik} x_{ik} x_{ik}^T - \sum_{j \in C_i} p_{ij} x_{ij} x_{ij}^T \right)$$

- $\max_{L} g(\mathbf{L}) = \max_{L} \sum_i \log p_i$ (KL divergence minimization)

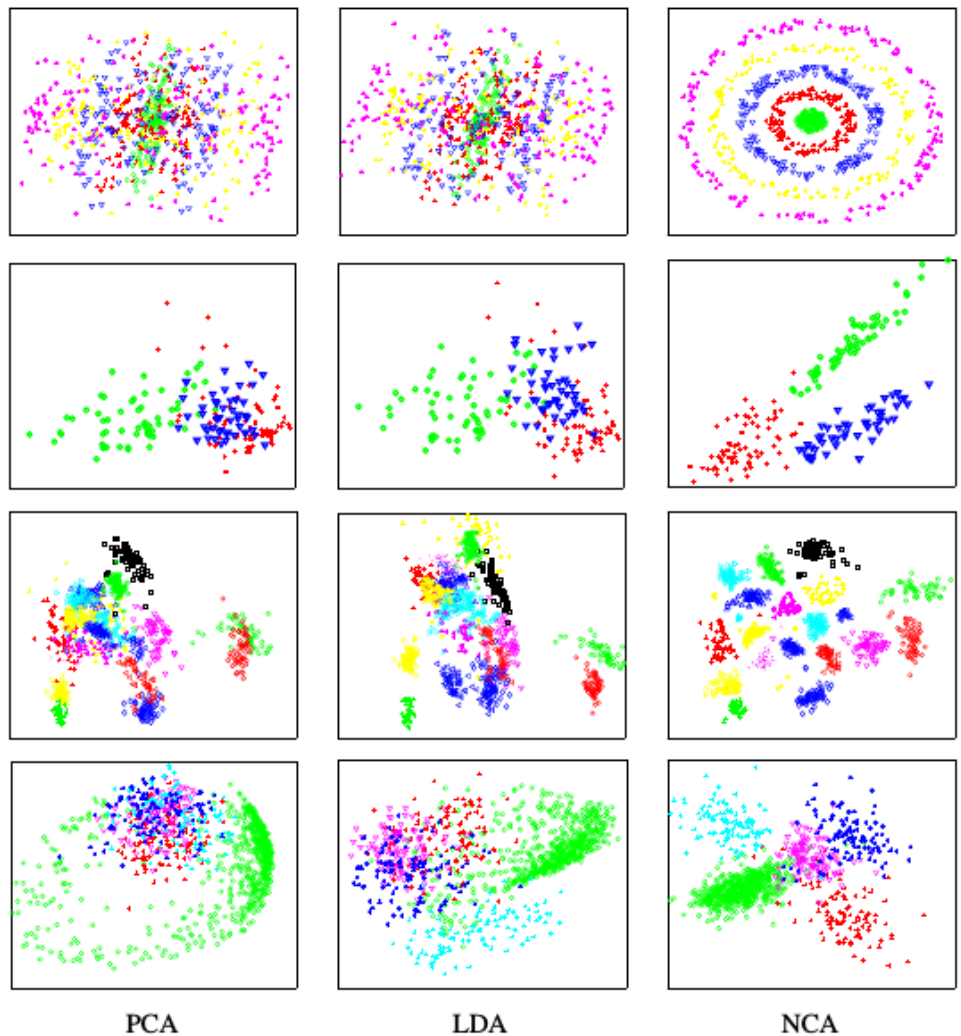
$$\frac{\vartheta \varepsilon}{\vartheta \mathbf{L}} = 2\mathbf{L} \sum_i \left(\sum_k p_{ik} x_{ik} x_{ik}^T - \frac{\sum_{j \in C_i} p_{ij} x_{ij} x_{ij}^T}{\sum_{j \in C_i} p_{ij}} \right)$$

Remarks around NCA

- We don't have to choose a parameter k ☺
- The stochastic nature makes ε_{NCA} differentiable w.r.t. $\mathbf{L}$ ☺
- But ε_{NCA} is not convex $\rightarrow$ no globally optimal $\mathbf{L}$ ☹

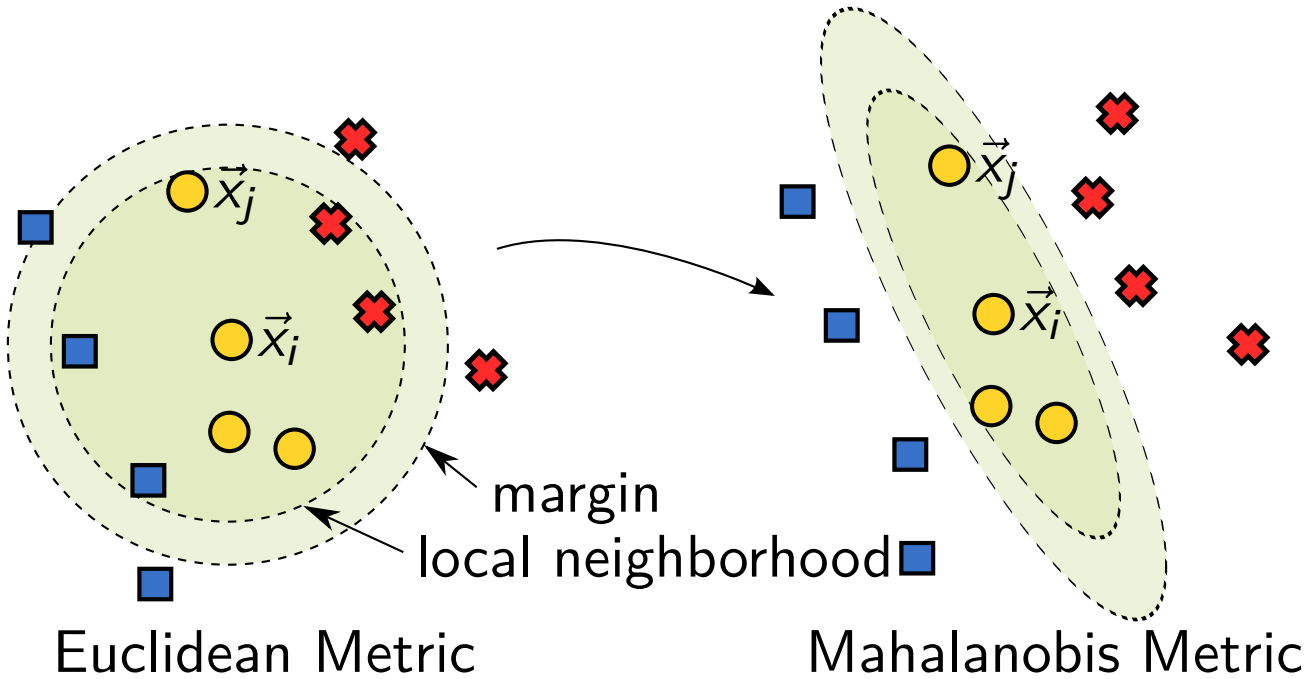
Data Visualization - PCA vs LDA vs NCA

Dataset	Dimensions
concentric rings	3
wine	13
faces	560
digits	256



Large Margin Nearest Neighbor [Weinberger et al., 2005]

- Idea: Enforce the **maximum margin** possible between intra-class and inter-class samples (as in SVMs)
- Considers triplets of points (x, x_+, x_-).



- Goal: Find a metric to maximize k-nn accuracy
- Advantage: Convex formulation w.r.t. $\mathbf{M}$ ☺

Large Margin Nearest Neighbor (cont'd)

- *Target neighbors* of $\vec{x}_i$: samples desired to be closest to $\vec{x}_i$
- *Impostors*: samples that violate the margin
- Triplet Loss function
  - Pulling target neighbors together

$$\varepsilon_{\text{pull}}(\mathbf{L}) = \sum_{i,j \rightsquigarrow i} ||\mathbf{L}(\vec{x}_i - \vec{x}_j)||^2$$

  - Pushing impostors away

$$\varepsilon_{\text{push}}(\mathbf{L}) = \sum_{i,j \rightsquigarrow i} \sum_{l} (1 - y_{il})[1 + ||\mathbf{L}(\vec{x}_i - \vec{x}_j)||^2 - ||\mathbf{L}(\vec{x}_i - \vec{x}_l)||^2]_+$$

  - Convex combination

$$\varepsilon(\mathbf{L}) = (1 - \mu)\varepsilon_{\text{pull}}(\mathbf{L}) + \mu\varepsilon_{\text{push}}(\mathbf{L}), \quad \mu \in [0, 1]$$

Large Margin Nearest Neighbor - Optimization

Distances as traces

$$d_{\mathbf{M}}(x_i, x_j) = (x_i - x_j)^T \mathbf{M}(x_i - x_j) = Tr(\mathbf{M}(x_i - x_j)(x_i - x_j)^T) = Tr(\mathbf{M}C_{ij})$$

Objective w.r.t. **M**

$$\varepsilon(\mathbf{M}) = (1-\mu)\sum_{i,j\rightsquigarrow i} Tr(\mathbf{M}C_{ij}) + \mu \sum_{i,j\rightsquigarrow i,l} (1-y_{il})[1+Tr(\mathbf{M}C_{ij}) - Tr(\mathbf{M}C_{il})]_+$$

Active triplets at iteration t

$$\mathcal{N}_t = \{(i,j,l) : 1 + Tr(\mathbf{M}_t C_{ij}) - Tr(\mathbf{M}_t C_{il}) > 0\}$$

Gradient at iteration t and subsequent gradients

$$\mathbf{G}_t = \nabla_{\mathbf{M}}\varepsilon(\mathbf{M}_t) = (1-\mu)\sum_{i,j\rightsquigarrow i} C_{ij} + \mu \sum_{(i,j,l)\in\mathcal{N}_t} (C_{ij} - C_{il})$$

$$\mathbf{G}_{t+1} = \mathbf{G}_t - \mu \sum_{(i,j,l)\in\mathcal{N}_t - \mathcal{N}_{t+1}} (C_{ij} - C_{il}) + \mu \sum_{(i,j,l)\in\mathcal{N}_{t+1} - \mathcal{N}_t} (C_{ij} - C_{il})$$

Information Theoretic Metric Learning [Davis et al., 2007]

Problem Formulation

$$\min_{\mathbf{M}} D_{ld}(\mathbf{M}, \mathbf{M}_0) \qquad s.t.$$

$$d_{\mathbf{M}}(x_i, x_j) \leq u \quad \text{if} \quad (i,j) \in \mathcal{S} \qquad \text{(similarity constraints)}$$

$$d_{\mathbf{M}}(x_i, x_j) \geq l \quad \text{if} \quad (i,j) \in \mathcal{D} \qquad \text{(dissimilarity constraints)}$$

$$\mathbf{M} \succeq 0$$

- Stein's loss / log det divergence (convex in $\mathbf{M}$ ☺)

$$D_{ld}(\mathbf{M}, \mathbf{M}_0) = Tr(\mathbf{M}\mathbf{M}_0^{-1}) - \log\det(\mathbf{M}\mathbf{M}_0^{-1}) - d$$

- Equivalent to KL divergence between two multivariate Gaussians with equal means and covariances $\mathbf{M}_0, \mathbf{M}$.
$KL(p(x|\mu, \mathbf{M}_0)||p(x|\mu, \mathbf{M})) = \frac{1}{2} D_{ld}(\mathbf{M}, \mathbf{M}_0)$

Information Theoretic Metric Learning - Optimization

Bregman projections

- Cycle through all constraints once.
- At each iteration project the solution onto the current constraint involving (x_i, x_j):
- $\mathbf{M}_{t+1} = \mathbf{M}_t + \beta \mathbf{M}_t (x_i - x_j)(x_i - x_j)^T \mathbf{M}_t$

Remarks

- Each constraint projection costs $O(d^2)$, therefore one cycle through all constraints costs $O(cd^2)$.
- No eigen-decomposition required ☺
- Automatic enforcement of positive semi-definiteness through rank-one updates ☺
- Easy to incorporate a slack variable for each constraint ☺

Outline

Metric Learning and Kernel Methods

Kernel methods

- Express similarity with the Gram matrix $\mathbf{K}$ which is $n \times n$.
- The feature space Φ is usually high-dimensional (theoretically can be infinite-dimensional).
- The training takes place in the kernel space. The algorithm no longer sees the raw inputs $\mathcal{X}$.
- Algorithms scale as $O(n^2)$ or $O(n^3)$

Metric Learning

- Learns a transformation $\mathbf{L}$, which is $p \times d$ or a Mahalanobis matrix $\mathbf{M}$ which is $d \times d$, like the covariance matrix C .
- Usually $p < d \rightarrow$ learning also results in dimensionality reduction.
- Algorithms scale as $O(d^2)$ or $O(d^3)$

Metric learning can be combined with kernel methods for better results.

Low-Rank Kernel learning [Kulis et al., 2006]

Problem Formulation

$$\min_{\mathbf{K}} D_{ld}(\mathbf{K}, \mathbf{K}_0) \qquad s.t.$$

$$\mathbf{K}(x_i, x_j) \leq u \quad \text{if} \quad (i,j) \in \mathcal{S} \qquad \text{(similarity constraints)}$$

$$\mathbf{K}(x_i, x_j) \geq l \quad \text{if} \quad (i,j) \in \mathcal{D} \qquad \text{(dissimilarity constraints)}$$

$$\mathbf{K} \succeq 0$$

Bregman update

$$\mathbf{K}_{t+1} = \mathbf{K}_t + \beta \mathbf{K}_t (e_i - e_j)(e_i - e_j)^T \mathbf{K}_t$$

Theorem

Let $\mathbf{K}_0 = X^T \mathbf{M}_0 X$. If $\mathbf{M}^*$ is the optimal Mahalanobis metric learned by ITML and $\mathbf{K}^*$ the optimal kernel matrix, then

$$\mathbf{K}^* = X^T \mathbf{M}^* X$$

Metric Learning in kernel space

Problem Formulation

- Assume input kernel function $k(x, y) = \phi(x)^T \phi(y)$.
- Want to learn metric
  $d_{\mathbf{M}}(x, y) = (\phi(x) - \phi(y))^T \mathbf{M} (\phi(x) - \phi(y))$.
- Equivalently: learn a new kernel function of the form
  $\tilde{k}(x, y) = \phi(x)^T \mathbf{M} \phi(y)$.
- Learned kernel $\tilde{k}$ can be shown to be of the form

$$\tilde{k}(x, y) = k(x, y) + \sum_i \sum_j \mathbf{M}_{ij} k(x, x_i) k(x_j, y)$$

- Can update parameters $\mathbf{M}_{ij}$ while optimizing the kernel formulation

Metric Learning Variants

Most metric learning algorithms improve by looking at pairs, triplets or even quadruplets of points.
Many noteworthy algorithms exist:

- Relevant Component Analysis (RCA) [Shental et al., 2002]
- Maximally Collapsing Metric Learning (MCML) [Globerson and Roweis, 2005]
- Information Theoretic Metric Learning (ITML) [Davis et al., 2007]
- LogDet Exact Gradient Online (LEGO) [Jain et al., 2009]
- Siamese Network [Chopra et al., 2005]
- Triplet Network [Hoffer and Ailon, 2015]
- ...

This is definitely not an exhaustive list.

Outline

Dimensionality Reduction

Some fundamental DR methods are learning a metric that optimizes an objective:

- PCA (maximizes variance)
- LDA (maximizes between / within scatter ratio)
- NCA (maximizes knn accuracy)

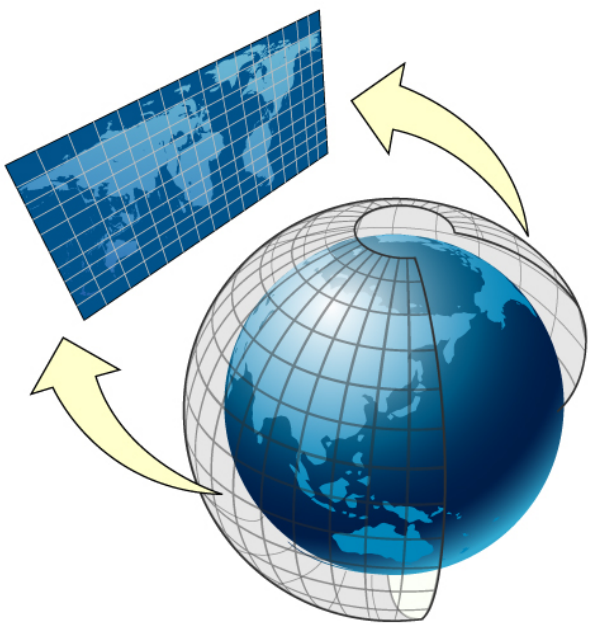
But most DR methods do not *explicitly* learn a metric:

- Multi-dimensional scaling (MDS) [Torgerson, 1952]
- Locality-Sensitive Hashing (LSH) [Gionis et al., 1999]
- Self-Organizing Maps (SOM) [Kohonen, 1998]
- Auto-encoders [Rumelhart et al., 1985]
- ...

Multidimensional Scaling [Torgerson, 1952]

Inverse Problem: Given a matrix of (dis-)similarities $D \in \mathbb{R}^{N \times N}$, find a (low-dimensional) embedding of N points.
Goal of MDS is to find coordinates of the data points in some subspace of $\mathbb{R}^n$, such that the given *proximities* are *preserved*.

A famous problem in cartography:
Find a 2-dimensional map of the earth, so that distances between cities are distorted as little as possible.
Notice that the original distances are not Euclidean, but *geodesic* (measured along the earth's surface).



Multi-dimensional Scaling (cont'd)

We are given an $N \times N$ matrix D of distances d_{ij} between all pairs of points. *Metric* MDS minimizes the distortion of distances in terms of a residual sum of squares, called the "stress":

$$stress(x_1, x_2, \ldots, x_N) = \sqrt{\frac{\sum_{i,j}(d_{ij} - ||x_i - x_j||)^2}{\sum_{i,j} d_{ij}^2}} \tag{2}$$

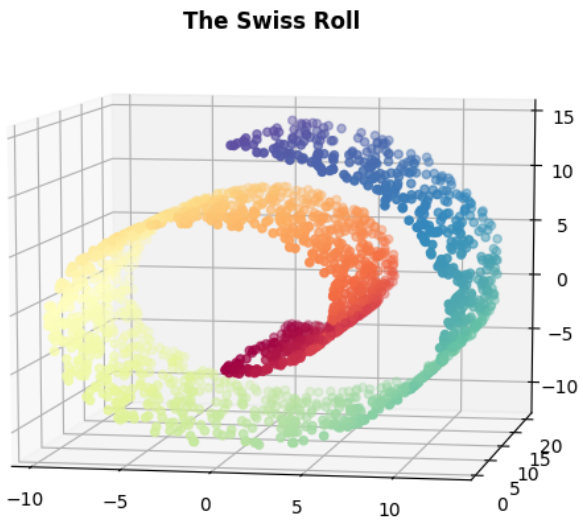
so

$$\{x_1, x_2, \ldots, x_N\}^* = \arg\min_{\{x_i\}} stress(x_1, x_2, \ldots, x_N) \tag{3}$$

- No unique solution. For example, all rotations of a solution would produce the same distances.
- MDS is often used for data visualization.

Manifold Learning

A standard approach to *non-linear* dimensionality reduction.
Main Idea: The data lie on a *surface* of dimension much lower than the original inputs.



Famous methods

- kernel-PCA [Schölkopf et al., 1997]
- Isomap [Tenenbaum et al., 2000]
- Locally-Linear Embedding (LLE) [Roweis and Saul, 2000]
- Laplacian Eigenmaps [Belkin and Niyogi, 2001]

Fun Fact: All these methods can be cast as kernel-PCA.

State-of-the-art method for high-dimensional data visualization: t-distributed Stochastic Neighbor Embedding (t-SNE) [Maaten and Hinton, 2008]

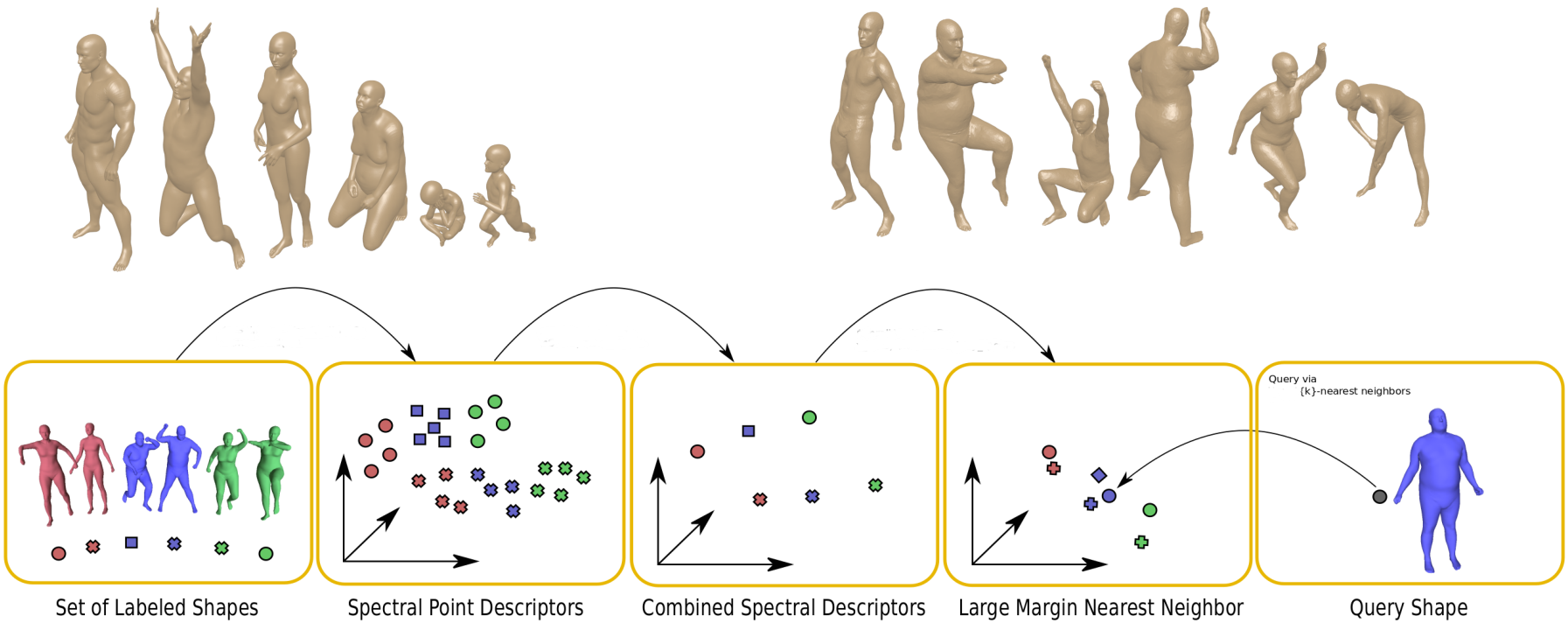
Outline

Non-rigid 3D Shape Retrieval via LMNN [Chiotellis et al., 2016]

Dataset: SHREC'14 [Pickup et al., 2014]

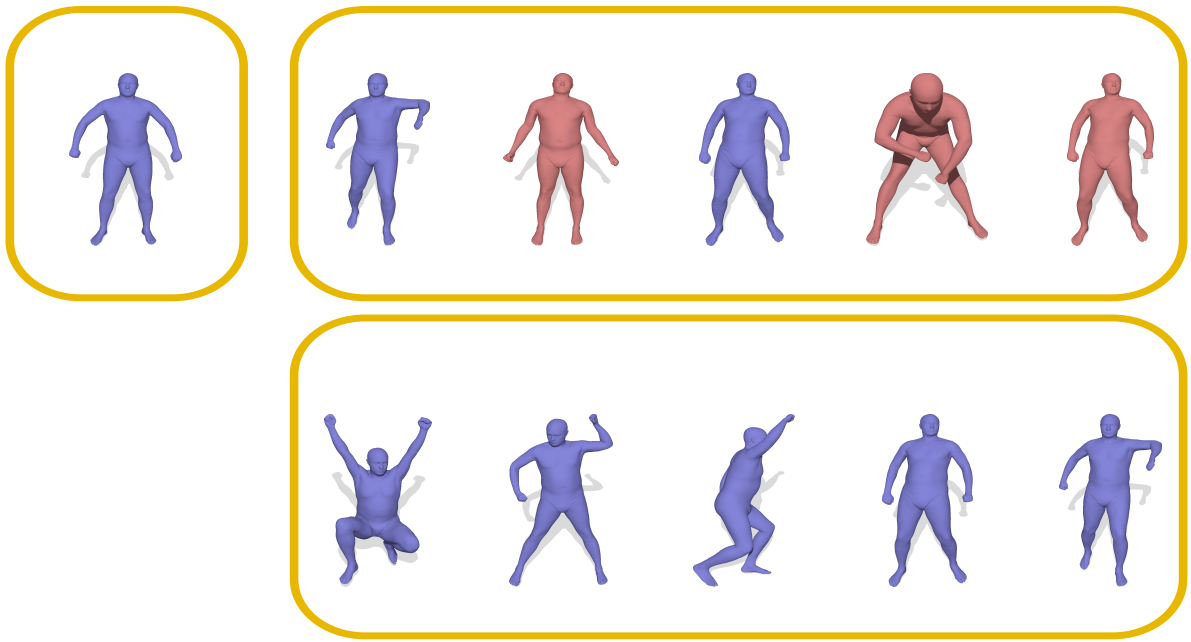
Synthetic dataset: 300 models
(**15** persons $\times$ 20 poses)

Real dataset: 400 models
(**40** persons $\times$ 10 poses)



Non-rigid 3D Shape Retrieval via LMNN (cont'd)

Retrieval Example



Top left: *A query model.*

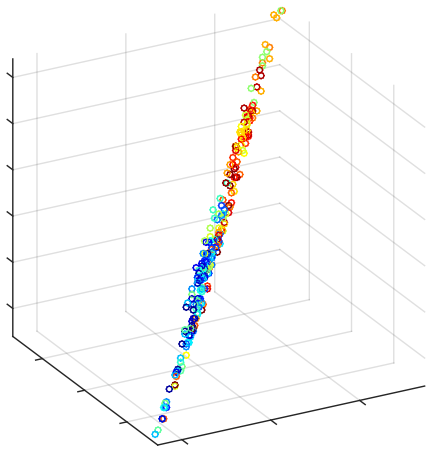
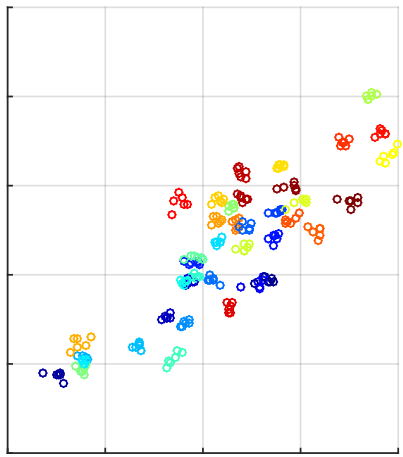
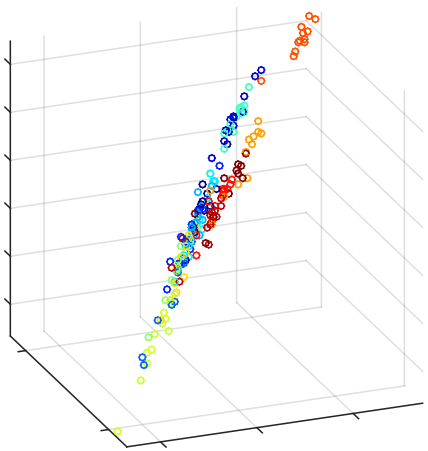
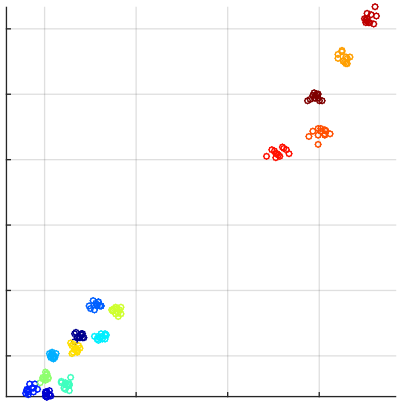
Top row: *5 best matches retrieved by the Supervised Dictionary Learning method [Litman et al., 2014].*

Bottom row: *5 best matches retrieved by the proposed method (CSD+LMNN).*

Blue: *Matches from query class.* ***Red:*** *Matches from other classes.*

Non-rigid 3D Shape Retrieval via LMNN (cont'd)

Embeddings Visualization

Dataset	$y_f(\mathcal{S})$ before learning	learned $L \cdot y_f(\mathcal{S})$
SHREC'14 Real		
SHREC'14 Synthetic		

Bibliography I

[Belkin and Niyogi, 2001] Belkin, M. and Niyogi, P. (2001).
Laplacian eigenmaps and spectral techniques for embedding and clustering.
In *NIPS*, volume 14, pages 585–591.

[Chiotellis et al., 2016] Chiotellis, I., Triebel, R., Windheuser, T., and Cremers, D. (2016).
Non-rigid 3d shape retrieval via large margin nearest neighbor embedding.
In *European Conference on Computer Vision*, pages 327–342. Springer.

[Chopra et al., 2005] Chopra, S., Hadsell, R., and LeCun, Y. (2005).
Learning a similarity metric discriminatively, with application to face verification.
In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 539–546. IEEE.

[Davis et al., 2007] Davis, J. V., Kulis, B., Jain, P., Sra, S., and Dhillon, I. S. (2007).
Information-theoretic metric learning.
In *Proceedings of the 24th international conference on Machine learning*, pages 209–216. ACM.

[Fisher, 1936] Fisher, R. A. (1936).
The use of multiple measurements in taxonomic problems.
Annals of eugenics, 7(2):179–188.

[Gionis et al., 1999] Gionis, A., Indyk, P., Motwani, R., et al. (1999).
Similarity search in high dimensions via hashing.
In *VLDB*, volume 99, pages 518–529.

[Globerson and Roweis, 2005] Globerson, A. and Roweis, S. (2005).
Metric learning by collapsing classes.
In *Nips*, volume 18, pages 451–458.

Bibliography II

[Goldberger et al., 2004] Goldberger, J., Hinton, G. E., Roweis, S. T., and Salakhutdinov, R. (2004).
Neighbourhood components analysis.
In *Advances in neural information processing systems*, pages 513–520.

[Hoffer and Ailon, 2015] Hoffer, E. and Ailon, N. (2015).
Deep metric learning using triplet network.
In *International Workshop on Similarity-Based Pattern Recognition*, pages 84–92. Springer.

[Jain et al., 2009] Jain, P., Kulis, B., Dhillon, I. S., and Grauman, K. (2009).
Online metric learning and fast similarity search.
In *Advances in neural information processing systems*, pages 761–768.

[Kohonen, 1998] Kohonen, T. (1998).
The self-organizing map.
Neurocomputing, 21(1):1–6.

[Kulis et al., 2006] Kulis, B., Sustik, M., and Dhillon, I. (2006).
Learning low-rank kernel matrices.
In *Proceedings of the 23rd international conference on Machine learning*, pages 505–512. ACM.

[Litman et al., 2014] Litman, R., Bronstein, A., Bronstein, M., and Castellani, U. (2014).
Supervised learning of bag-of-features shape descriptors using sparse coding.
In *Computer Graphics Forum*, volume 33, pages 127–136. Wiley Online Library.

[Maaten and Hinton, 2008] Maaten, L. v. d. and Hinton, G. (2008).
Visualizing data using t-sne.
Journal of Machine Learning Research, 9(Nov):2579–2605.

Bibliography III

[Pearson, 1901] Pearson, K. (1901).
On lines and planes of closest fit to system of points in space. philiosophical magazine, 2, 559-572.

[Pickup et al., 2014] Pickup, D., Sun, X., Rosin, P. L., Martin, R. R., Cheng, Z., Lian, Z., Aono, M., Ben Hamza, A., Bronstein, A., Bronstein, M., Bu, S., Castellani, U., Cheng, S., Garro, V., Giachetti, A., Godil, A., Han, J., Johan, H., Lai, L., Li, B., Li, C., Li, H., Litman, R., Liu, X., Liu, Z., Lu, Y., Tatsuma, A., and Ye, J. (2014).
SHREC'14 track: Shape retrieval of non-rigid 3d human models.
In *Proceedings of the 7th Eurographics workshop on 3D Object Retrieval*, EG 3DOR'14. Eurographics Association.

[Roweis and Saul, 2000] Roweis, S. T. and Saul, L. K. (2000).
Nonlinear dimensionality reduction by locally linear embedding.
science, 290(5500):2323–2326.

[Rumelhart et al., 1985] Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1985).
Learning internal representations by error propagation.
Technical report, DTIC Document.

[Schölkopf et al., 1997] Schölkopf, B., Smola, A., and Müller, K.-R. (1997).
Kernel principal component analysis.
In *International Conference on Artificial Neural Networks*, pages 583–588. Springer.

[Shental et al., 2002] Shental, N., Hertz, T., Weinshall, D., and Pavel, M. (2002).
Adjustment learning and relevant component analysis.
In *European Conference on Computer Vision*, pages 776–790. Springer.

Bibliography IV

[Tenenbaum et al., 2000] Tenenbaum, J. B., De Silva, V., and Langford, J. C. (2000).
A global geometric framework for nonlinear dimensionality reduction.
science, 290(5500):2319–2323.

[Torgerson, 1952] Torgerson, W. S. (1952).
Multidimensional scaling: I. theory and method.
Psychometrika, 17(4):401–419.

[Weinberger et al., 2005] Weinberger, K. Q., Blitzer, J., and Saul, L. K. (2005).
Distance metric learning for large margin nearest neighbor classification.
In *Advances in neural information processing systems*, pages 1473–1480.