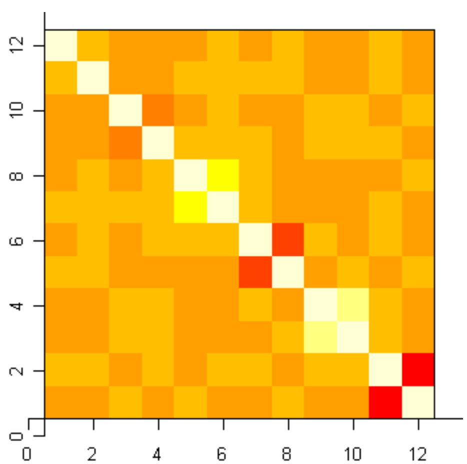
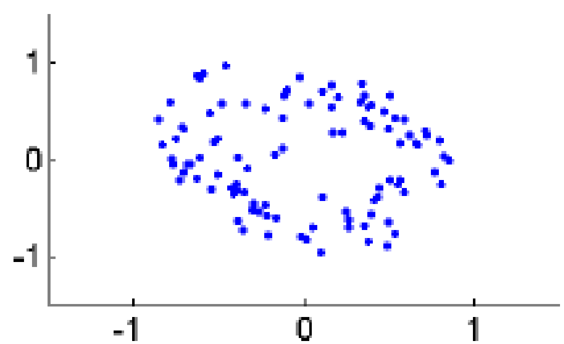
MDS Embedding

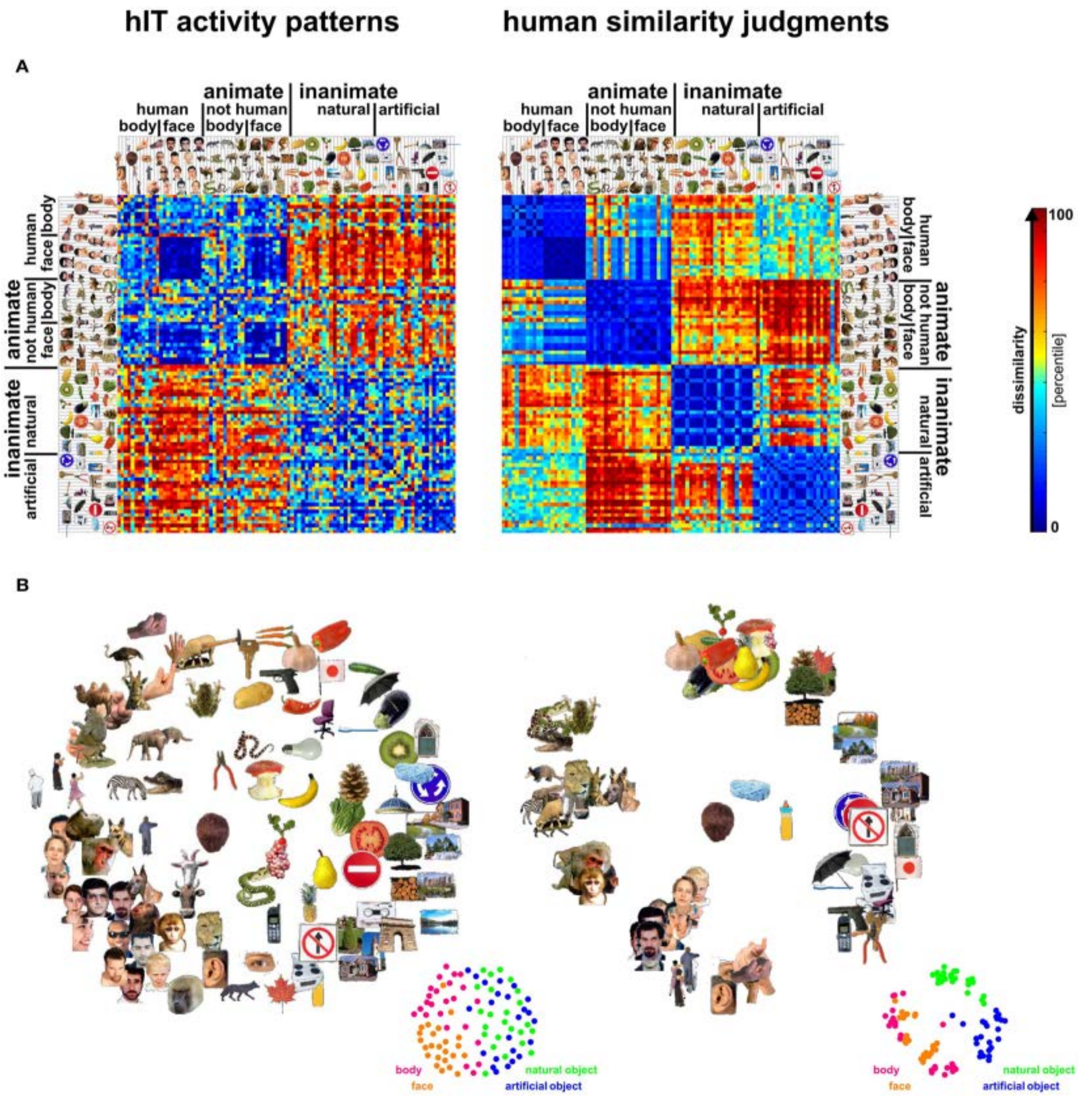
MDS takes as input a distance matrix D, containing all $N \times N$ pair of distances between elements xi, and embed the elements in N dimensional space such that the inter distances Dij are preserved as much as possible by $||xi - xj||$ in the embedded space.

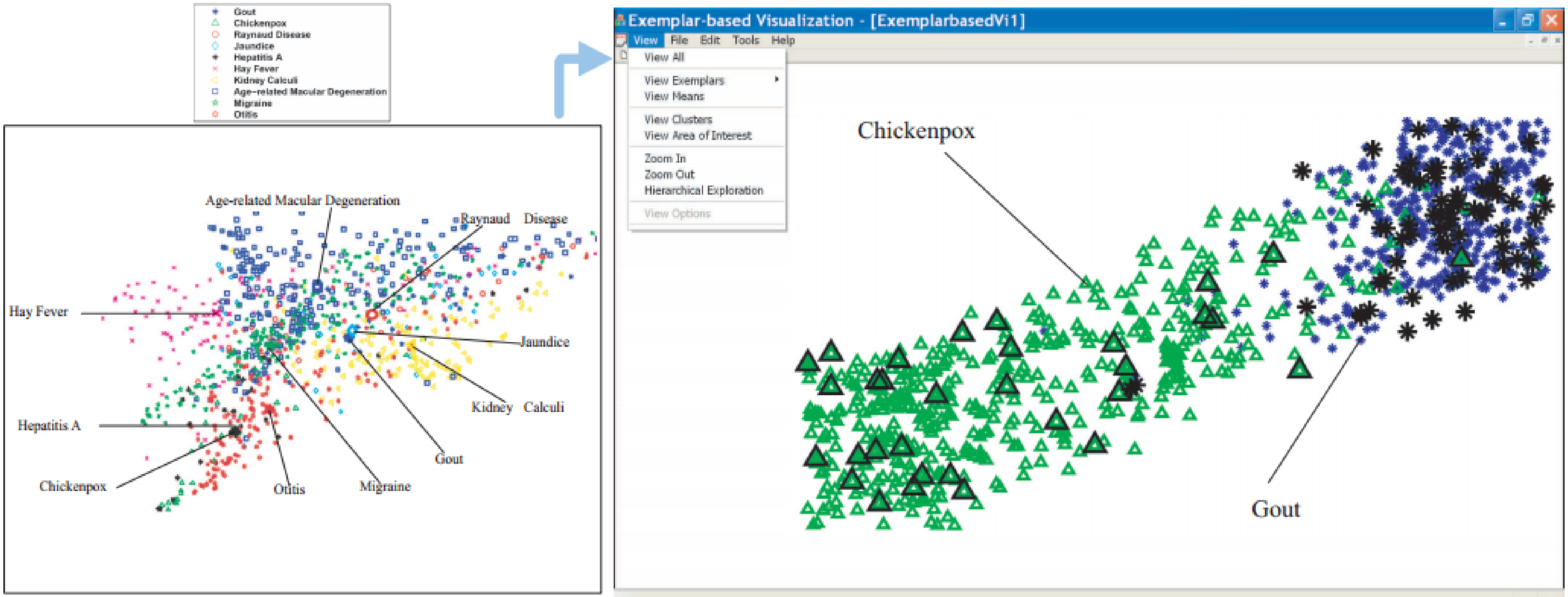




	Banff	Calgary	Columbia Icefield	Edmonton	Field, B.C.	Jasper	Lake Louise	Radium Hot Springs	Golden	Revelstoke
Calgary	128									
Columbia Icefield	188	316								
Edmonton	423	295	461							
Field, B.C.	85	213	157	508						
Jasper	291	419	100	361	260					
Lake Louise	58	186	130	481	27	233				
Radium Hot Springs	132	260	261	555	157	361	130			
Golden	134	262	207	557	49	307	76	105		
Revelstoke	282	410	355	705	197	455	224	253	148	
Vancouver	856	984	928	1279	771	798	794	818	713	565

Distances shown are in Kilometres.
To convert to miles multiply by 0.6

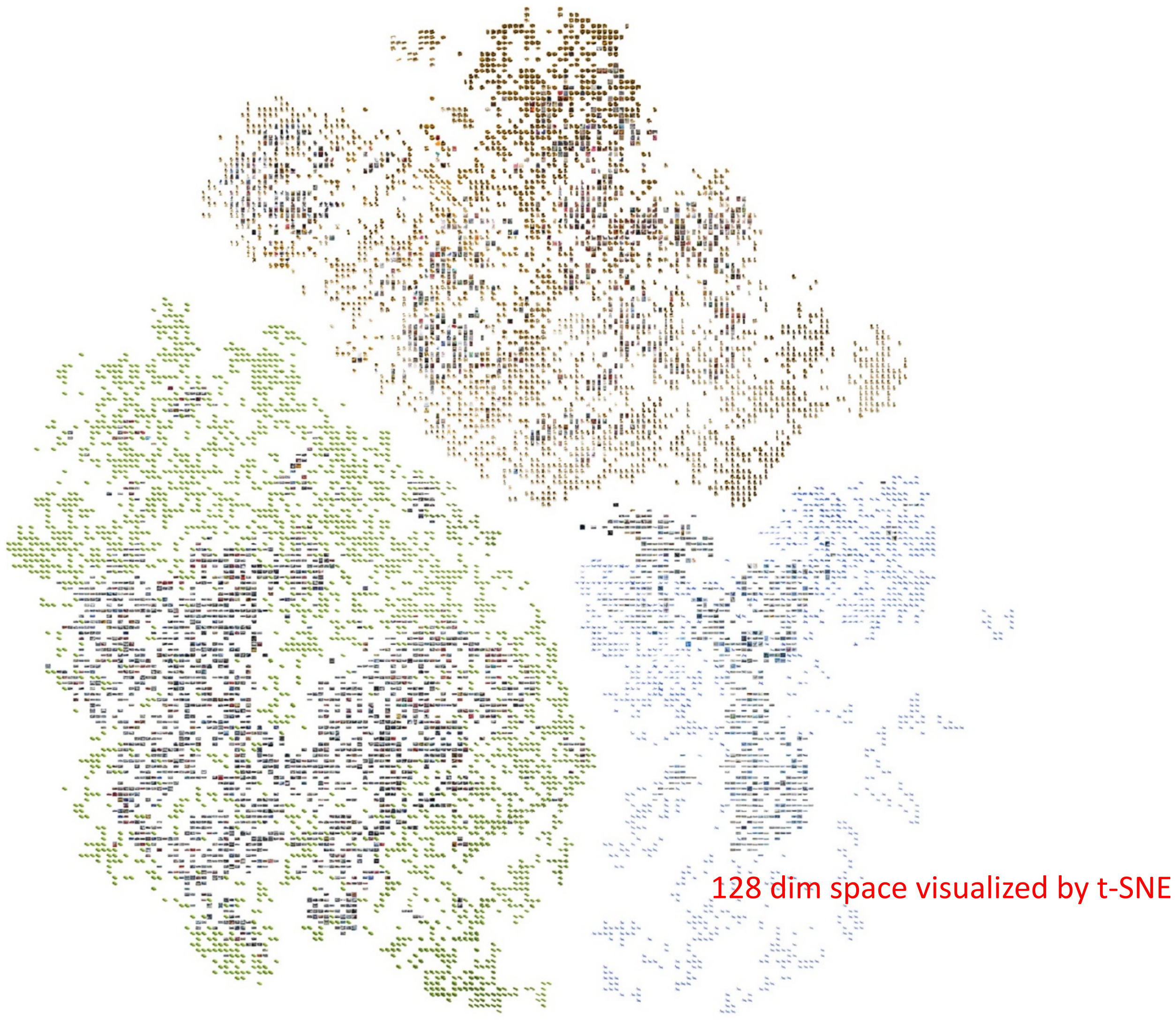

hIT activity patterns
human similarity judgments
A
animate | inanimate
human | not human
body | face | body | face
natural | artificial
dissimilarity
[percentile]
100
0
B
body
face
natural object
artificial object



(a) (b)

Fig. 5. Visualization of abstracts in *10PubMed* (15,565 documents, 10 topics) by EV. Each point represents an abstract; each color shape represents a disease; and the corresponding big color shape indicates the means of an abstract group. Visualization of (a) 1000 exemplars with their means, (b) two distinct groups of diseases: “Gout” and “Chickenpox” with the selected exemplars (100 in total), emphasized by the bigger black shapes.

Joint Embeddings of Shapes and Images



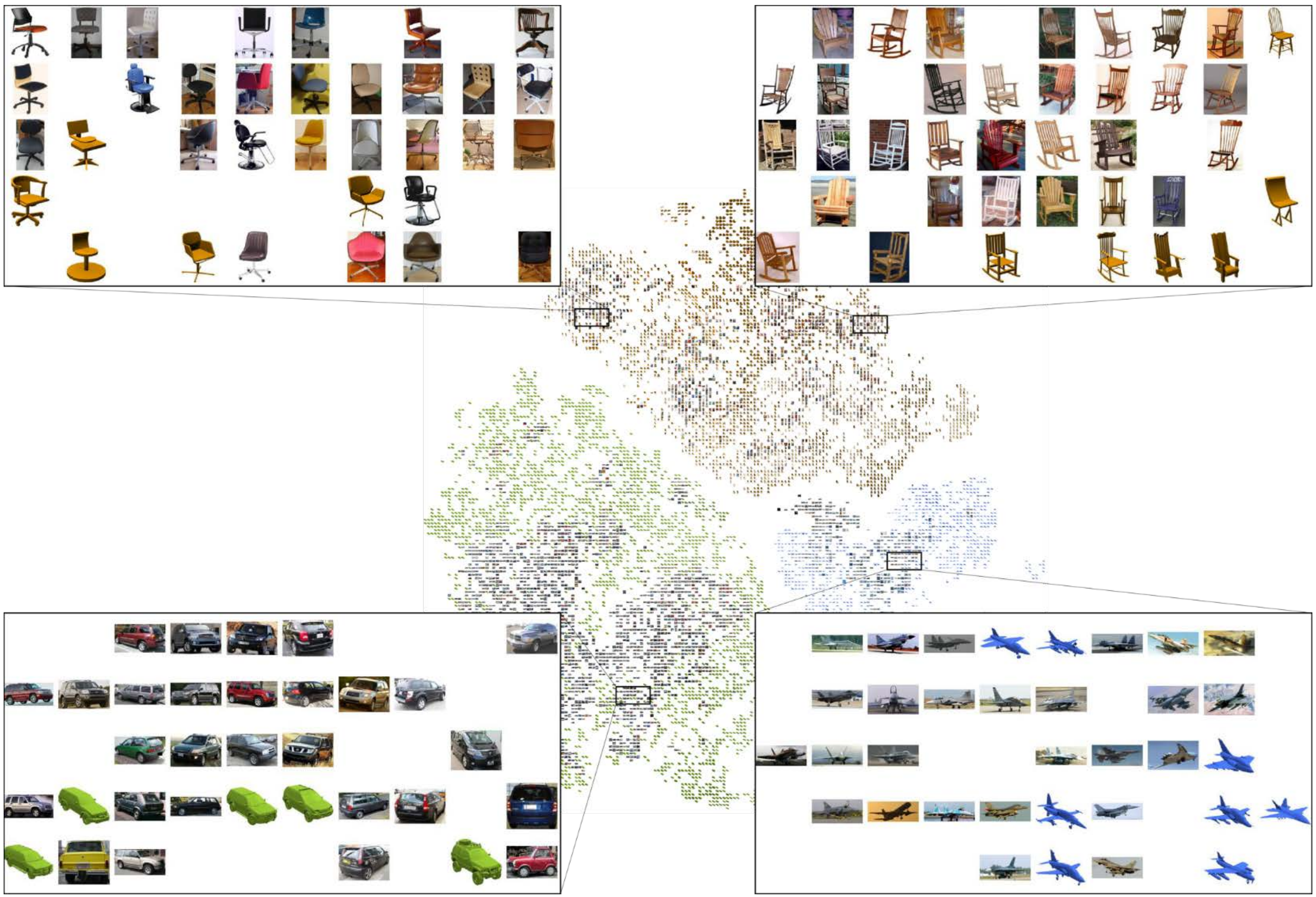
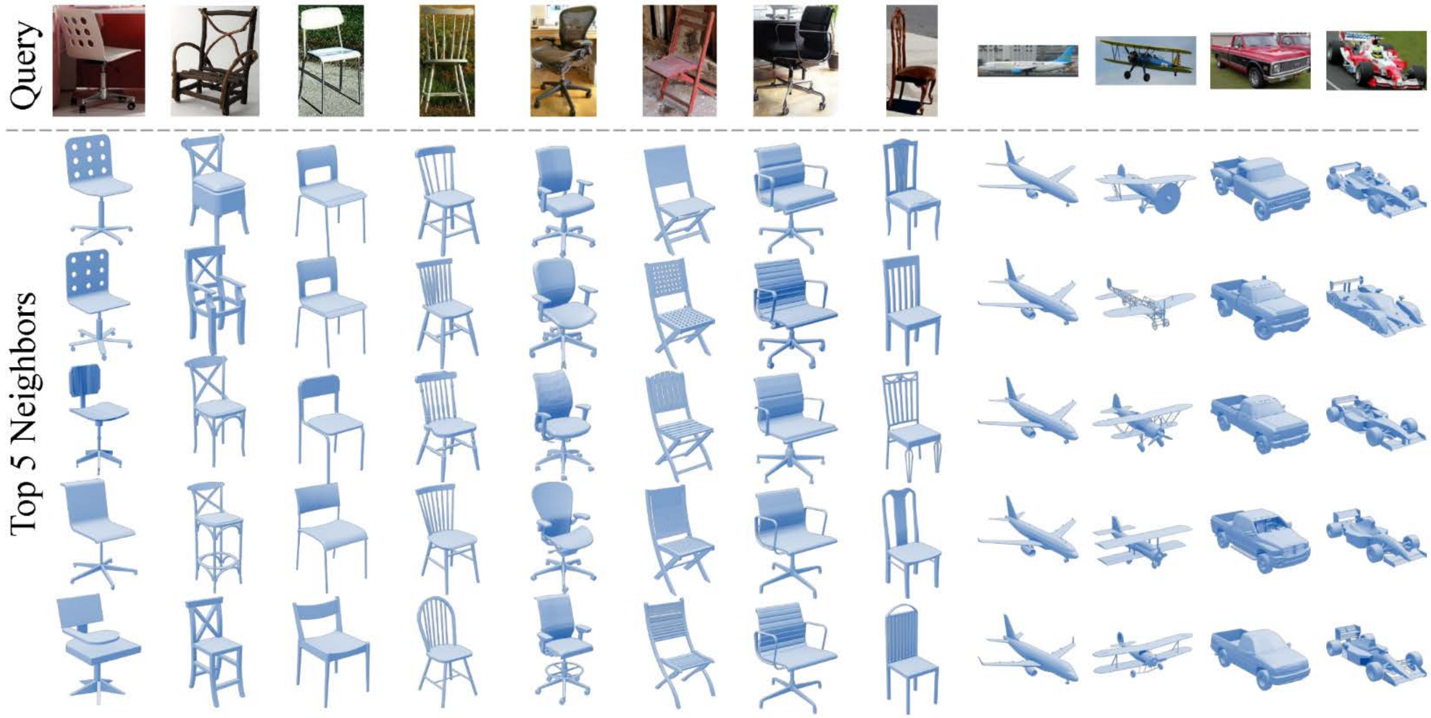
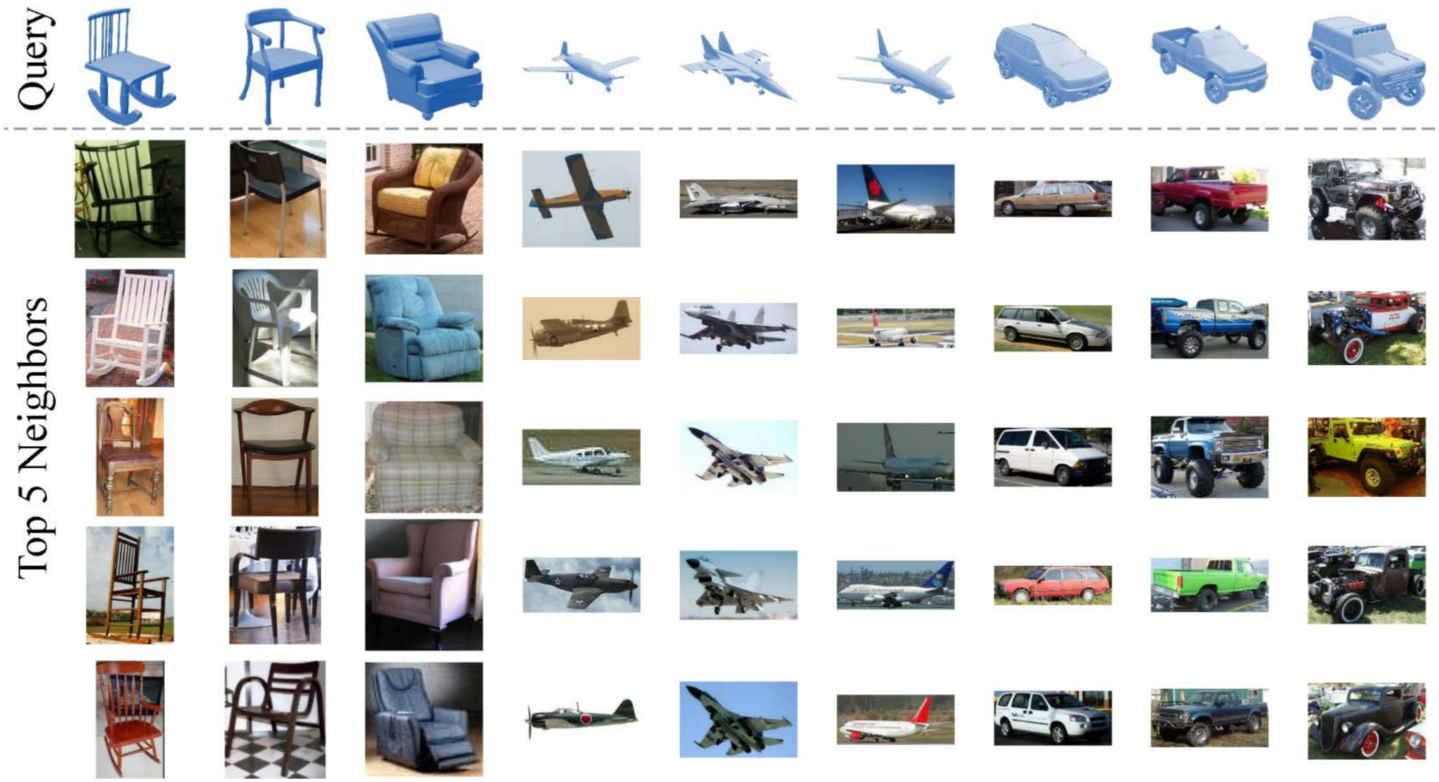


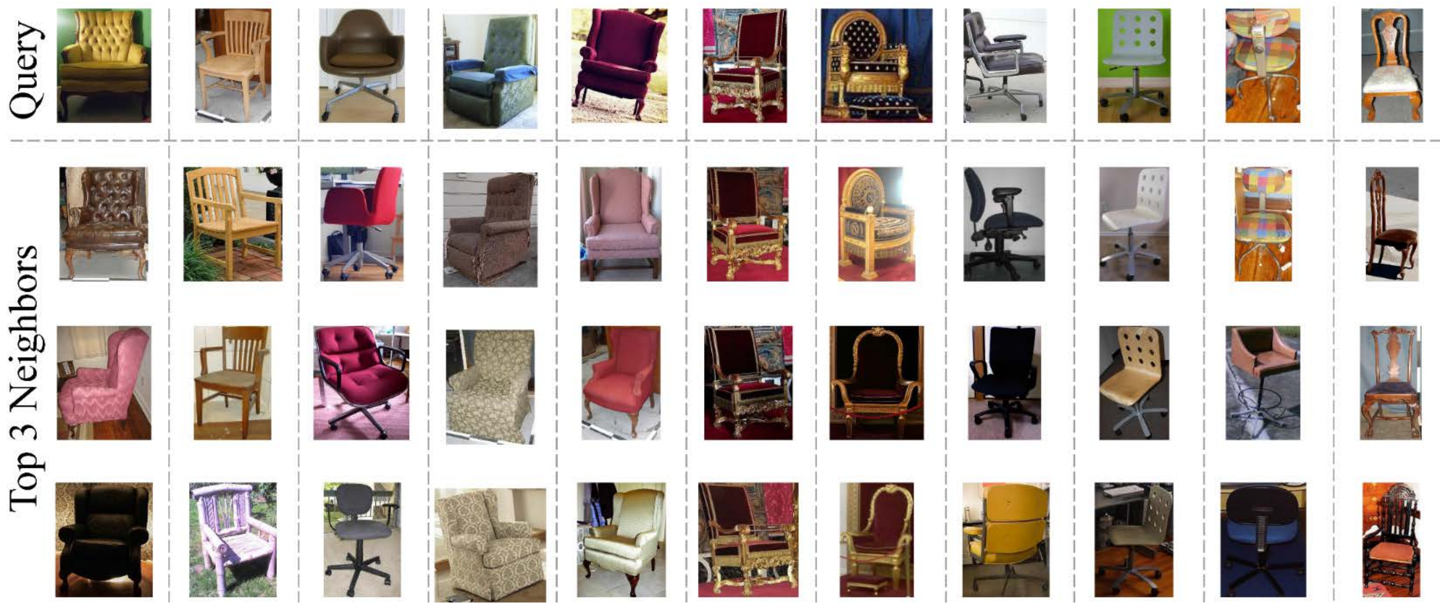
Image based Shape Retrieval



Shape based Image Retrieval

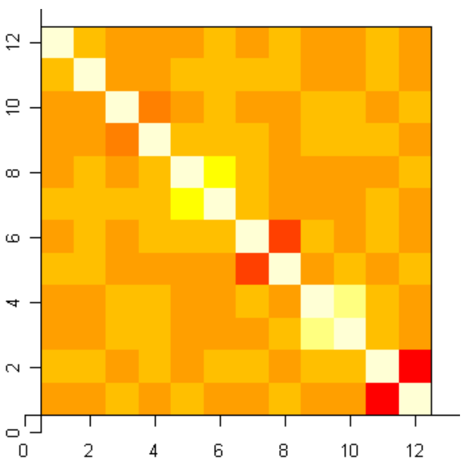


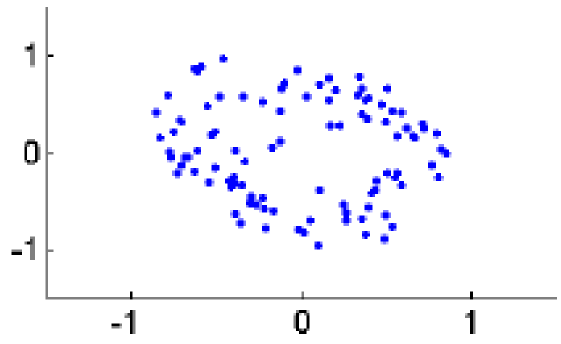
Cross-View Image Retrieval



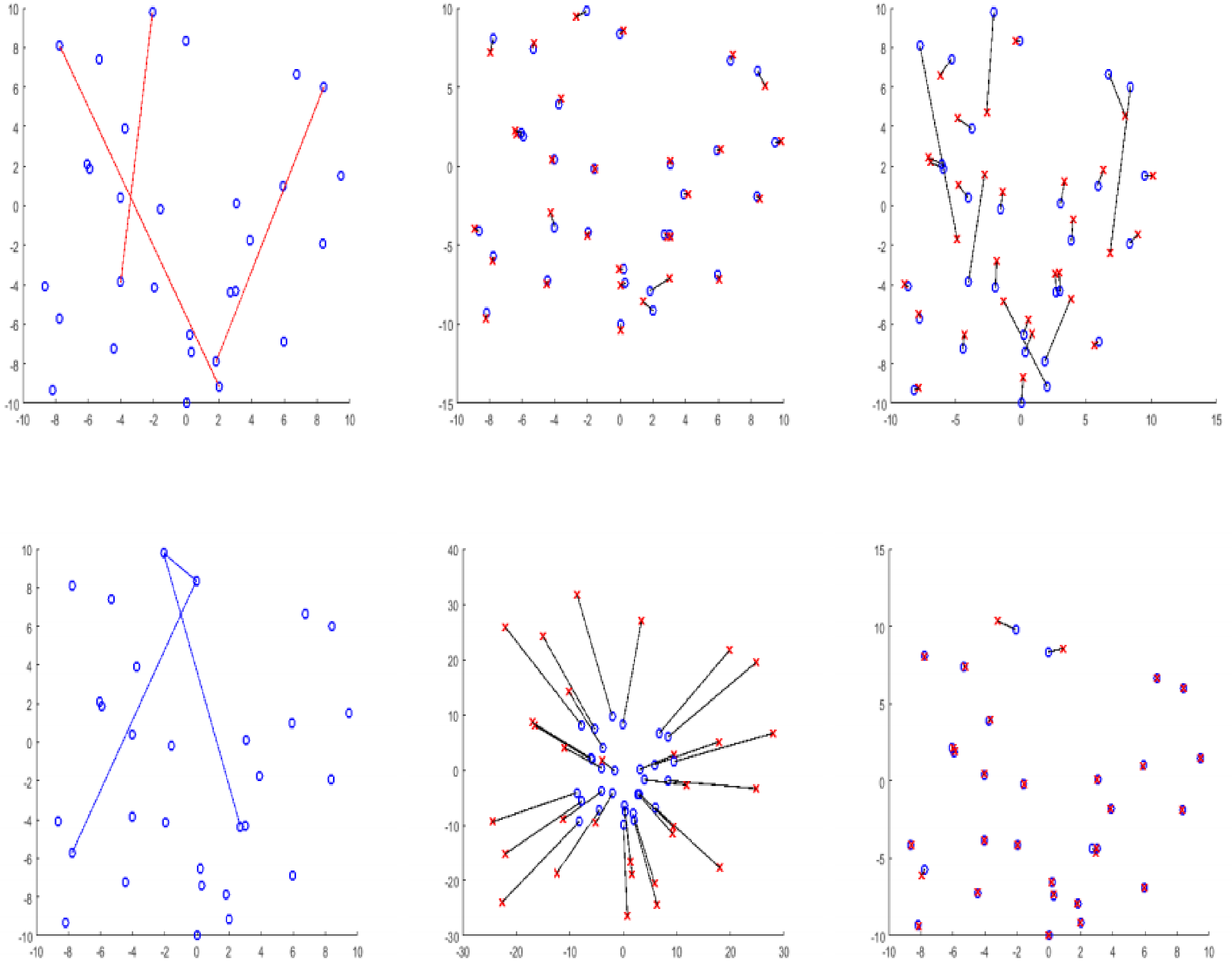
MDS Embedding

$$Stress_D(x_1, x_2, \ldots x_N) = \sum_{i \neq j} (D_{ij} - ||x_i - x_j||)^2$$



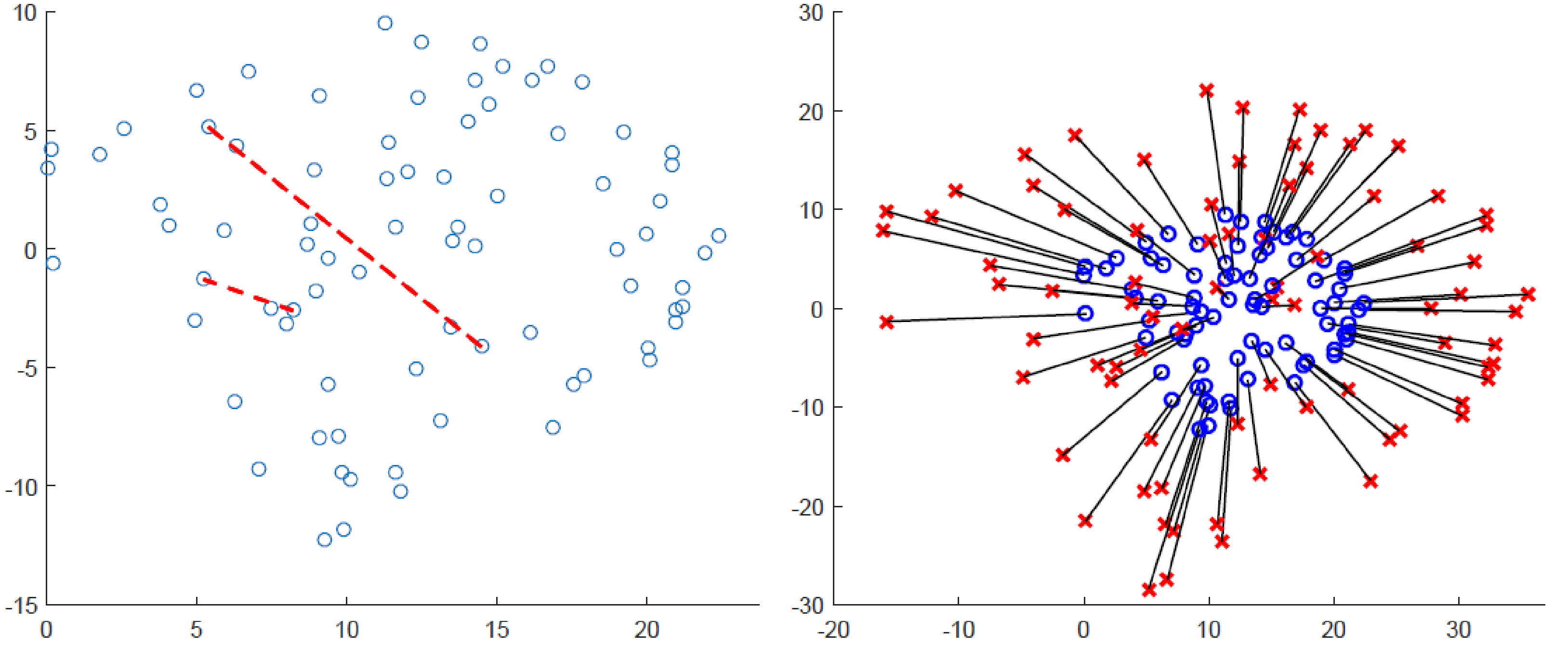


Common MDS do not handle outliers



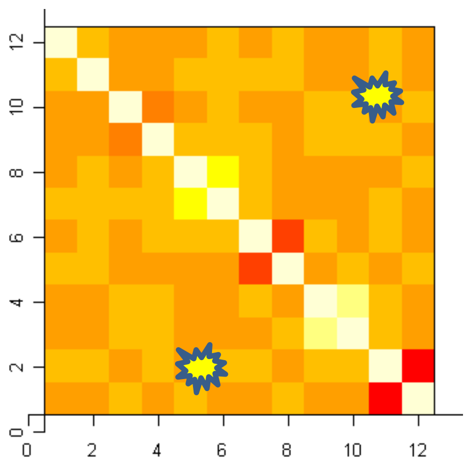
input **SMACOF** **Sammon**

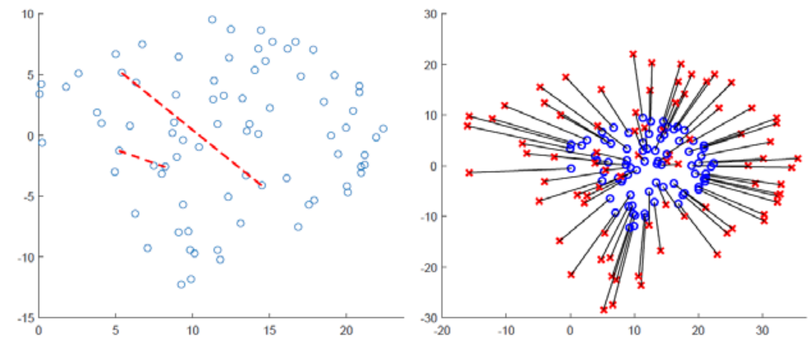
Two outlier distances lead to significant distortion in the embedding



In many real-world scenarios, input distances may be noisy or contain outliers, due to malicious acts, system faults, or erroneous measures.

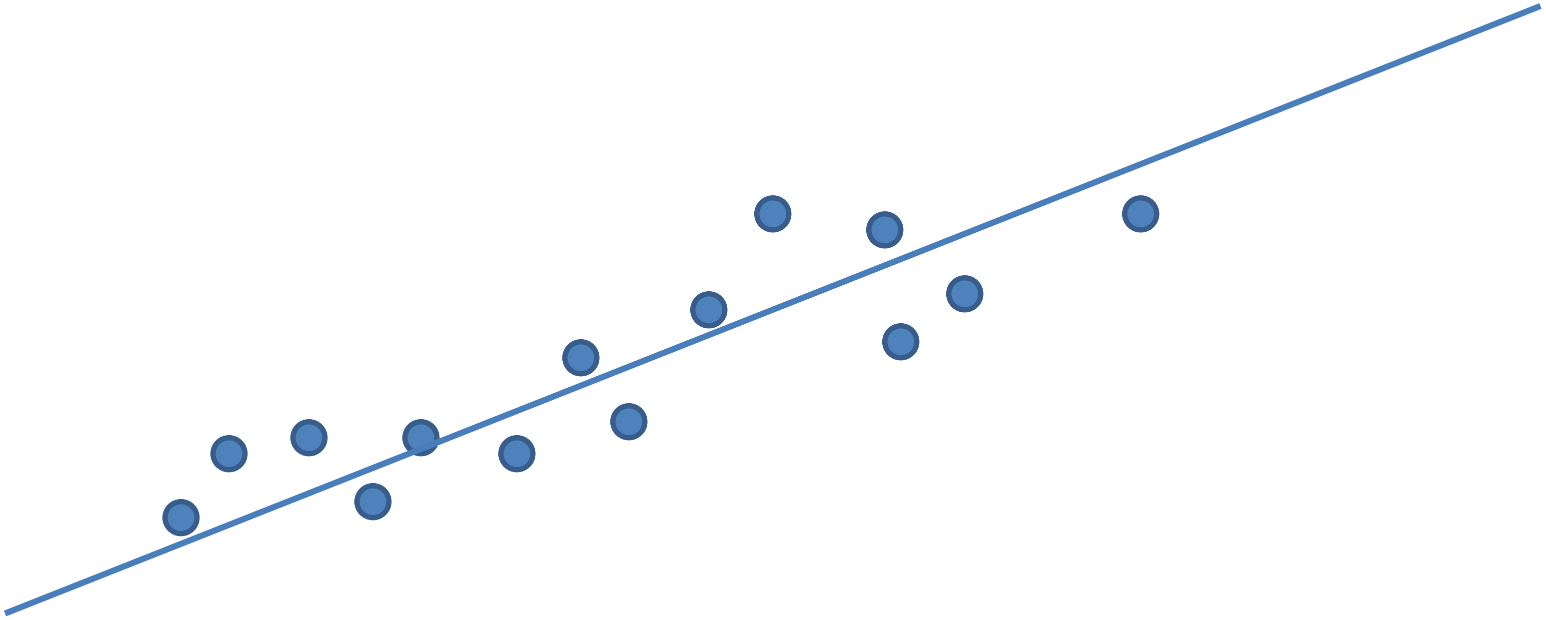
Two outlier distances lead to significant distortion in the embedding



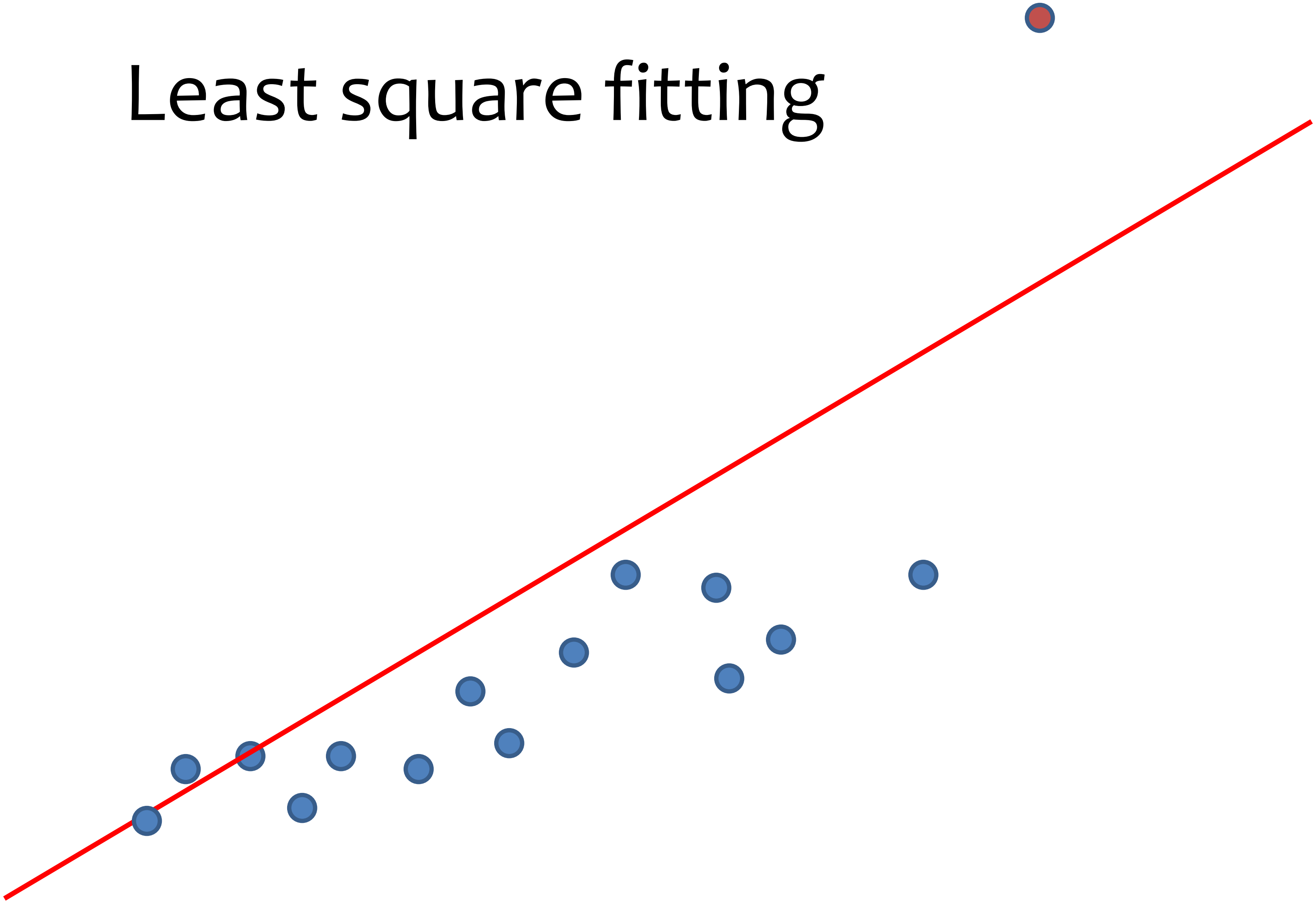


In many real-world scenarios, input distances may be noisy or contain outliers, due to malicious acts, system faults, or erroneous measures.

Least square fitting



Least square fitting



RANSAC

- Generate Lines using Pairs of Points.
- Count number of points within ε of line.
- Pick the best line.

Video

RANSAC

Sadly can’t be applied to MDS – a lot of data is needed for generating an embedding.
Almost every sample will still have outliers.

Forero and Giannakis method

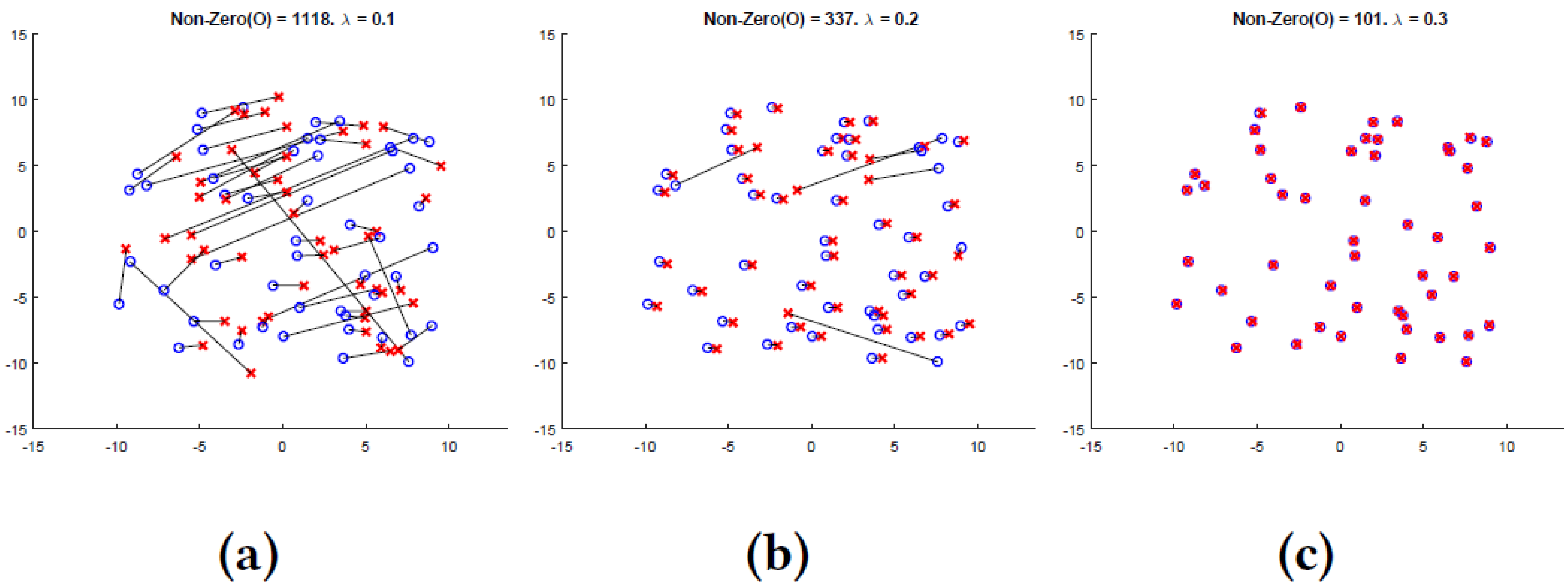
$$\sum_{i<j}(D_{ij} - ||x_i - x_j|| - O_{ij})^2 + \lambda \sum_{i<j} \mathbb{1}(O_{ij} \neq 0)$$

O_{ij} The non-zero entries represent the outlier pairs

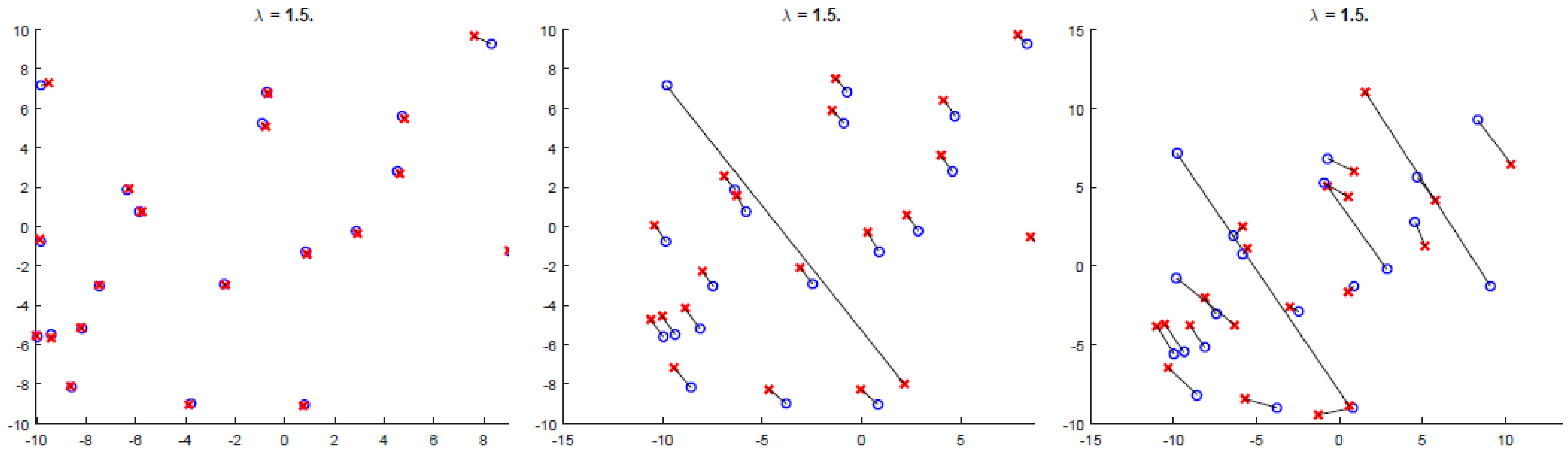
λ Lasso regression parameter (when bigger there are less outliers)

- Tuning the regularization parameter λ is not a simple task.
- There are NxN unknowns instead of just dxN, thus it is significantly harder to solve accuratly and thus very sensitive to the initial guess.

Different λ applied to the same dataset with the same initial guess, leads to different embedding qualities.



Same λ applied to the same datasets with different initial guesses, yields different embedding qualities.

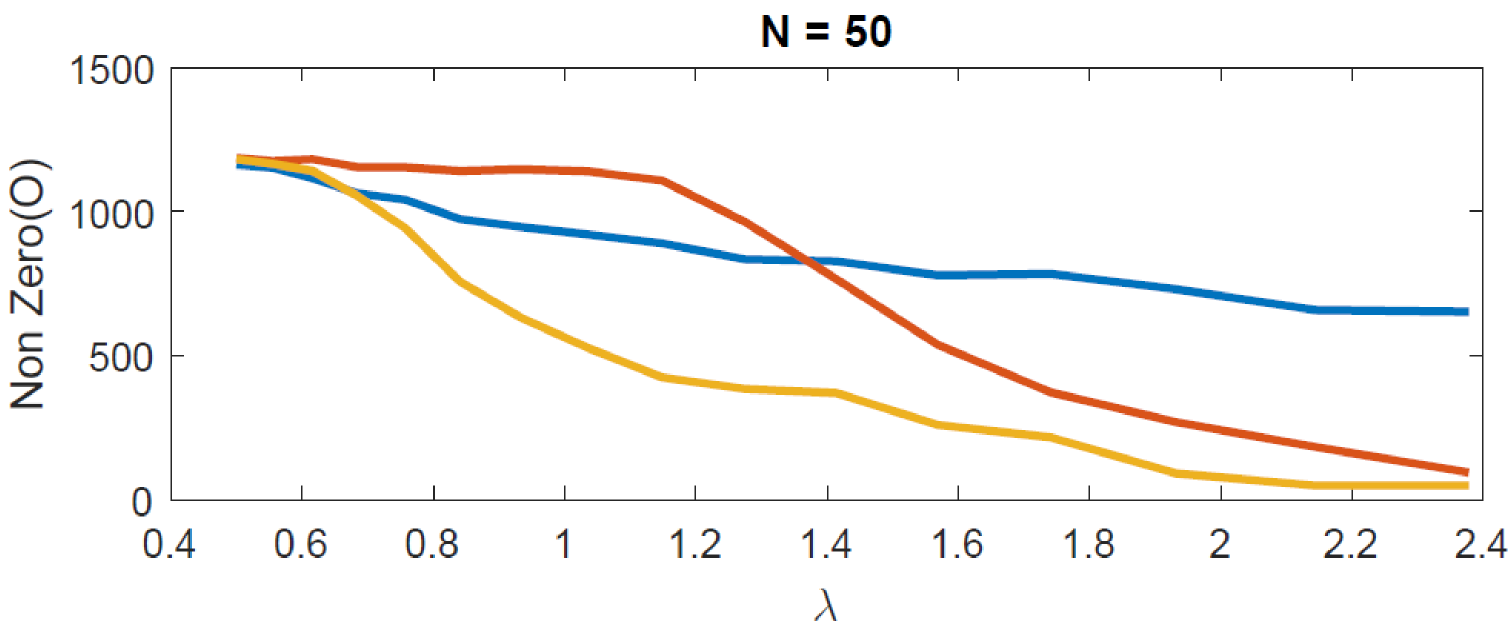


(d) (e) (f)

FG12 method is overly sensitive to the initial guess.

This graph presents the number of non-zero elements in O (which represent outliers) as a function of λ.

The three plots were generated using different initial guesses that were uniformly sampled.



Embed and remove pairs which are overly stressed…

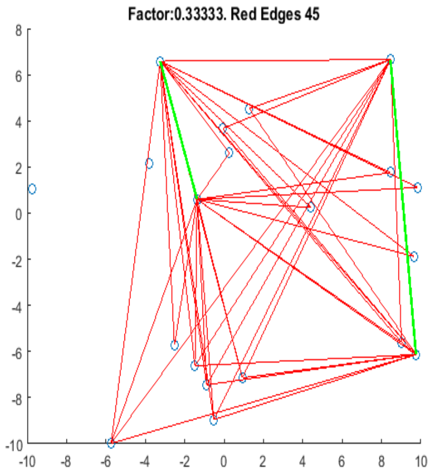
$$\sum_{i \neq j} (D_{ij} - ||x_i - x_j||)^2$$

Sadly, the overly stressed edges are not necessarily outliers.
(for example long edge that became a short one can cause a lot of short edges to deform in the embedding).

Also other stress weighting has their shortcomings – we tested that method for a while.

Geometric Reasoning

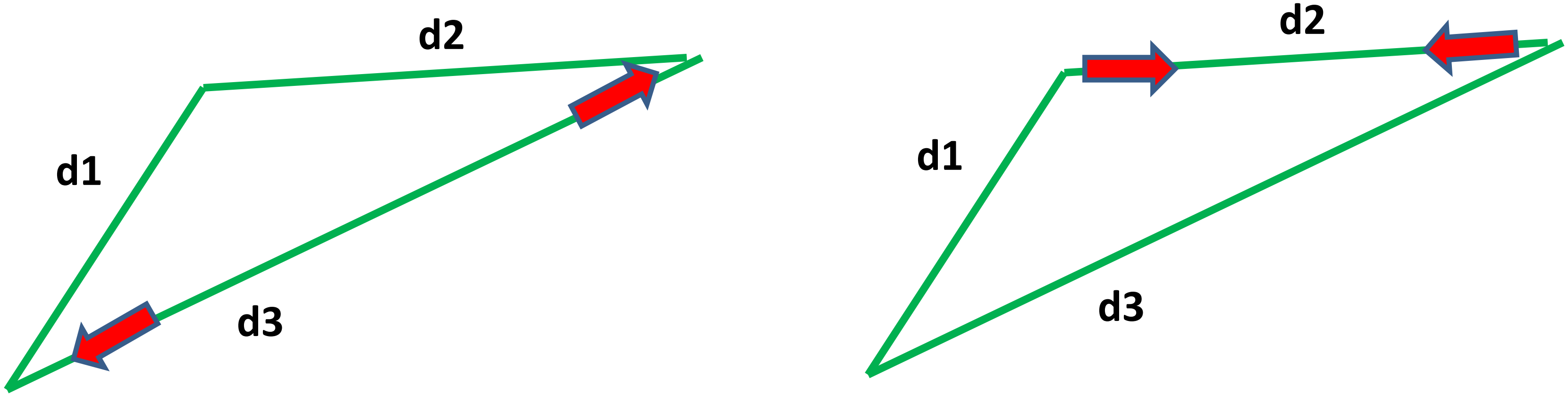
An outlier distance tends to break many triangles.
We detect those outliers and filter them.



Broken Triangles

For triangle with edge length $d1 \leq d_2 \leq d3$

If $d_1 + d_2 < d_3$ then the triangle is broken

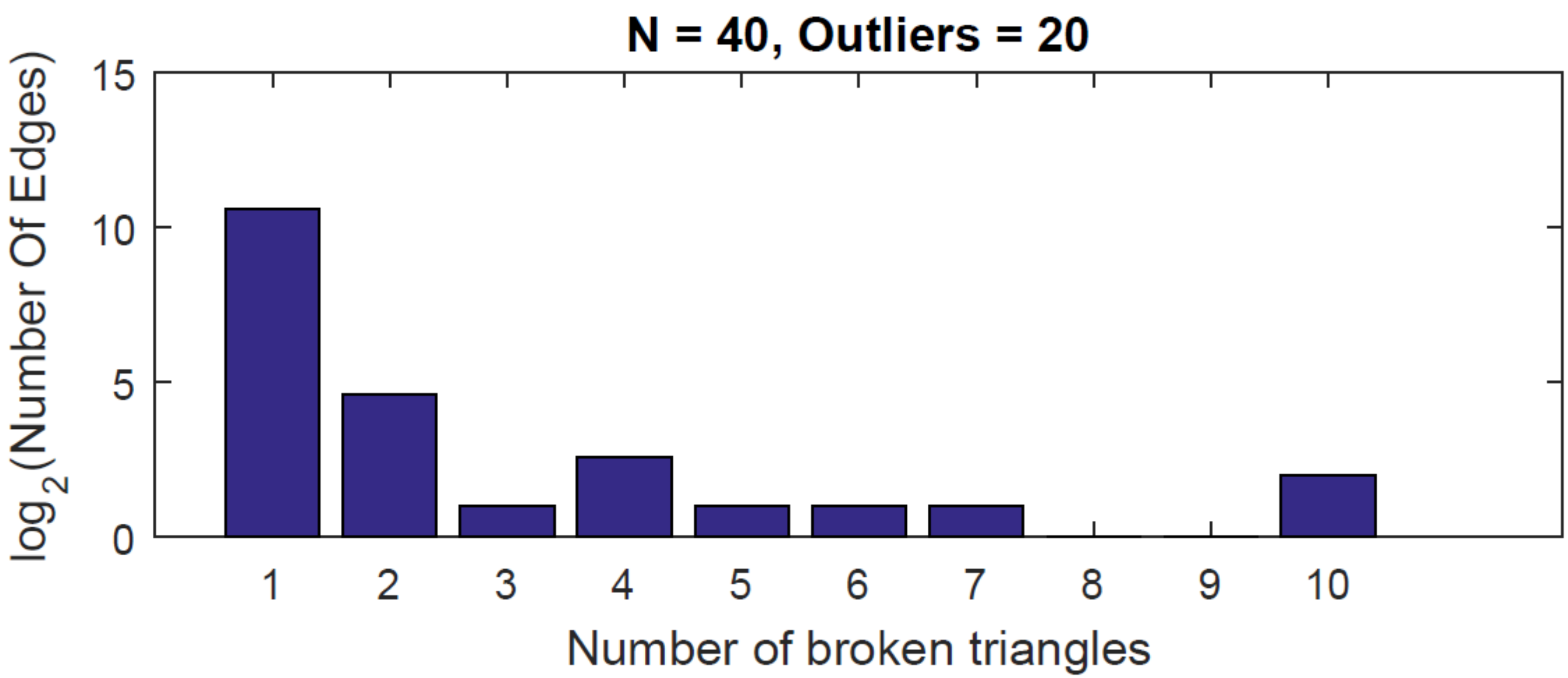


Broken Triangles

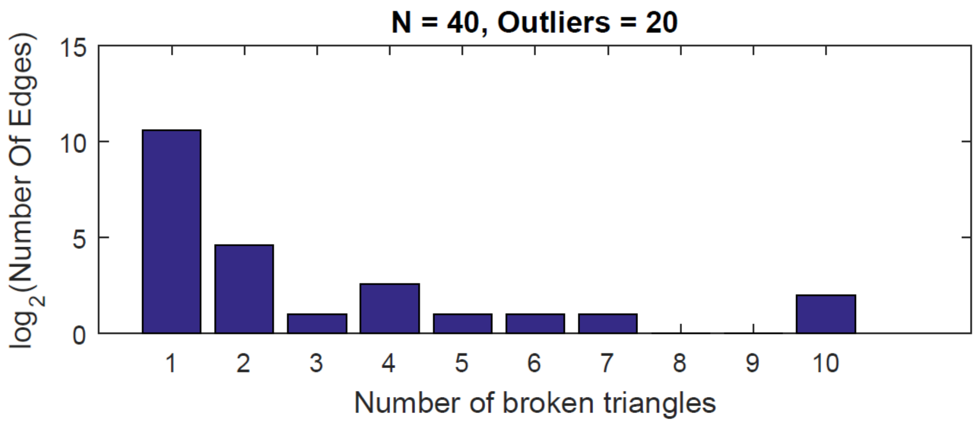
An edge in a broken triangle is not necessarily an outlier

Not every outlier edge necessarily breaks a triangle

Histogram of Broken Triangles



Histogram of Broken Triangles



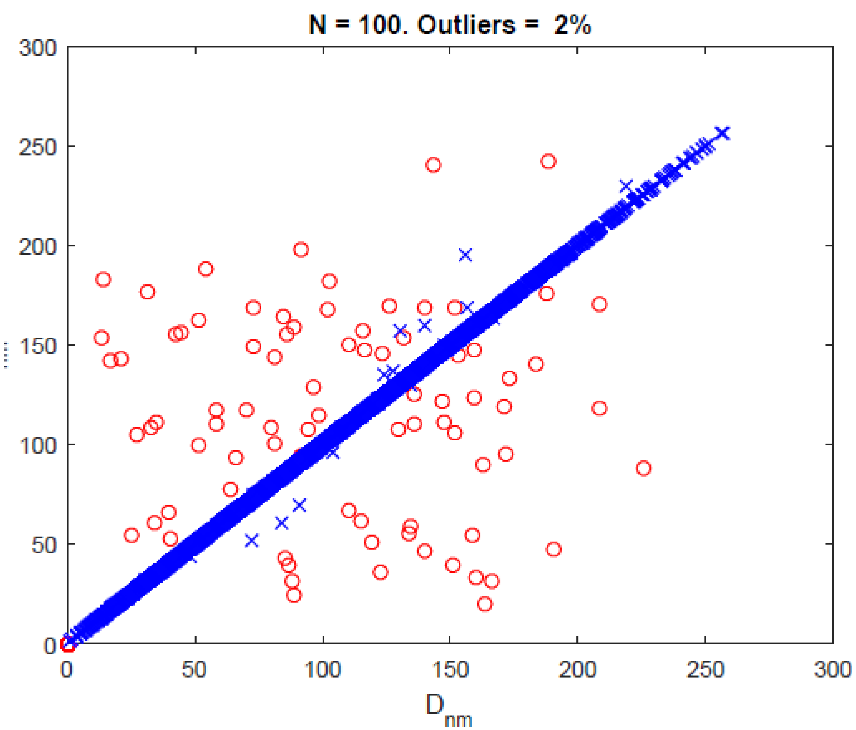
We set ϕ to be the smallest value that satisfies the following two requirements:

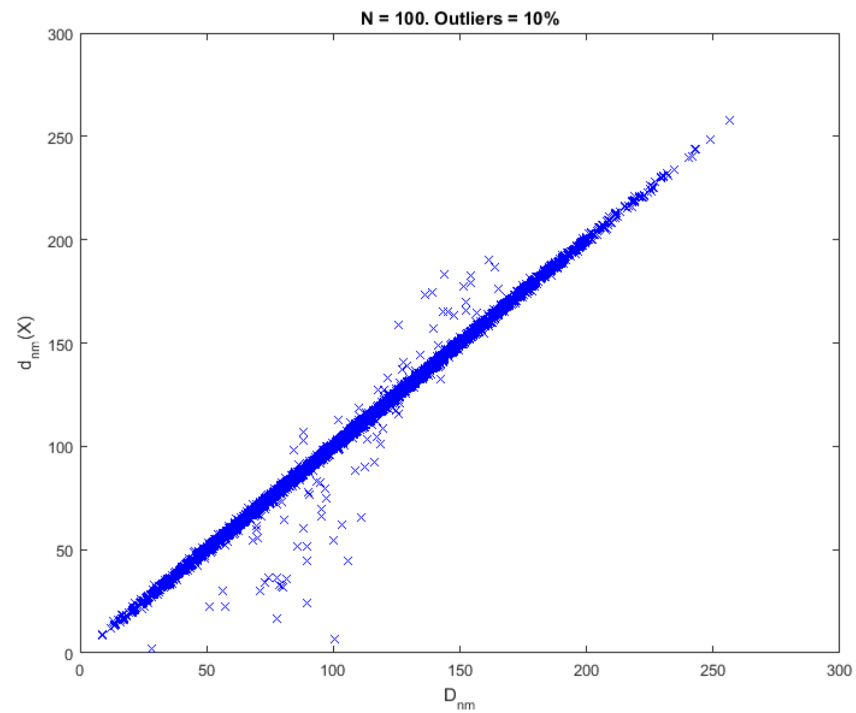
$$\sum_{b=1}^{\phi} H(b) \geq |E|/2$$

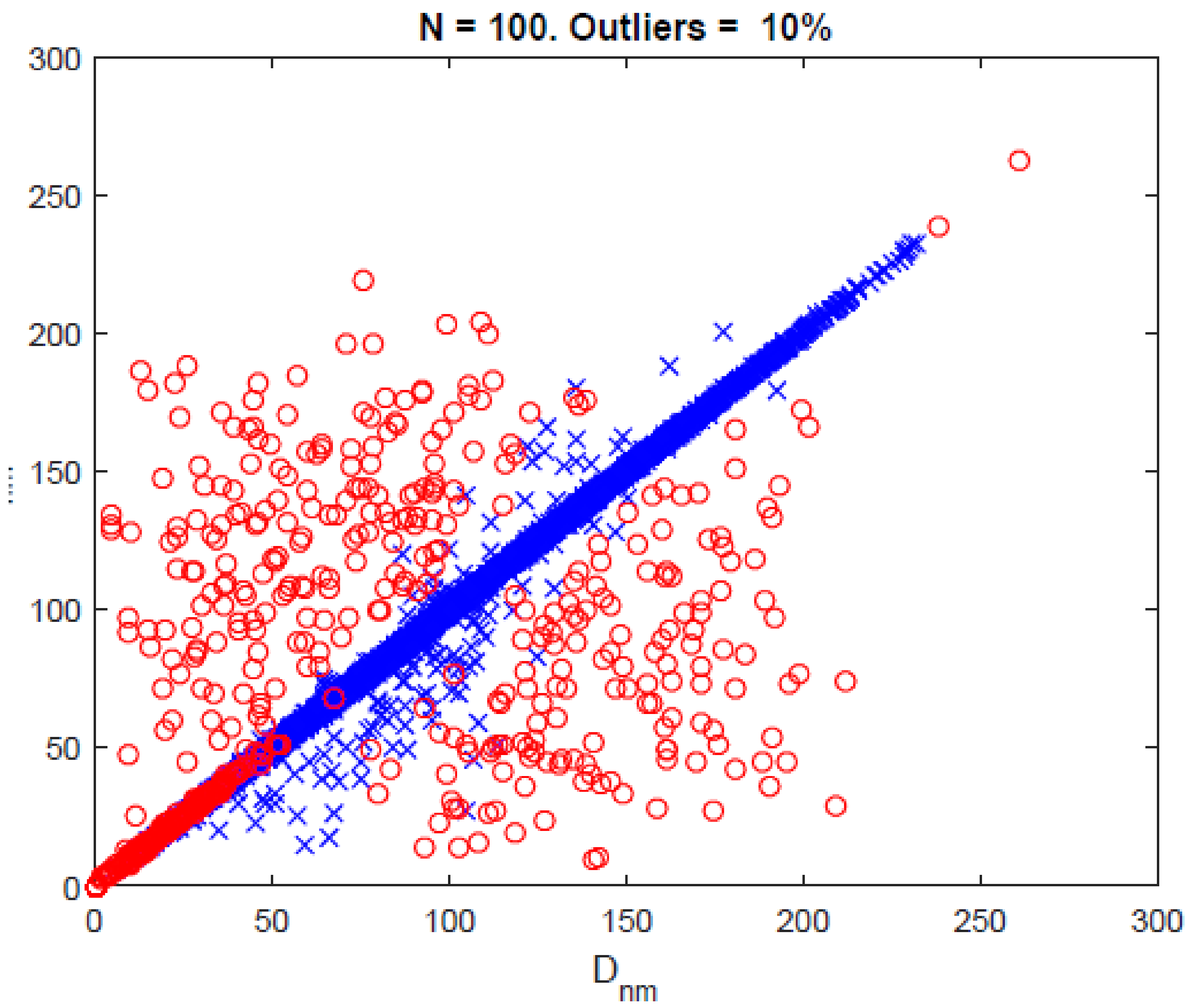
$$H(\phi + 1) > H(\phi)$$

Shepard Diagram

Each point represents a distance. The X-axis represents the input distances and the Y-axis represents the distance in the embedding result.

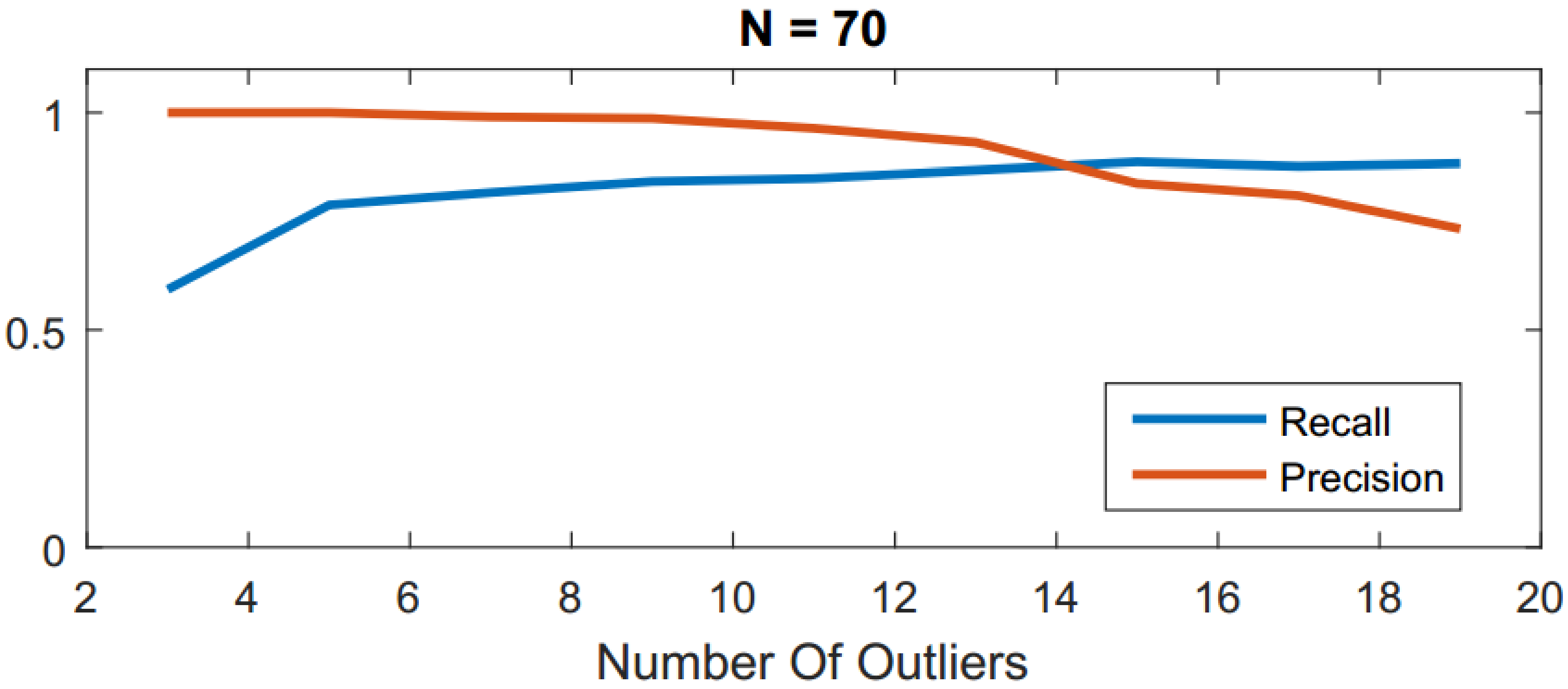






The Red dots are the distances classified as outliers. Some of the are on diagonal – those are the false positives.

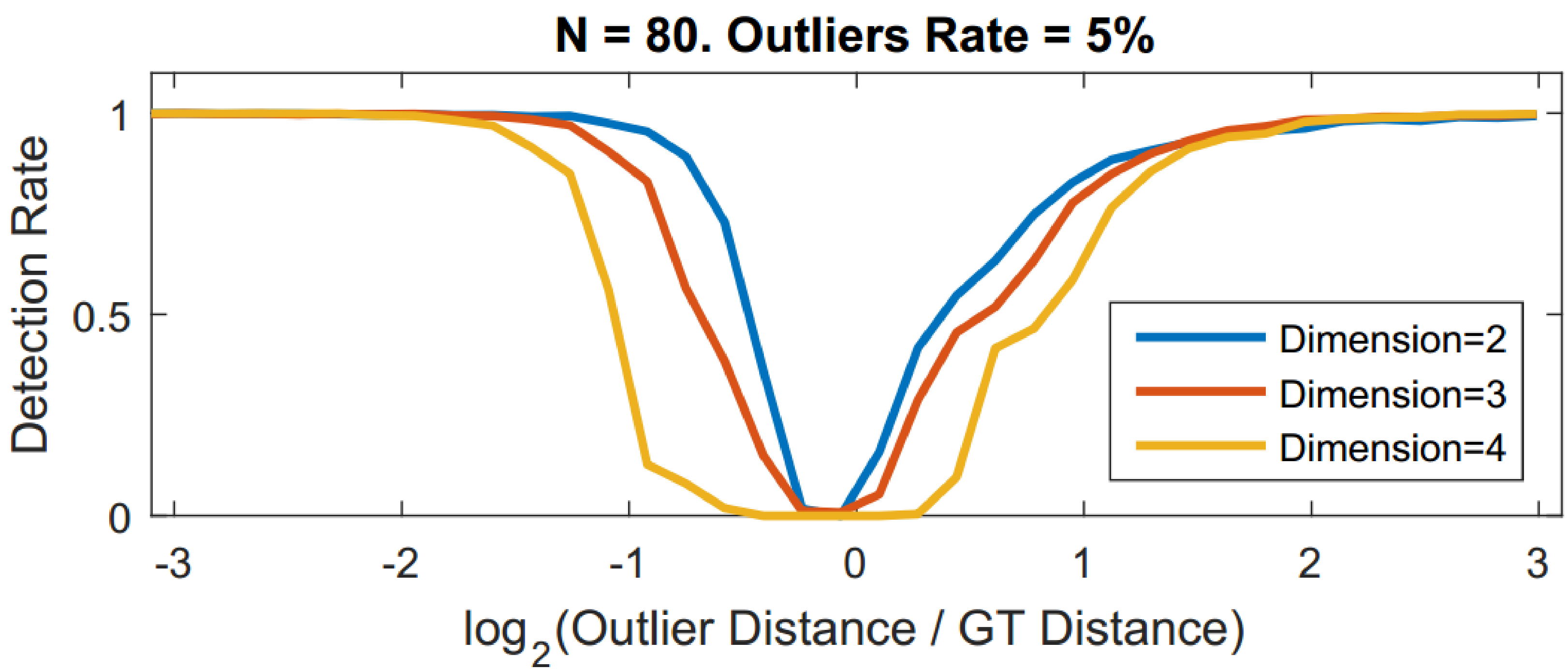
Precision and Recall



Threshold Performance

The outlier detection rate as a function of
the shrinkage enlargement of the outliers relative to the
ground-truth value. Edges that are strongly deformed (either squeezed or enlarged) are likely to be detected.

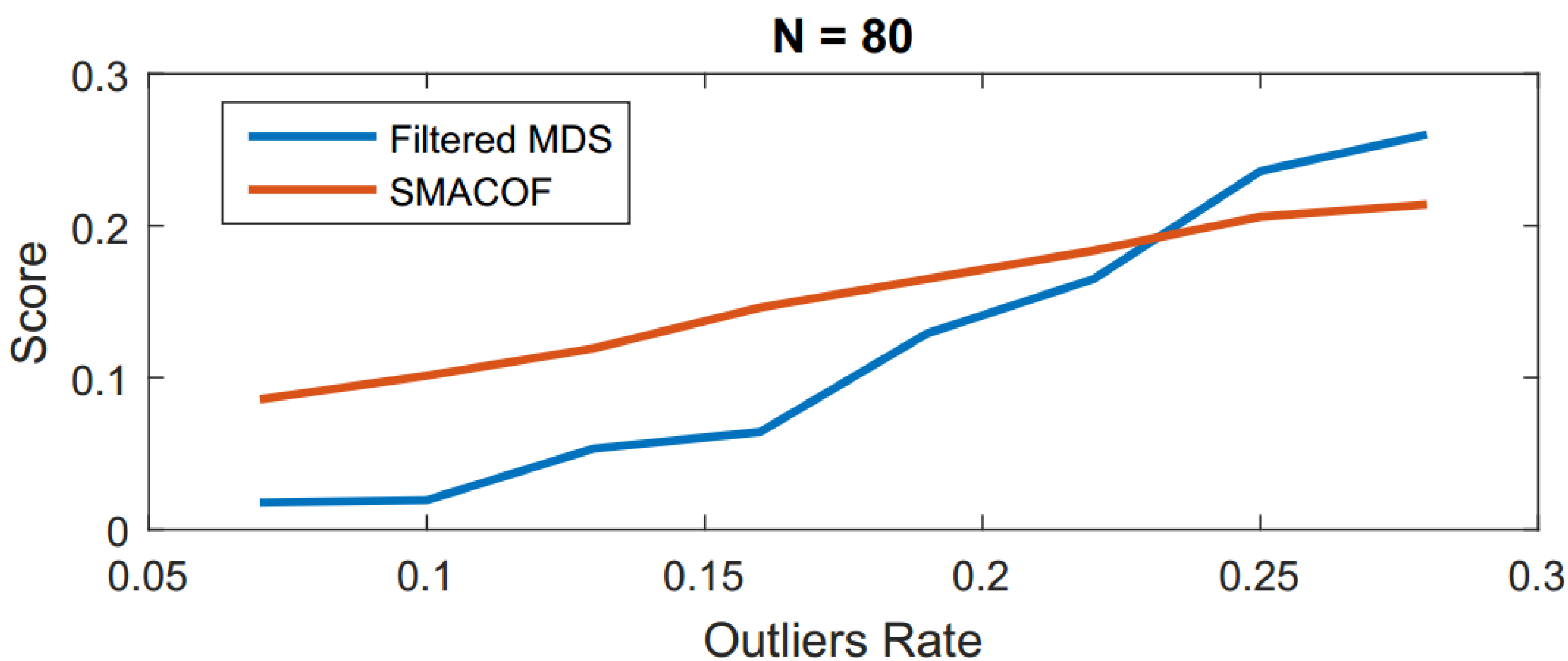
Note: the X-axis is logarithmic: *log*2(*Dout* /*DGT*).

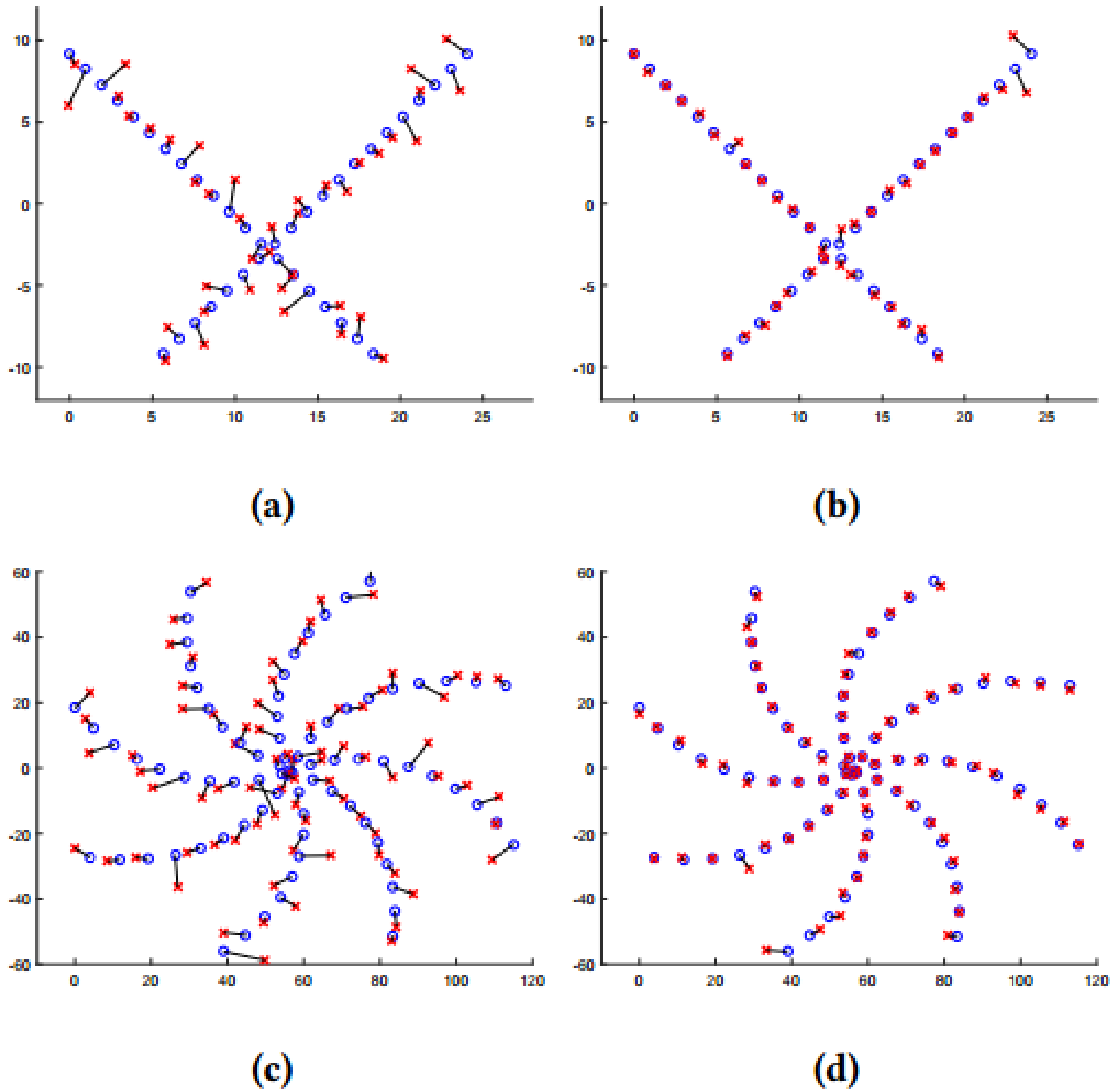


Qualitative Comparison

A comparison between SMACOF and our method as a function of outlier rate.
Up to 22% our method has better performance.

$$Score = \sum_{i \neq j} S_{ij}, \qquad S_{ij} = \left| \log \frac{\| X_i - X_j \|}{D_{ij}} \right|$$



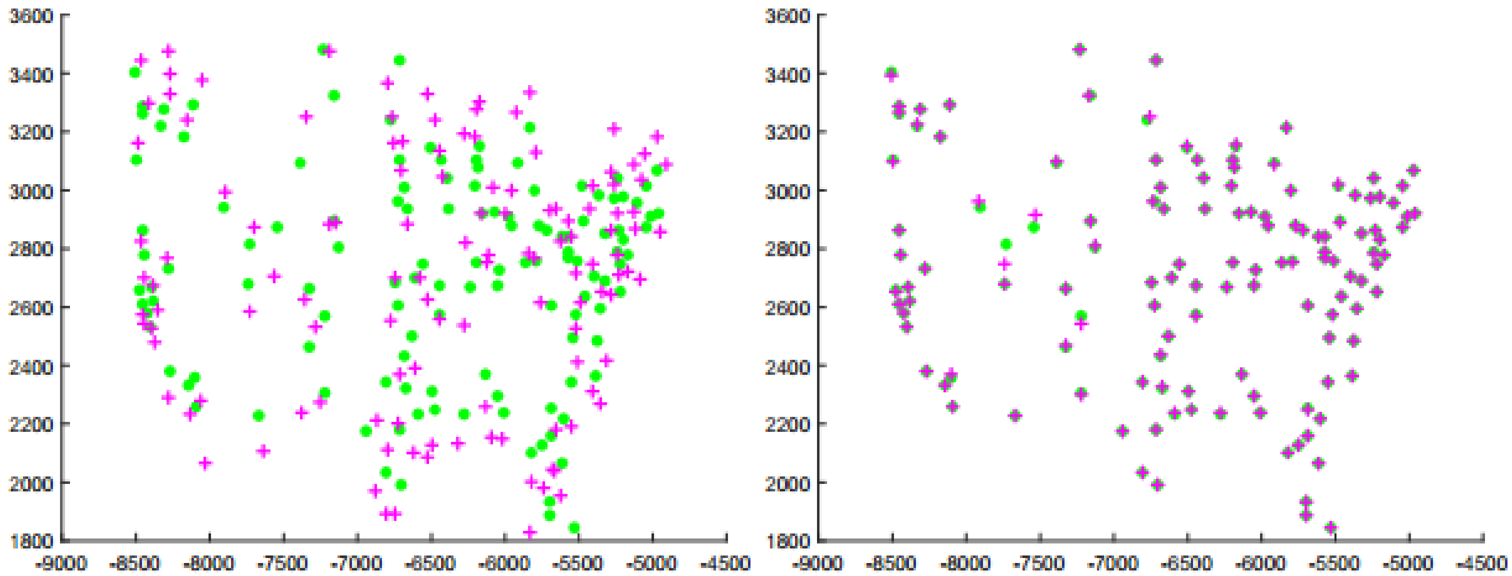


The embedding of a ’PLUS’ shaped dataset with 10% outliers, and a ’SPIRAL’ shaped dataset with 15% outliers.

(a,c) SMACOF (b,d) Our technique.

128 US Cities

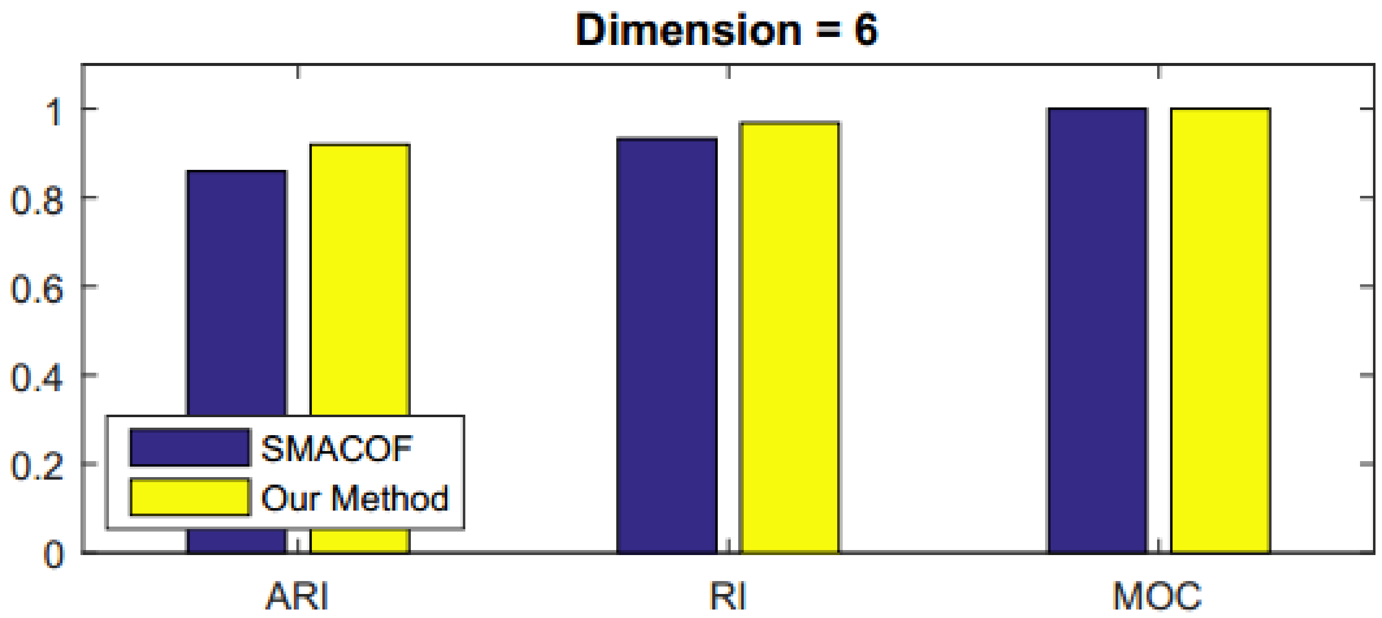
Two-dimensional embedding of SGB128 distances with 10% outliers.
The green dots are the ground-truth locations and the magenta dots represent the embedded points.
(a) SMACOF (b) Our Filtering technique.



(a) (b)

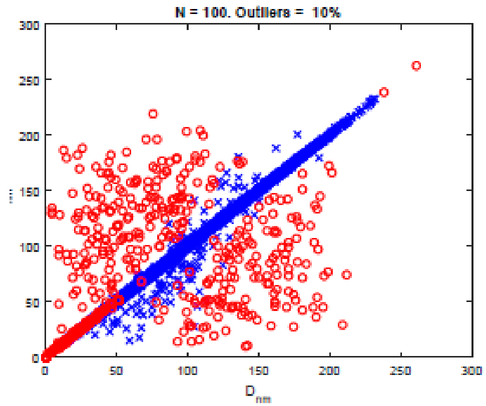
Protein Dataset

Average cluster index value of 10 executions.
The embedding dimension is set to 6, since for lower dimensions SMACOF fails due to co-located points.



Outlier Detection for Robust Multi-dimensional Scaling

Thank You



Outlier Detection for Robust Multi-dimensional Scaling

Leonid Blouvshtein and Daniel Cohen-Or