UPC
UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Facultat d'Informàtica de Barcelona

FIB

DISTANCE-BASED DIMENSIONALITY REDUCTION FOR BIG DATA

ADRIÀ CASANOVA LLOVERAS

Thesis supervisor
PEDRO DELICADO USEROS (Department of Statistics and Operations Research)

Thesis co-supervisor
CRISTIAN PACHÓN GARCIA (Department of Statistics and Operations Research)

Degree
Master's Degree in Data Science

Master's thesis

Facultat d'Informàtica de Barcelona (FIB)

Universitat Politècnica de Catalunya (UPC) - BarcelonaTech

Abstract

Dimensionality reduction aims to project a data set into a low-dimensional space. Many techniques have been proposed, most of them based on the inter-individual distance matrix. When the number of individuals is really large, the use of distance matrices is prohibitive. There are algorithms that extend MDS (a classical dimensionality reduction method based on distances) to the big data setting. In this TFM, we adapt these algorithms to any generic distance-based dimensionality reduction method.

Contents

1 Introduction, Motivation, and Objectives

1.1 Introduction

- Dimensionality Reduction definition, goal and applications.
- Examples of DR methods.
- Key points and limitations of DR methods.

1.2 Motivation

- When the number of individuals is really large, the use of distance matrices is prohibitive.
- There are algorithms that extend MDS to the big data setting.

1.3 Objectives

- Adapt these algorithms to any generic distance-based dimensionality reduction method.

2 State of the Art

2.1 Introduction

- There are many DR algorithms, linear and non-linear, but they use the distance matrix of datapoints. In big datasets, this matrix cannot fit in the system's RAM, so DR methods are not feasible. Moreover, time complexity can be prohibitive in some cases as well.

- Delicado and Pachón-García proposed new versions of MDS that handled this problem and compared them with prior algorithms [5].

- Regarding non-linear methods, Landmark Isomap [9] was proposed to adapt Isomap to large data settings. Later, in 2024, Reichmann, Hägele and Weiskopf generalized Interpolation MDS to any DR method that would return a map between high- and low-dimensional spaces, such as PCA, non-classicla MDS, t-SNE, UMAP or Autoencoders.

- Finally, t-SNE has been optimized for big data environments in Python through iterative implementations. Their details, however, are out of the scope of this project.

2.2 A few Dimensionality Reduction Techniques

2.2.1 Non-classical MDS. The SMACOF algorithm

SMACOF (Scaling by MAjorizing a COmplicated Function) is a multidimensional scaling algorithm that minimizes metric stress using a majorization technique [3]. Also known as the Guttman Transform, this technique is more powerful for this problem than general optimization methods, such as gradient descent.

Algorithm 1 SMACOF

```
Require: $D_{\mathcal{X}} = (\delta_{ij})$, the matrix of observed distances; $q$, the embedding's dimensionality; $n\_iter$, the maximum number of iterations; and $\epsilon$, the convergence threshold.
Ensure: $\tilde{\mathcal{Y}}$, a configuration in a $q$-dimensional space.
Initialize $\tilde{\mathcal{Y}}^{(0)} \in \mathbb{R}^{n \times q}$
$k \leftarrow 0$
repeat
    Compute distance matrix of $\tilde{\mathcal{Y}}^{(k)}$: $d_{ij}(\tilde{\mathcal{Y}}^{(k)}) = \|x_i - x_j\|$
    Compute the Metric STRESS: $STRESS_M(D_{\mathcal{X}}, \tilde{\mathcal{Y}}^{(k)}) = \sqrt{\frac{\sum_{i<j}(\delta_{ij} - d_{ij})^2}{\sum_{i<j}\delta_{ij}^2}}$
    Compute the Guttman Transform: $\tilde{\mathcal{Y}}^{(k+1)} = n^{-1}B(\tilde{\mathcal{Y}}^{(k)})\tilde{\mathcal{Y}}^{(k)}$ where $B(\tilde{\mathcal{Y}}^{(k)}) = (b_{ij})$:
        $$b_{ij} = \begin{cases} -\delta_{ij}/d_{ij}(\tilde{\mathcal{Y}}^{(k)}) & \text{if } i \neq j \text{ and } d_{ij}(\tilde{\mathcal{Y}}^{(k)}) > 0 \\ 0 & \text{if } i \neq j \text{ and } d_{ij}(\tilde{\mathcal{Y}}^{(k)}) = 0 \\ -\sum_{j \neq i} b_{ij} & \text{if } i = j \end{cases}$$
    $k \leftarrow k + 1$
until $k \geq n\_iter$ or $|STRESS_M(D_{\mathcal{X}}, \tilde{\mathcal{Y}}^{(k-1)}) - STRESS_M(D_{\mathcal{X}}, \tilde{\mathcal{Y}}^{(k)})| < \epsilon$
return $\tilde{\mathcal{Y}}^{(k)}$
```

2.2.2 Local MDS

Local MDS [4] is a variant of non-classical multidimensional scaling that differs in how large distances are treated. Specifically, a repulsive term between distant points is added to the stress function to further separate points in the low-dimensional configuration.

Algorithm 2 Local MDS

```
Require: $D_{\mathcal{X}}$, the matrix of observed distances; $q$, the embedding's dimensionality; $k$,
    the size of neighborhoods; and $\tau$, the weight of the repulsive term.
Ensure: $\tilde{\mathcal{Y}}$, a configuration in a $q$-dimensional space.
Compute the symmetrized k-NN graph of $D_{\mathcal{X}}$, $\mathcal{N}$
Calculate $t = \frac{|\mathcal{N}|}{|\mathcal{N}^C|} \cdot \text{median}_{\mathcal{N}} \left(D_{i,j}\right) \cdot \tau$
Minimize
    $$\sum_{(i,j)\in N} \left(D_{i,j} - \|\mathbf{y}_i - \mathbf{y}_j\|\right)^2 - t \sum_{(i,j)\notin N} \|\mathbf{y}_i - \mathbf{y}_j\|$$
return The solution to the optimization problem, $\tilde{\mathcal{Y}}$.
```

Parameters τ - which must be in the unit interval - and k may be tuned with k'-cross validation thanks to the LCMC (Local Continuity Meta-Criteria). (*POSSIBLE ANNEX*)

2.2.3 Isomap

Isomap [10] is a nonlinear technique that preserves geodesic distances between points in a manifold. The key insight of Isomap is that large distances between objects are estimated from the shorter ones by the shortest path length. Then, shorter and estimated-larger distances have the same importance in a final MDS step.

Algorithm 3 Isomap

```
Require: $D_{\mathcal{X}}$, the matrix of observed distances; $q$, the embedding's dimensionality; and
    $\epsilon$ or $k$, the bandwith.
Ensure: $\tilde{\mathcal{Y}}$, a configuration in a $q$-dimensional space.
Find the $\epsilon$-NN or k-NN graph of $\mathcal{X}$, $G$.
Compute the distance matrix of $G$, $D_G$.
Embed $D_G$ to a $q$-dimensional space with MDS.
return The output configuration of MDS.
```

The only tuning parameter of Isomap is the bandwith (ϵ or k), but there is no consensus on what is the best method to choose it.

2.2.4 t-SNE

t-SNE (t-Distributed Stochastic Neighbor Embedding) [6] is a nonlinear dimensionality reduction technique that preserves local neighborhoods by modeling similarities between points as conditional probabilities. The difference between these probability distributions in high and low-dimensional spaces is then minimized.

t-SNE focuses on retaining the local structure of the data while ensuring that every point y_i in the low-dimensional space will have the same number of neighbors, making it particularly effective for visualizing clusters. The use of the Student t-distribution in the

```
Algorithm 4 t-SNE
Require: $\mathcal{X} \in \mathbb{R}^{n \times p}$, the high-dimensional configuration; $q$, the embedding's dimensionality; perplexity $Perp$.
Ensure: $\tilde{\mathcal{Y}}$, a configuration in a $q$-dimensional space.
For every datapoint $i$, find $\sigma_i$ so that the conditional probability ditribution
    $p_{j|i} = \frac{\exp(-\|x_i - x_j\|^2 / 2\sigma_i^2)}{\sum_{k \neq i} \exp(-\|x_i - x_k\|^2 / 2\sigma_i^2)}$
    has perplexity $2^{-\sum_j p_j \log_2 p_j} = Perp$
Symmetrize conditional distributions: $p_{ij} = \frac{p_{j|i} + p_{i|j}}{2n}$ if $i \neq j$, $p_{ii} = 0$
Consider Student t-distributed joint probabilities for the low-dimensional data $y_i$:
    $q_{ij} = \frac{(1 + \|y_i - y_j\|^2)^{-1}}{\sum_{h \neq k} (1 + \|y_h - y_k\|^2)^{-1}}$
Minimize the sum of Kullback-Leibler divergences between the joint distributions over all datapoints:
    $C(\tilde{\mathcal{Y}}) = \sum_i \sum_j p_{j|i} \log \frac{p_{j|i}}{q_{j|i}}$
return The solution to the optimization problem, $\tilde{\mathcal{Y}}$.
```

low-dimensional space addresses the *crowding problem* by allowing dissimilar points to be modeled far apart [6].

Perplexity is interpreted as the average effective number of neighbors of the high-dimensional datapoints x_i and typical values are between 5 and 50.

2.3 Multidimensional Scaling for Big Data

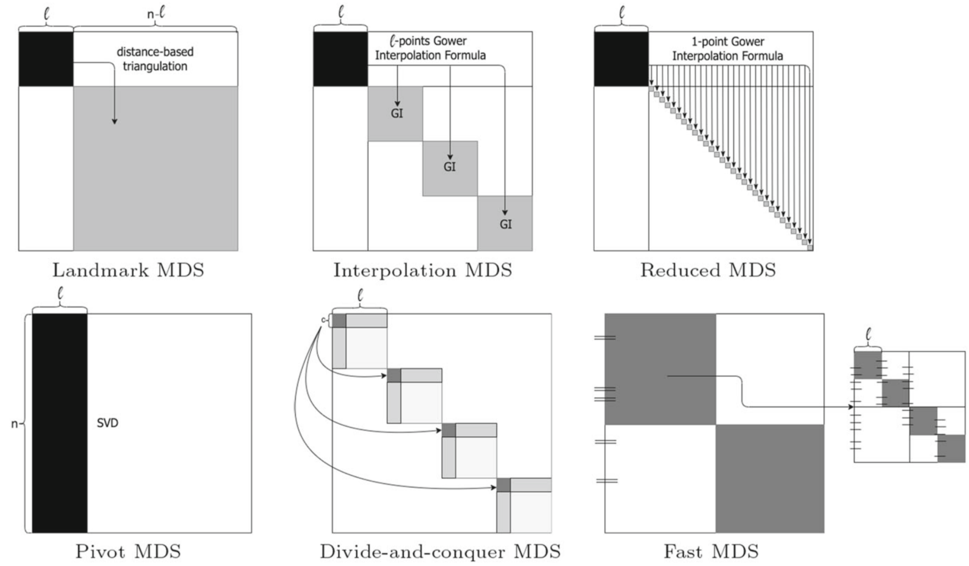
Delicado and Pachón-García [5] compared four existing versions of MDS with two newly proposed (Divide-and-conquer MDS and Interpolation MDS) to handle large data. As can be seen in figure 1, these can be grouped into four categories:

- **Interpolation-based**: Landmark MDS, Interpolation MDS and Reduced MDS apply classical multidimensional scaling to a subset of $l \ll n$ points and then interpolate the projection of the remaining data. They differ in how the interpolation is computed: Landmark MDS uses distance-based triangulation; Interpolation MDS, the l-points Gower Interpolation Formula; and Reduced MDS, the 1-point Gower Interpolation Formula.

- **Approximation-based**: Pivot MDS approximates the SVD of the full inner product matrix with the SVD of the inner product matrix between a subset of $l \ll n$ points and all the points in the dataset.

- **Divide-and-conquer**: In Divide-and-conquer MDS, the dataset is randomly partitioned into subsets of up to $l \ll n$ points into which MDS is independently applied. Then, the resulting embeddings are aligned with Procrustes transformations.

- **Recursive**: Fast MDS is similar in spirit to Divide-and-conquer MDS, but it partitions the data recursively.



Landmark MDS

Interpolation MDS

Reduced MDS

Pivot MDS

Divide-and-conquer MDS

Fast MDS

Figure 1: Schematic representation of the six MDS algorithms described by Delicado and Pachón-García [5].

2.4 Landmark Isomap and the Out-of-Core Dimensionality Reduction Framework

Silva and Tenenbaum [9] first introduced Landmark MDS in 2002 by applying it to Isomap (L-Isomap). This way, they reduced the time complexity of both classical multidimensional scaling and Isomap from $\mathcal{O}(n^3)$ to $\mathcal{O}(n^2 l)$, where $l \ll n$ is the amount of landmark points.

Note that, similarly to Landmark MDS, other big data versions of MDS can be used with Isomap. Nonetheless, interpolation-based and approximation-based algorithms cannot be trivially generalized to nonlinear dimensionality reduction methods.

Later, in 2024, Reichmann, Hägele and Weiskopf [8] proposed the Out-of-Core Dimensionality Reduction Framework. Similar to Interpolation MDS, this algorithm applies a DR method that produces a mapping between high- and low-dimensional spaces (i.e. PCA, MDS, t-SNE, UMAP, Autoencoder) to a small subset of the data and then projects the remaining datapoints in blocks. In order to obtain the aforementioned mappings, Reichmann et. al. gathered different projection mechanisms for every method. PCA and autoencoders learn a parametric mapping between the original data and the embedding, so projecting new points is straightforward. For non-classical MDS, stress is minimized for a single point while keeping others fixed, a process known as *single scaling* in the literature [1]. A similar strategy is used for t-SNE [2] and UMAP [7], which leverage the k-NN of the projecting point to initizalize the optimizer.

3 Specification and Design of the Solution

- Jusitfy D&C in non-linear DR methods. Most techniques shown in `bigmds` are not generalizable to non-linear DR methods different to Isomap (since it uses MDS). However, D&C and recurrence can work with more general algorithms. Finally, we chose D&C because recurrence can create very small partitions that may be problematic.

- D&C algorithm

4 Development of the Proposal

- DR methods implementations: packages used, problems found during development, tuning of parameters, experiments' methodology

5 Experimentation and Evaluation of the Proposal

- Swiss Roll.
- MNIST.
- Time complexity measurements (including MDS).
- Evaluation and comparison with state of the art.

6 Analysis of Sustainability and Ethical Implications

DESCRIPTION OF THIS SECTION FROM THE REGULATION:

It must include an analysis of the impact of the following gender-related technical aspects:

- issues related to data management and analysis
- issues related to equity, where possible biases are identified and assessed both in the data and in the processes carried out in relation to data management and analysis
- actions carried out to eliminate or mitigate such biases

ACTUAL CONTENT:

- D&C can be used in normal computers, while traditional DR methods require supercomputers to work on big datasets. Hence, we will reduce computing emissions.
- Maybe (it needs testing) DR methods could emphasize biases in the data. This happens because when projecting only a few coordinates, small clusters could be left behind in the remaining not projected coordinates and do not show in the final embedding. Hence, in theory, small communities could become invisible.

7 Conclusions

References

[1] Wojciech Basalaj. "Incremental multidimensional scaling method for database visualization". In: *Visual Data Exploration and Analysis VI.* Ed. by Robert F. Erbacher, Philip C. Chen, and Craig M. Wittenbrink. Vol. 3643. International Society for Optics and Photonics. SPIE, 1999, pp. 149–158. DOI: 10.1117/12.342830. URL: https://doi.org/10.1117/12.342830.

[2] Gordon J. Berman et al. "Mapping the stereotyped behaviour of freely moving fruit flies". In: *Journal of The Royal Society Interface* 11.99 (2014), p. 20140672. DOI: 10.1098/rsif.2014.0672. eprint: https://royalsocietypublishing.org/doi/pdf/10.1098/rsif.2014.0672. URL: https://royalsocietypublishing.org/doi/abs/10.1098/rsif.2014.0672.

[3] Ingwer Borg and Patrick J. F. Groenen. *Modern Multidimensional Scaling: Theory and Applications.* Springer Series in Statistics. New York: Springer, 1997.

[4] Lisha Chen and Andreas Buja and. "Local Multidimensional Scaling for Nonlinear Dimension Reduction, Graph Drawing, and Proximity Analysis". In: *Journal of the American Statistical Association* 104.485 (2009), pp. 209–219. DOI: 10.1198/jasa.2009.0111. eprint: https://doi.org/10.1198/jasa.2009.0111. URL: https://doi.org/10.1198/jasa.2009.0111.

[5] Pedro Delicado and Carlos Pachón-García. "Multidimensional scaling for big data". In: *Advances in Data Analysis and Classification* (2024). ISSN: 1862-5355. DOI: 10.1007/s11634-024-00591-9.

[6] Laurens van der Maaten and Geoffrey Hinton. "Visualizing Data using t-SNE". In: *Journal of Machine Learning Research* 9.86 (2008), pp. 2579–2605. URL: http://jmlr.org/papers/v9/vandermaaten08a.html.

[7] Leland McInnes et al. "UMAP: Uniform Manifold Approximation and Projection". In: *Journal of Open Source Software* 3.29 (2018), p. 861. DOI: 10.21105/joss.00861. URL: https://doi.org/10.21105/joss.00861.

[8] Luca Reichmann, David Hägele, and Daniel Weiskopf. "Out-of-Core Dimensionality Reduction for Large Data via Out-of-Sample Extensions". In: *2024 IEEE 14th Symposium on Large Data Analysis and Visualization (LDAV).* 2024, pp. 43–53. DOI: 10.1109/LDAV64567.2024.00008.

[9] Vin Silva and Joshua Tenenbaum. "Global Versus Local Methods in Nonlinear Dimensionality Reduction". In: *Advances in Neural Information Processing Systems.* Ed. by S. Becker, S. Thrun, and K. Obermayer. Vol. 15. MIT Press, 2002. URL: https://proceedings.neurips.cc/paper_files/paper/2002/file/5d6646aad9bcc0be55b2c82f69750387-Paper.pdf.

[10] Joshua B. Tenenbaum, Vin de Silva, and John C. Langford. "A Global Geometric Framework for Nonlinear Dimensionality Reduction". In: *Science* 290.5500 (2000), pp. 2319–2323. DOI: 10.1126/science.290.5500.2319. URL: https://www.science.org/doi/abs/10.1126/science.290.5500.2319.

8 Annexes