Variance-based Pivot Selection for Metric Spaces

Yan Ruan [1], Detao Ji [2], Kun Luo [1], Yuhang Lou [1], Minhua Lu [1], Rui Mao [1]

[1]College of Computer Science and Software Engineering, Shenzhen University, China

[2]School of Computer Science & Technology, Beijing Institute of Technology, Zhuhai, China

Background

- In metric-space indexing, data can be represented by distances to reference points (pivots).
- The intrinsic dimension of the data is often estimated first and used as the number of pivots to select, after which the pivots are chosen according to certain heuristics.

The Problem

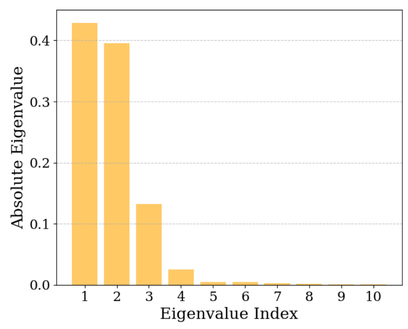
- Existing dimension estimation methods are usually inaccurate or unstable.
- Current pivot selection heuristics may produce redundant pivots or show inconsistent performance across datasets.

Our Contributions

- ***Variance-based framework***: Designed for intrinsic dimension estimation, combining eigenvalue truncation, matrix centering, and a novel elbow rule (S-Elbow).
- ***MVGSO***: A pivot selection method that incrementally chooses representative, low-redundancy pivots via Gram–Schmidt orthogonalization.

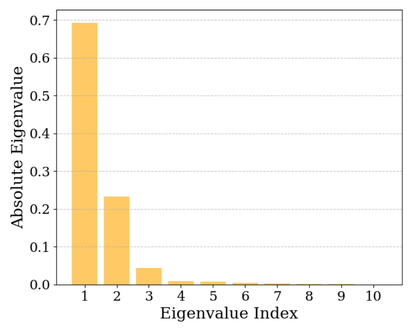
Dimension Estimation Methods

PCA-based Method



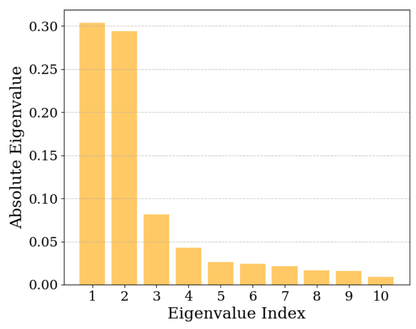
Original 2D data shows 3 components after PCA on distance matrix.

Truncation



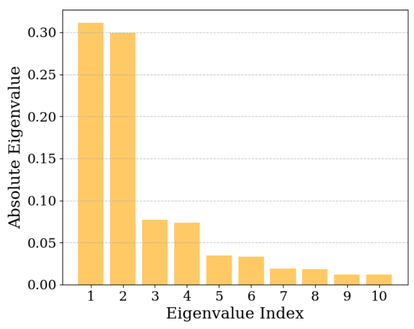
Removing the first eigenvalue.

Double-Centering



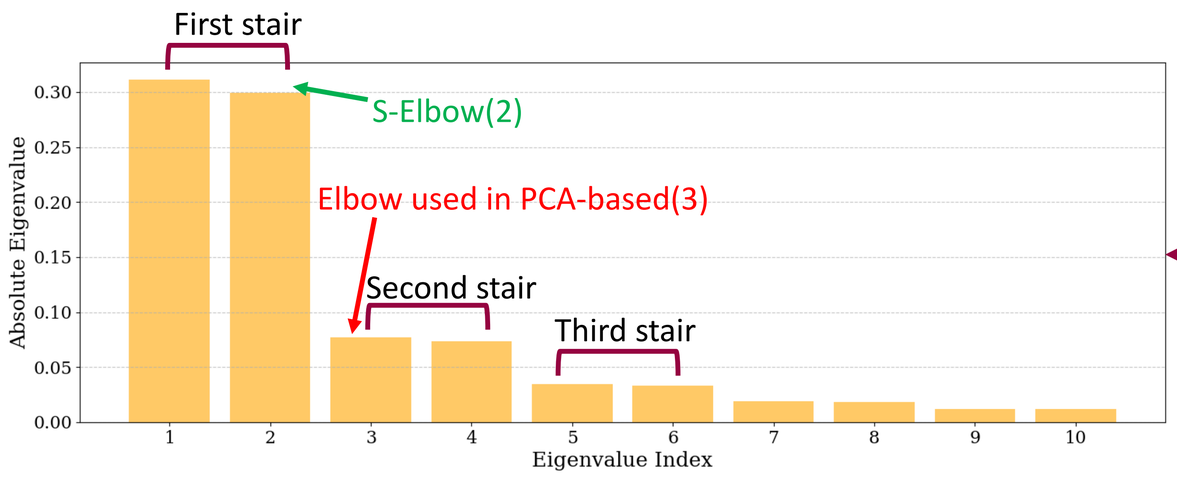
Applying the double-centering step of cMDS yields two clear components.

Column-Centering



Finds the dimension more simply.

Column-centering reveals clear stair structure and the first stair corresponds to the intrinsic dimension.

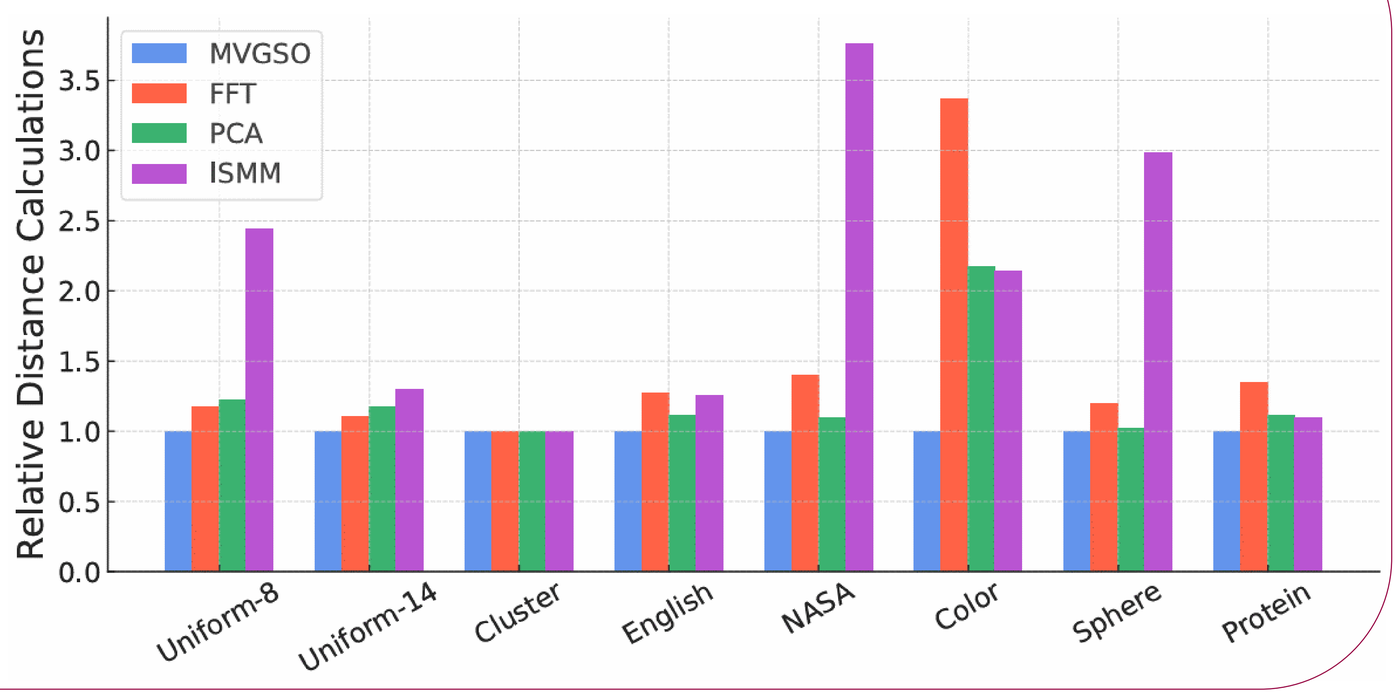


We propose S-Elbow to detect stair boundary for dimension estimation.

Pivot Selection Methods

We compare our proposed ***MVGSO*** with common pivot selection heuristics:

- ***FFT***: Simple but but sensitive to initialization and unstable in pruning.
- ***PCA***: Projects onto principal directions, but higher-order components are often redundant with limited query-time gains.
- ***ISMM***: Maximizes pairwise distance; effective but somewhat dataset-dependent.
- ***MVGSO***: Incrementally selects pivots with maximum variance, applying Gram–Schmidt to reduce redundancy, , and achieves **up to 70% fewer distance computations** with **~10% average** improvement over the best baselines



Supported by the NSFC project “Theory and methods of graph universal representation through metric space”.