Feature Selection

David Mount

For CMSC 828K:
Algorithms and Data Structures
for Information Retrieval

Introduction

Information retrieval for multimedia and other non-text databases:

- **Image and multimedia:** Images, video, audio.
- **Medical databases:** ECGs, X-rays, MRI scans.
- **Time series data:** Financial data, sales, meteorological, geological data.
- **Biology:** Genome, proteins.

Query-by-example: Find objects that are similar to a given query object.

Distance and Metrics

Similarity is defined in terms of a **distance function**. An important subclass are **metrics**.

Metric Space: A pair (X,d) where X is a set and d is a distance function such that for x,y in X:

$$d(x,y) = d(y,x) \geq 0$$

$$d(x,y) = 0 \Leftrightarrow x = y$$

$$d(x,y) + d(y,z) \geq d(x,z) \quad \text{(Triangle Inequality)}$$

Hilbert Space: A vector space and a norm $|v|$, which defines a metric $d(u,v)=|v-u|$.

Examples of Metrics

Minkowski L_p Metric: Let $X = \mathbf{R}^k$.

$$d(x,y) = \left[\sum_i |x_i - y_i|^p \right]^{1/p}$$

L_1 (Manhattan): $d(x,y) = \sum |x_i - y_i|$

L_2 (Euclidean): $d(x,y) = \sqrt{\sum (x_i - y_i)^2}$

L_{inf} (Chessboard): $d(x,y) = \max_i |x_i - y_i|$

More Metrics

Hamming Distance: Let $X = \{0,1\}^k$. Number of 1-bits in the exclusive-or

$$x \oplus y.$$

Edit Distance: Let $X = \Sigma^*$. Minimum number of single character changes (insert, delete, swap) to convert x into y.

Objective

Efficiency issues:

- **Edit distance** requires **quadratic time** to compute.
- **Data structures:** There are many data structures for storing vector data.
- **Curse of dimensionality:** Most nearest neighbor algorithms have running times that grow **exponentially** with dimension.

Objective: Embed objects into a low-dimensional space and use a Minkowski metric, allowing for a small **distortion** in distances.

Isometries and Embeddings

Isometry: A **distance preserving** mapping f from metric space (A, d_A) to (B, d_B):

$$d_B(f(x), f(y)) = d_A(x, y)$$

Contractive: A mapping f that **does not increase** distances:

$$\frac{1}{c} d_A(x, y) \le d_B(f(x), f(y)) \le d_A(x, y)$$

c is the **distortion** of the embedding.

Feature Selection

Main Problem: Given some finite metric space (X, d), where $|X|=n$, map it to some Hilbert space, say, $(\mathbf{R}^k, L_p)$. Ideally both the dimension and the distortion should be as low as possible.

Feature Selection: We can think of each coordinate of the resulting vector as describing some feature of the object. Henceforth, feature means coordinate.

Overview

We present methods for mapping a metric space (X, d) to $(\mathbf{R}^k, L_p)$. Let $|X| = n$.

- **Multidimensional Scaling:** A classical method for embedding metric spaces into Hilbert spaces.
- **Lipschitz Embeddings:** From any finite metric space to $(\mathbf{R}^k, L_p)$ where k is $O(\log^2 n)$ with $O(\log n)$ distortion.
- **SparseMap:** A practical variant of LLR embeddings.
- **KL-transform and FastMap:** Methods based on projections to lower dimensional spaces.

Multidimensional Scaling

Finding the best mapping of a metric space (X,d) to a k-dimensional Hilbert space $(\mathbf{R}^k,d')$ is a nonlinear optimization problem. The objective function is the **stress** between the distances:

$$\text{Stress}(d,d') = \sqrt{\frac{\sum \left(d'(f(x),f(y)) - d(x,y)\right)^2}{\sum d(x,y)^2}}$$

Standard incremental methods (e.g., **steepest descent**) can be used to search for a mapping of minimal stress.

Limitations

MDS has a number of limitations that make it difficult to apply to information retrieval.

- All **O(n^2) distances** in the database need to be computed in the embedding process. Too large for most real databases.
- In order to map a query point into the embedded space, a similar optimization problem needs to be solved, requiring computation of **O(n) distances**.

Lipschitz Embeddings

Each coordinate or **feature** is the distance to the closest point of a subset of X.

Bourgain (1985) Any n-point metric space (X, d) can be embedded into $O(\log n)$-dimensional Euclidean space with $O(\log n)$ distortion.

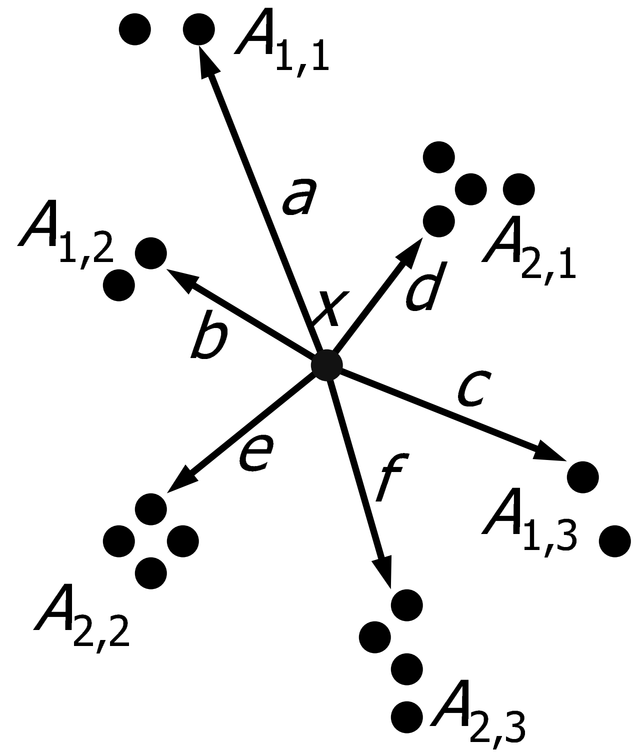
Linial, London, and Rabinovich (1995) Generalized to any L_p metric, and showed how it to compute it by a **randomized** algorithm. Dimension increases to $O(\log^2 n)$.

LLR Embedding Algorithm

Let $q = O(\log n)$. [Constant affects prob of success.]

For $i = 1,2,\ldots,\lg n$ do
  For $j = 1$ to q do
    $A_{i,j}$ = random subset of X of size 2^i.

Map x to the vector $\{d_{i,j}\}/Q$ where $d_{i,j}$ is the distance from x to the closest point in $A_{i,j}$ and Q is the total number of subsets.



$$x \to \frac{1}{6}(a,b,c,d,e,f)$$

Why does it work?

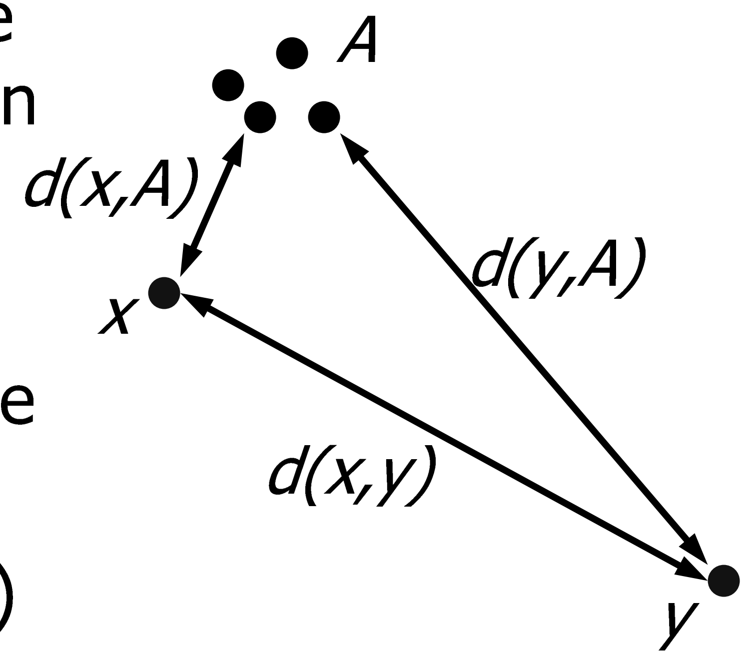
Assume L_1 metric.

Let $d(x,A)$ denote the distance from x to the closest point in A.

Observe that for any x,y and any subset A, by the triangle inequality we have

$$\left|d(x,A) - d(y,A)\right| \le d(x,y)$$

The mapping is **contractive**.



Why does it work?

Let $B(x, r_x)$ denote the set of points within distance r_x of x. Suppose $r_y < r_x$ and A is a subset such that

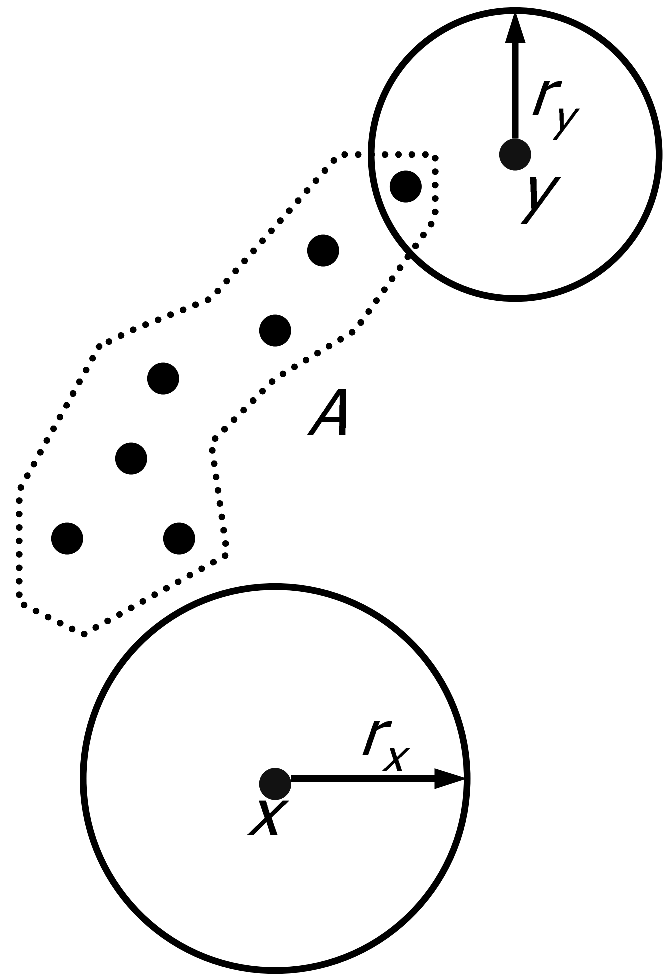
$$\left(A \cap B(x, r_x) = \emptyset\right) \wedge$$

$$\left(A \cap B(y, r_y) \neq \emptyset\right)$$

then

$$\left|d(x, A) - d(y, A)\right| \geq r_x - r_y$$

Will show this happens often.



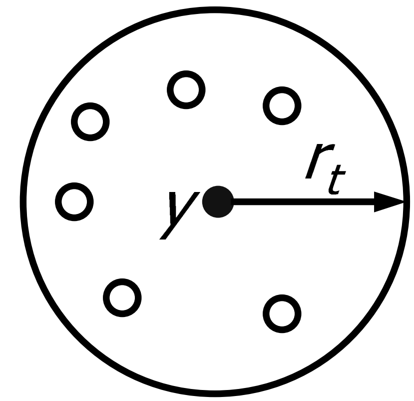
Why does it work?

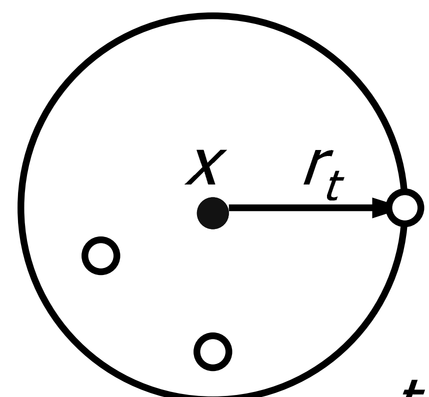
Fix points x and y. For t = 1,2,… let r_t be the smallest value such that

$$\min\left(\left|B(x, r_t)\right|, \left|B(y, r_t)\right|\right) \geq 2^t$$

Repeat as long as $r_t < d(x,y)/2$.

Note that for all r_t, $B(x, r_t)$ and $B(y, r_t)$ are disjoint.





$t = 2$

Why does it work?

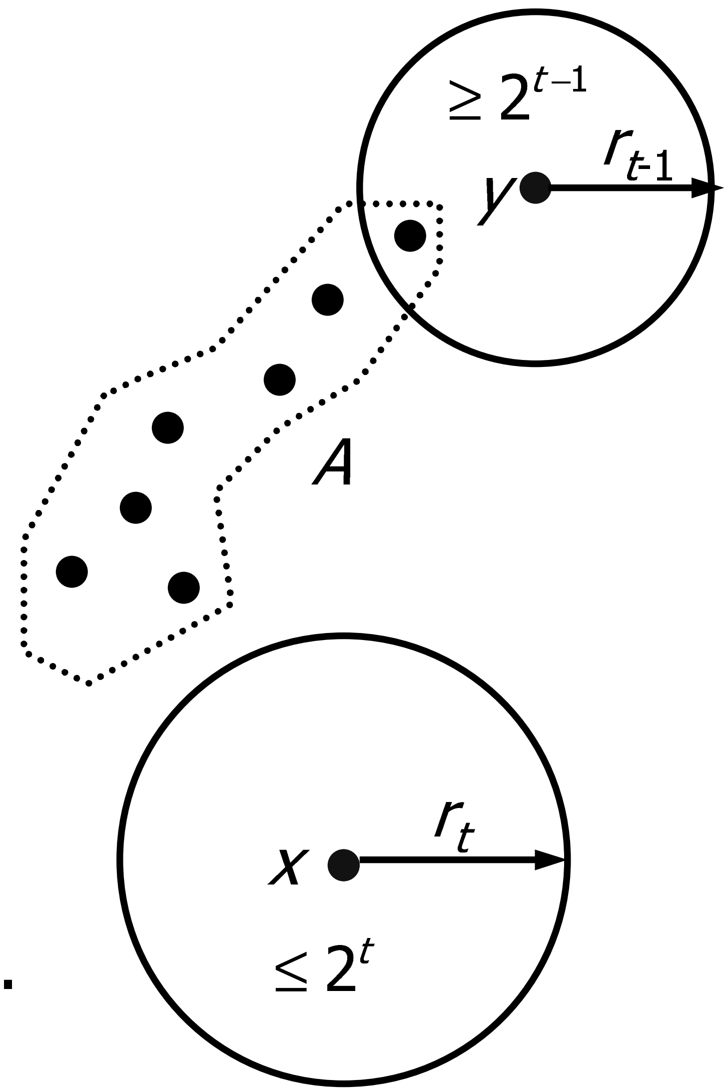
When the algorithm picks a random subset A of size roughly $n/2^t$, with constant probability (>1/10):

$$\left(A \cap B(x, r_t) = \emptyset\right) \wedge$$

$$\left(A \cap B(y, r_{t-1}) \neq \emptyset\right)$$

or vice versa.

Repeating this $q = O(\log n)$ times the expected number of such sets is at least $q/10$.



Why does it work?

Repeating over all q subsets in this group, with high probability we have:

$$\sum_{i=1}^{q} \left| d(x,A_i) - d(y,A_i) \right| \geq \frac{\log n}{10} \left(r_t - r_{t-1} \right)$$

Repeating this for all the size groups t=1,2,…, with high probability we have

$$d_{\mathrm{L1}}(x',y') = \sum_{i,j}^{Q} \left| d(x,A_{i,j}) - d(y,A_{i,j}) \right| / Q$$

$$\geq \frac{\log n}{10 \cdot Q} \sum_t \left(r_t - r_{t-1} \right) \geq \frac{\log n}{20 \cdot Q} d(x,y)$$

Review of the proof

Since $Q = O(\log^2 n)$ the final distortion is $O(\log n)$.

When we sample subsets of size $n/2^t$, we have a **constant probability** of finding subsets whose corresponding coordinate differs by at least $r_t - r_{t-1}$.

We repeat the sample $O(\log n)$ times so that this occurs with **high probability**. There are $O(\log n)$ size groups so the total dimension is $O(\log^2 n)$.

The sum of the $r_t - r_{t-1}$ terms **telescope** totaling to $d(x,y)/2$, and repeated sampling adds an $O(\log n)$ factor.

How low can you go?

Folklore: Any n-point metric space can be embedded into $(\mathbf{R}^n, L_{inf})$ with distortion **1**. (A point is mapped to a vector of distances to the other points.)

Bourgain: There is an n-point metric space such that any embedding in $(\mathbf{R}^k, L_2)$ has distortion of at least $\mathbf{O(\log n / \log \log n)}$.

Johnson-Lindenstrauss: You can embed n points in $(\mathbf{R}^t, L_2)$ into $(\mathbf{R}^k, L_2)$ where $k = O((\log n)/\varepsilon^2)$ with distortion $\mathbf{1+}\varepsilon$. (Select k unit vectors from $\mathbf{R}^t$ at random. Each coordinate is the length of the orthogonal projection onto each vector.)

Limitations

Limitations of the LLR embedding:

- **$O(\log n)$ distortion**: Experiments show that the actual distortions may be much smaller.
- **$O(\log^2 n)$ dimension**: This is a real problem.
- **$O(n^2)$ distance computations** must be performed in the process of embedding and embedding a query point requires **$O(n)$ distance computations**: Too high if distance function is complex.

SparseMap

SparseMap (Hristescu and Farach-Colton) is a variant of the LLR-embedding:

- **Incremental construction of features:** Once the first k coordinates have been computed, we can use these (rather than the distance function) to perform…
- **Distance Approximation:** For computing the distance from a point x to a subset $A_{i,j}$.
- **Greedy Resampling:** Keep only the best (most discriminating) coordinates.

Distance Approximation

Suppose that the first k' coordinates have been selected. This gives a partial distance estimate

$$d'_{k'}(x,y) = \sqrt{\sum_{i=1}^{k'} (x_i - y_i)^2}$$

To compute $d(x,A)$ for a new subset A:

- Compute $d'_{k'}(x,y)$ for each y in A.
- Select y with the smallest such distance estimate.
- Return $d(x,y)$.

Only 1 distance calculation, as opposed to $|A|$. Total number is of distances is $O(kn)$.

Greedy Resampling

Suppose that k coordinates have been computed. Want to keep only the **best** coordinates.

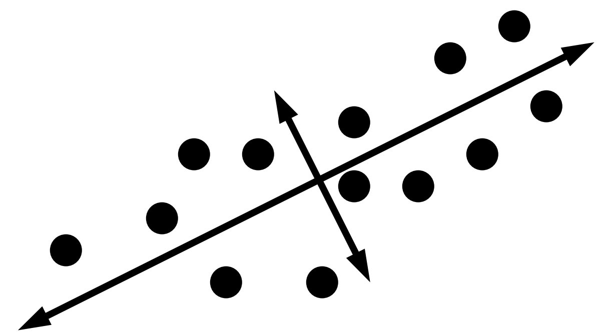
- **Sample** some number of random pairs (x,y) from the database, and compute their distances $d(x,y)$.
- **Select** the coordinate (subset A) that minimizes the **stress** between the true and estimated distances:

$$\text{Stress}(d,d') = \sqrt{\frac{\sum (d(x,y) - d'(x,y))^2}{\sum d'(x,y)^2}}$$

- **Repeat** selecting the coordinate which together with the previous coordinates minimizes stress.

KL-Transform

Karhunen-Loeve transform:
Given *n* points (vectors) in $\mathbf{R}^t$, project them into a lower dimensional space $\mathbf{R}^k$, so as to minimize the mean squared error.



KL-Transform

- Translate the points so that their mean coincides with the origin.
- Let X denote the matrix whose columns are the resulting vectors. Compute the **covariance matrix** $\Sigma = XX^T$.
- Compute the **eigenvectors** Φ_i and the **eigenvalues** λ_i of Σ in decreasing order.
- Project the columns of X orthogonally onto the subspace spanned by Φ_i, $i = 1,\ldots,k$.

FastMap

The KL-transform assumes that points are in $\mathbf{R}^t$. What if we have a finite metric space (X, d)?

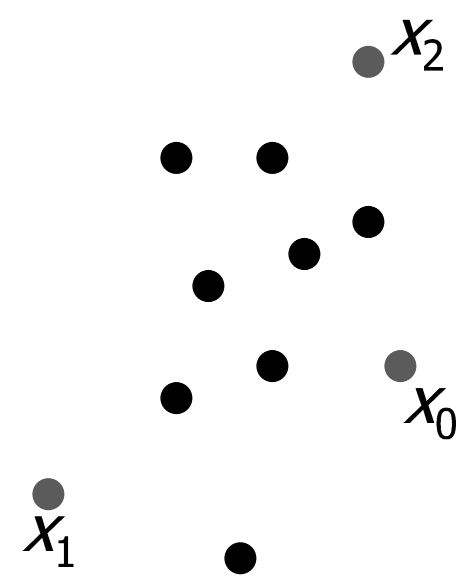
Faloutsos and Lin (1995) proposed FastMap as metric analogue to the KL-transform. Imagine that the points are in a Euclidean space.

- Select two **pivot points** x_a and x_b that are far apart.
- Compute a **pseudo-projection** of the remaining points along the "line" $x_a x_b$.
- **"Project"** the points to an orthogonal subspace and **recurse**.

Selecting the Pivot Points

The pivot points should lie along the principal axes, and hence should be far apart.

- Select any point x_0.
- Let x_1 be the furthest from x_0.
- Let x_2 be the furthest from x_1.
- Return (x_1, x_2).



Pseudo-Projections

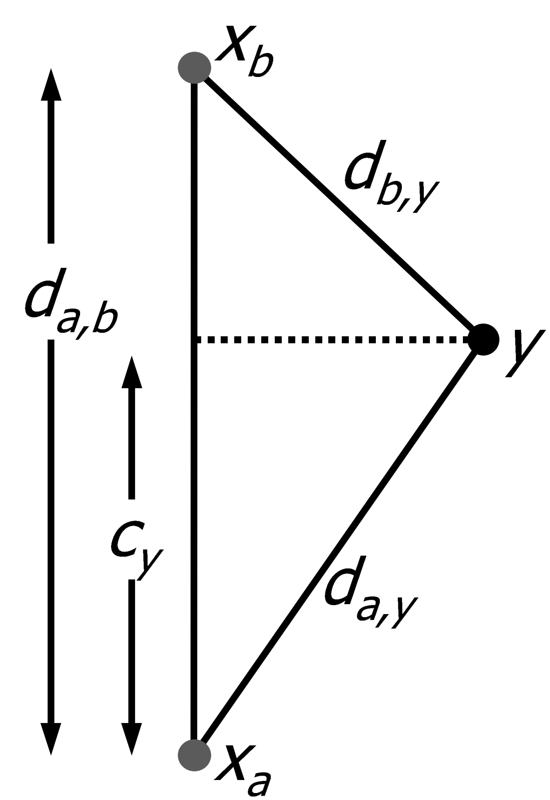
Given pivots (x_a, x_b), for any third point y, we use the **law of cosines** to determine the relation of y along $x_a x_b$.

$$d_{by}^2 = d_{ay}^2 + d_{ab}^2 - 2c_y d_{ab}$$

The **pseudo-projection** for y is

$$c_y = \frac{d_{ay}^2 + d_{ab}^2 - d_{by}^2}{2d_{ab}}$$

This is first coordinate.

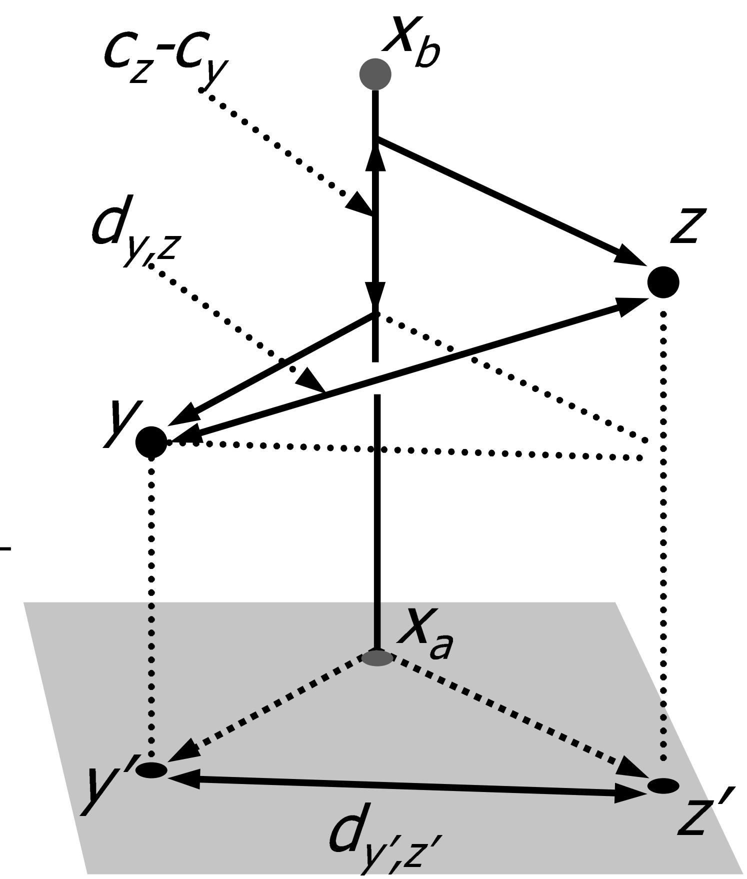


"Project to orthogonal plane"

Given distances along $x_a x_b$ we can compute distances within the "orthogonal hyperplane" using the Pythagorean theorem.

$$d'(y',z') = \sqrt{d^2(y,z) - (c_z - c_y)^2}$$

Using $d'(.,.)$, recurse until k features chosen.



Experimental Results

Faloutsos and Lin present experiments showing that FastMap is **faster** than MDS for a given level of stress, but has **much higher** stress than MDS. All data sets were relatively small.

Hristescu and Farach-Colton present experiments on protein data showing that SparseMap is considerably **faster** than FastMap, and it performs marginally **better** with respect to stress and cluster preservation.

Summary

- **Lipschitz Embeddings:** Embedding metric spaces using distances to subsets.
  - $O(\log^2 n)$ dimensions and $O(\log n)$ distortion.
  - Likely the best from a theoretical point of view.
  - SparseMap, a practical variant of this idea.
- **Johnson-Lindenstrauss:** Can embed any *n*-point set in Euclidean space into $O(\log n)$ dimensions.
- **KL-transform:** Embedding that minimizes squared errors. FastMap mimics this idea in metric spaces.

Bibliography

G. Hjaltson and H. Samet, "Contractive embedding methods for similarity searching in general metric spaces", manuscript.

J. Bourgain, "On Lipschitz embedding of finite metric spaces in Hilbert space", *Israel J. of Math*., 52, 1985, 46-52.

G. Hristescu and M. Farach-Colton, "Cluster preserving embeddings of proteins", DIMACS Tech. Rept. 99-50.

Bibliography

N. Linial, E. London and Y. Rabinovich, "The Geometry of Graphs and some of its algorithmic applications", *Combinatorica*, 15, 1995, 215-245.

C. Faloutsos and K.-I. Lin, "FastMap: A Fast Algorithm for Indexing, Data-Mining and Visualization of Traditional and Multimedia Datasets", *Proc. ACM SIGMOD*, 1995, 163-174.

K. Fukunaga, *Introduction to Statistical Pattern Recognition*, Academic Press, 1972.

F. W. Young and R. M. Hamer, *Multidimensional Scaling: History, Theory and Applications*, Lawrence Erlbaum Associates, Hilldale, NJ, 1987.