Goal: Compare different out-of-sample extensions for CMDS

CMDS attempts to preserve inter-point distances in a lower dimension

Spectral solution

Given a squared distance matrix $D^{(2)} \in \mathbb{R}^{n\times n}$ (Euclidean distances)

1) Apply a double-centering
$$B = -\frac{1}{2} C D^{(2)} C,$$
where $C = I - ee^{T}/n$ for $e = (1, \ldots, 1)^{T}$.

2) Take the k largest eigenvalues and corresponding eigenvalues of B. The rank K SVD is
$$B = U_k \Lambda_k U_k^{T} = (U_k \Lambda^{1/2})(\Lambda^{1/2} U_k^{T}) = Y^{T} Y$$
for $Y = \Lambda_k^{1/2} U_k^{T}$, where
$$\Lambda_k^{1/2} = \mathrm{diag}(\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_k}) \text{ and } U_k = [e_1 | \cdots | e_k].$$

Constraint $\sum y_i = 0$. Assume a solution exists, i.e. $\|y_i - y_j\| = d_{ij}$ for $i, j \in [n]$. Then
$$d_{ij}^2 = \|y_i - y_j\|^2 = (y_i - y_j)^{T}(y_i - y_j)$$
$$= \|y_i\|^2 + \|y_j\|^2 - 2 y_i^{T} y_j$$
$$\Rightarrow \sum_i d_{ij}^2 = \sum_i \|y_i\|^2 + n\|y_j\|^2 - 2 \sum_i y_i^{T} y_j$$
$$= \sum_i \|y_i\|^2 + n\|y_j\|^2$$

Similarly,

$$\sum_j d_{ij}^2 = n\|y_i\|^2 + \sum_j \|y_j\|^2$$

Now

$$d_{..}^2 = \sum_i \sum_j d_{ij}^2 = \sum_i n\|y_i\|^2 + \sum_j \|y_j\|^2$$

$$= n\sum_i \|y_i\|^2 + n\sum_j \|y_j\|^2$$

$$= 2n \sum_t \|y_t\|^2$$

$$\Rightarrow \sum_t \|y_t\|^2 = \frac{1}{2n} d_{..}^2$$

$$\|y_j\|^2 = \frac{1}{n} d_{.j}^2 - \frac{1}{2n^2} d_{..}^2$$

$$\|y_i\|^2 = \frac{1}{n} d_{i.}^2 - \frac{1}{2n^2} d_{..}^2$$

Finally,

$$y_i^T y_j = \underbrace{\frac{1}{2}\left(\frac{1}{n} d_{i.}^2 + \frac{1}{n} d_{.j}^2 - \frac{1}{n^2} d_{..}^2 - d_{ij}^2\right)}_{:= g_{ij}}$$

Gram matrix $G = (g_{ij}) \in \mathbb{R}^{n \times n}$

embedding matrix $Y = [y_1 | \cdots | y_n]^T \in \mathbb{R}^{n \times k}$

$YY^T = G$

Properties of G:

1) $G = -\frac{1}{2} C D^{(2)} C$ for $C = I_n - ee^T/n$ $D^{(2)} = (d_{ij}^2)$

2) rows and columns of G sum to 0

$\Rightarrow G$ has an eigenvalue of 0

In order for a solution to exist, G must be positive definite so we can write

$$G = U \Lambda U^T = \underbrace{(U \Lambda^{1/2})}_{Y} \underbrace{(\Lambda^{1/2} U^T)}_{Y^T}$$

Thm. If $G = -\frac{1}{2} C D^{(2)} C$ is positive semidefinite and $k \geq r = \text{rank}(G)$, then the CMDS problem has the following exact solution.

$$Y = U \Lambda_k^{1/2}$$

$$= [\sqrt{\lambda_1} e_1 | \cdots | \sqrt{\lambda_r} e_r | \underbrace{\sqrt{\lambda_{r+1}}}_{=0} e_{r+1} | \cdots | \underbrace{\sqrt{\lambda_k}}_{=0} e_k]$$

where (λ_i, e_i) are the eigenpairs of G.

Note. In practice, choosing a low dimension k a priori will leave no exact solution. Hence we use the least k-rank approximation of G via Eckart-Young:

$$G^k = U \Lambda_k U^T \Rightarrow Y_k = U \Lambda_k^{1/2}$$

Note. The map

$$D^{(2)} \mapsto -\frac{1}{2} C D^{(2)} C =: G$$

transforms squared distances into inner products.

Minimization Solution

$$\underset{y_i \in \mathbb{R}^k}{\operatorname{argmin}} \sum_{i,j} (\|y_i - y_j\| - d_{ij})^2$$

constrained to $\sum y_i = 0$
- optional, can translate projections

Out-of-Sample Strategies

Option 1a: minimize loss (one point)

Given data X and their projections Y, a new point x' is mapped

$$x' \mapsto y' = \underset{y \in \mathbb{R}^k}{\arg\min} \sum_i (\|y - y_i\| - d(x', x_i))^2$$

Pros:
- intuitive
- effectively preserves distances
- constructs an embedding function on the entire data space

Cons:
- $O(n)$
- doesn't preserve the distances between out-of-sample points ☆
- minimization problem may have local minima

Option 1b: minimize loss (k points)

Given data X and their projections Y, a new set of k points $x_1', \dots, x_k'$ is mapped so that

$$2 \sum_{i=1}^{n} \sum_{j=1}^{k} \left(\| y_i - y_j' \| - d(x_i, x_j') \right)^2 + \sum_{i=1}^{k} \sum_{j=1}^{k} \left(\| y_i' - y_j' \|^2 - d(x_i', x_j') \right)^2$$

is minimized.

Pros:

- also preserves distances between new points

Cons:

- doesn't give a well-defined embedding function
- $O(nk + k^2)$
- minimization potentially time-consuming

Option 2: spectral method that preserves the centroid of the original samples (Trosset)

Ex. $k=1$

$\Delta_2 = (\delta_{ij}^2)$ squared distances between original n objects

$a_2 \in \mathbb{R}^n$ squared distances of the new object from the original n objects

$$A_2 = \begin{bmatrix} \Delta_2 & a_2 \\ a_2^t & 0 \end{bmatrix}$$

Def. Let $w \in \mathbb{R}^m$. $x_1, \ldots, x_m \in \mathbb{R}^d$ is <u>w-centered</u> if $\sum_{i=1}^m w_i x_i = 0$.

For w such that $e^t w \neq 0$, define

$$T_w(A_2) = -\frac{1}{2}\left(I - \frac{ew^t}{e^t w}\right) A_2 \left(I - \frac{we^t}{e^t w}\right).$$

Note the case $w = e$ is standard CMDS.

Thm. For any $w \notin e^\perp$, the $m \times m$ dissimilarity matrix A_2 is an EDM with embedding dimension p <=> there exists a w-centered spanning set $\{y_1, \ldots, y_m\}$ for which

$$T_w(A_2) = (y_i^t y_j).$$

Let $e = (1, \ldots, 1)^t \in \mathbb{R}^n$ and $f = (e^t, 1)^t \in \mathbb{R}^{n+1}$. Applying CMDS to A_2 approximates the fallible inner products $\tau_f(A_2)$, i.e. with respect to the centroid of all $n+1$ objects.

Since the out-of-sample problem requires us to maintain the original set of inner products, we instead set $w = (e^t, 0)^t$ and approximate the fallible inner products

$$B = \tau_w(A_2) = \begin{bmatrix} \tau_e(\Delta_2) & b \\ b^t & \beta \end{bmatrix}$$

with $x_1, \ldots, x_n \in \mathbb{R}^d$ fixed. This results in the nonlinear optimization problem

$$\min_{y \in \mathbb{R}^d} \left\| B - \begin{bmatrix} x_1^t \\ \vdots \\ x_n^t \\ y^t \end{bmatrix} \begin{bmatrix} x_1 & \cdots & x_n & y \end{bmatrix} \right\|^2$$

$$= \min_{y \in \mathbb{R}^d} 2 \sum_{i=1}^{n} (b_i - x_i^t y)^2 + (\beta - y^t y)^2 .$$

If the term $(\beta - y^t y)^2$ is dropped, the function that remains is convex. Furthermore, if X is full rank (otherwise a smaller d would suffice), the solution is

$$\hat{y} = (X^t X)^{-1} X^t b .$$

Pros:

- approximate out-of-sample solution is easily computable
  - could derive closed form

Cons:

- approximates inner products rather than distances

Big Idea: We want to compute the inner products for the new points with respect to the centroid of the projections of the original data. The projections of the new points are chosen to approximate this extended Gram matrix.

Questions:

- Does approximate inner products lead to approximate distances? (*)
- When approximating B, how much does fixing $x_1, \ldots, x_n$ raise the minimum?
  - $x_1, \ldots, x_n$ best approximates $T_c(\Delta_2)$ but the best approximation of $T_w(A_2)$ could be a completely different set of points.
  - $$\min_{y \in \mathbb{R}^d} \left\| B - \begin{bmatrix} x_1^+ \\ \vdots \\ x_n^+ \\ y^+ \end{bmatrix} \begin{bmatrix} x_1 & \cdots & x_n & y \end{bmatrix} \right\|^2$$

$$\text{vs.} \quad \underbrace{\min_{Y \in \mathbb{R}^{p \times d}} \left\| B - YY^+ \right\|^2}_{\substack{\text{not quite CMDS of } A_2 \text{ because} \\ B \text{ represents inner products with} \\ \text{respect to centroid of original} \\ \text{data, not ALL data}}} \leftarrow \text{spectral solution?}$$

(*) $\|x-y\|^2 = \langle x,x \rangle + \langle y,y \rangle - 2\langle x,y \rangle$

Option 3: Eigenmap Extension (Bengio)

Consider a Hilbert space $\mathcal{H}_p$ of functions with
$\langle f, g \rangle_p = \int f(x) g(x) p(x)\, dx$
for density p. Associate with kernel K a linear operator K_p in $\mathcal{H}_p$:

$$(K_p f)(x) = \int K(x,y) f(y) p(y)\, dy$$

In practice, we use an "empirical" Hilbert space $\mathcal{H}_{\hat{p}}$.

Prop. Let $\tilde{K}$ be a kernel (not necessarily positive semi-definite) that gives rise to a symmetric matrix $\tilde{M}_{ij} = \tilde{K}(x_i, x_j)$ upon a dataset $D = \{x_1, \dots, x_n\}$. Let (λ_k, v_k) be the eigenpairs of $\tilde{M}$ and (λ_k', f_k) the eigenpairs of $\tilde{K}_{\hat{p}}$. Let $e_k(x) = y_k(x)\sqrt{\lambda_k}$ denote the embedding associated with a new point x. Then

a) $\lambda_k' = \frac{1}{n}\lambda_k$

b) $f_k(x) = \frac{\sqrt{n}}{\lambda_k} \sum_{i=1}^{n} v_{ki} \tilde{K}(x, x_i)$

c) $f_k(x_i) = \sqrt{n}\, v_{ki}$

d) $y_k(x) = \frac{f_k(x)}{\sqrt{n}} = \frac{1}{\lambda_k} \sum_{i=1}^{n} v_{ki} \tilde{K}(x, x_i)$

e) $y_k(x_i) = y_{ik}$, $e_k(x_i) = e_{ik}$

where does this come from?

Note. Eigenvectors / Eigenfunctions assumed to have unit norm and eigenvectors in non-increasing order.

For MDS, a normalized kernel can be defined by

$$\tilde{k}(a,b) = -\frac{1}{2}\left(d^2(a,b) - \mathbb{E}_x[d^2(x,b)] - \mathbb{E}_{x'}[d^2(a,x')] + \mathbb{E}_{x,x'}[d^2(x,x')]\right).$$

This is a continuous version of the double-centering formula.

Big Idea.