A Distribution-Aware MDS Heuristic Algorithm for Large-Scale Data

[Condensed Mathematical Formulation]

Georgios Vourvachakis

November 2024

1 Problem Statement and Motivation

The raison d'être of classical Multi-dimensional Scaling (MDS) is to project high-dimensional data into a lower-dimensional space while preserving pairwise distances between data points. This formulation of MDS aims to minimize a stress function that quantifies the discrepancy between the original distances and the distances in the embedded lower-dimensional space. The stress function is defined as:

$$stress = \sqrt{\frac{\sum_{i<j}(d_{ij} - \hat{d}_{ij})^2}{\sum_{i<j} d_{ij}^2}}$$

where d_{ij} represents the original distances and $\hat{d}_{ij}$ the distances in the embedded space. MDS opts to find an embedding that minimizes this stress, ensuring that the pairwise distances are as preserved as possible in the new space.

This formulation involves calculating the pairwise distances d_{ij} for all pairs of points in the dataset, which requires $O(n^2)$ operations. This is followed by solving an eigenvalue problem to determine the optimal projection, which has a computational complexity of $O(n^3)$ for the standard MDS algorithm. These high time and space complexities make the classical MDS approach infeasible for large datasets with thousands or millions of points.

The high computational cost of classical MDS motivates the development of more efficient methodologies capable of handling larger datasets. A critical factor contributing to this complexity is the full pairwise distance matrix, which becomes increasingly difficult to compute and process as the dataset size grows. To address this issue, a distribution-aware sampling heuristic approach is proposed. The key insight behind this approach is that not all pairwise distances need to be preserved with equal fidelity in the lower-dimensional embedding. Instead, a more efficient sampling strategy can be employed to select a subset

of the most informative points, ensuring that the distance structure of the data is approximated with lower computational costs.

The proposed distribution-aware sampling approach reduces the number of pairwise distances that need to be computed and optimized, effectively lowering the computational complexity. By focusing on a carefully selected subset of distances, the approach aims to retain the most critical features of the data distribution while achieving significant reductions in the overall computational burden. This heuristic approach, which relies on the distribution of the data, enables scalable MDS computations even for very large datasets, thus making it feasible to apply MDS to problems with hundreds of thousands or even millions of data points.

The motivation behind this methodology is to provide an efficient alternative to classical MDS by leveraging the idea that large-scale datasets often contain redundancies in the distance information that can be exploited to reduce computational requirements. By utilizing a sampling-based approach that is informed by the full/batch data distribution, we can preserve the critical structure of the data with much lower computational overhead, enabling the use of MDS in large-scale applications.

Main source of inspiration is **the bible of statistics** book by Cassella & Berger (2002). Statistical Inference, 2nd Edition.[12]

2 Methodological Foundation

2.1 Distribution−Aware Sampling

A sampling strategy is developed, preserving statistical properties through distribution fitting and validation. For each bin k, optimize the objective:

$$p_k(x) = \arg\min_{f \in \mathcal{F}} \{-\log \mathcal{L}(f|\mathbf{x}) : D_{KS}(F_k, \hat{F}k) < c\alpha\} \tag{1}$$

where:

- $\mathcal{F} = \{Gamma,\ Log-normal,\ Normal,\ Beta\}$
- $\mathcal{L}(f|\mathbf{x})$ is the likelihood function
- D_{KS} is the Kolmogorov-Smirnov test statistic
- c_α is the critical value at significance level α
- F_k is the empirical CDF of bin k
- $\hat{F}_k$ is the CDF estimator

Key references:

1. Distribution fitting and likelihood estimation: D'Agostino, R.B., & Stephens, M.A. (1986). *Goodness-of-fit Techniques*. Marcel Dekker. [1]
2. Maximum likelihood with KS constraints: Clauset, A., Shalizi, C.R., & Newman, M.E. (2009). Power-law distributions in empirical data. *SIAM Review*, 51(4), 661-703. [2]

2.2 Stratified Sampling Implementation

The implementation employs optimal allocation stratified sampling:

$$n_k = N_{sample} \cdot \frac{N_k \sigma_k}{\sum_{i=1}^{L} N_i \sigma_i}$$

where:

- n_k is the allocated sample size for stratum k
- N_k is the population size of stratum k
- σ_k is the standard deviation within stratum k
- L is the total number of strata/bins

Key references:

1. Optimal allocation in stratified sampling: Cochran, W.G. (1977). *Sampling Techniques*, 3rd Edition. Wiley. [3]
2. Modern stratification methods: Tillé, Y. (2006). *Sampling Algorithms*. Springer. [4]

3 Optimization Strategy

3.1 Memory Optimization

Memory complexity reduction from $O(n^2)$ to $O(m^2)$ is achieved through:

$$m = \min\left(5000, \sqrt{\frac{M_{available}}{8\beta}}\right)$$

where β is a safety factor accounting for auxiliary data structures.

Key references:

1. Memory-efficient algorithms: Vitter, J.S. (1985). Random sampling with a reservoir. *ACM Transactions on Mathematical Software*, 11(1), 37-57. [5]

3.2 Distribution Matching Quality

The quality of fit is evaluated using multiple criteria

Kolmogorov-Smirnov test statistic:

$$D_n = \sup_x |F_n(x) - F(x)|$$

Anderson-Darling test statistic:

$$A^2 = n \int_{-\infty}^{\infty} \frac{(F_n(x) - F(x))^2}{F(x)(1 - F(x))} dF(x)$$

Key references:

1. Kolmogorov-Smirnov test: Massey Jr, F.J. (1951). The Kolmogorov-Smirnov test for goodness of fit. *Journal of the American Statistical Association*, 46(253), 68-78. [6]

2. Anderson-Darling test: Anderson, T.W., & Darling, D.A. (1954). A test of goodness of fit. *Journal of the American Statistical Association*, 49(268), 765-769. [7]

4 Results Statistical Analysis/Evaluation

4.1 Distribution Preservation

For each stratum k, the null and alternative Hypothesis are simply stated as:

$$H_0 : F_k(x) = F_{0k}(x) \ for\ all\ x$$
$$H_1 : F_k(x) \neq F_{0k}(x) \ for\ some\ x$$

Using the Benjamini-Hochberg procedure for multiple testing control:

$$\alpha_{BH}(i) = \frac{i}{m}\alpha$$

where:

- i is the rank of the p-value
- m is the number of tests
- α is the desired false discovery rate

Key references:

1. Multiple testing procedures: Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B*, 57(1), 289-300. [8]

2. Distribution comparison methods: Gibbons, J.D., & Chakraborti, S. (2011). *Nonparametric Statistical Inference*, 5th Edition. Chapman and Hall/CRC. [9]

4.2 Computational Efficiency

The computational complexity comparison:

$$T_{reduced} = O(m^3) + O(n \log n) \ll T_{original} = O(n^3)$$

where the speedup factor is approximately:

$$\eta(m; n) \approx \frac{n^3}{m^3 + n \log n}, \quad m \ll n$$

Key references:

1. Computational complexity analysis: Cormen, T.H., Leiserson, C.E., Rivest, R.L., & Stein, C. (2009). *Introduction to Algorithms*, 3rd Edition. MIT Press. [10]

2. Efficient sampling algorithms: Drineas, P., Kannan, R., & Mahoney, M.W. (2006). Fast Monte Carlo algorithms for matrices I: Approximating matrix multiplication. *SIAM Journal on Computing*, 36(1), 132-157. [11]

5 Future Directions

5.1 Alternative Manifold Learning Methods

One shall consider extending the approach to other methods:

1. t-SNE with modified Kullback-Leibler divergence:

$$C = \sum_i \sum_j p_{ij} \log \frac{p_{ij}}{q_{ij}}$$

2. UMAP with fuzzy topological structure:

$$UMAP_{loss} = \sum_{(i,j)} p_{ij} \log \left(\frac{p_{ij}}{q_{ij}} \right) + (1 - p_{ij}) \log \left(\frac{1 - p_{ij}}{1 - q_{ij}} \right)$$

5.2 Enhanced Distribution Metrics

We can also incorporate various metrics (comment on associated jupyter notebook about their ramifications):

1. Wasserstein distance:

$$W_p(P_r, P_g) = \inf_{\gamma \in \Pi(P_r, P_g)} E_{(x,y)\sim\gamma}[\|x - y\|^p]^{1/p}$$

2. KL divergence for distribution comparison:

$$D_{KL}(P\|Q) = \sum_x P(x) \log \frac{P(x)}{Q(x)}$$

6 Recommendations for Future Work

1. Implement adaptive bin sizing based on local density:

$$n_{bins} = \lceil \sqrt{n} \cdot \sigma_{local} \rceil$$

2. Add GPU acceleration for distance calculations:

$$d_{ij} = \|\mathbf{x}_i - \mathbf{x}_j\|_2 \quad (parallelized)$$

3. Develop online learning capabilities with incremental updates:

$$X_{t+1} = X_t + \eta \nabla_{X_t} \{stress(X_t)\}$$

7 Conclusion

This heuristic, data-specific implementation successfully addresses the scalability challenges inherent in traditional Multidimensional Scaling (MDS), which often becomes infeasible when applied to large datasets due to its high computational cost. By introducing a distribution-aware sampling approach, the method effectively reduces the dimensionality of large datasets while preserving key statistical properties, particularly the pairwise distances crucial to the quality of MDS embeddings. The method achieves this without compromising the statistical fidelity of the data, ensuring that the integrity of the underlying structure is maintained in the lower-dimensional space.

The condensed mathematical formulation presented in this work serves as a robust foundation for future enhancements. Not only does it optimize memory and computational efficiency, but it also opens avenues for further exploration of alternative manifold learning methods. For instance, incorporating the Wasserstein distance metric into the methodology offers the potential for more accurate distribution matching, as it accounts for the geometry of the data space in a

more nuanced manner than traditional metrics like the Kolmogorov-Smirnov statistic. Similarly, integrating Uniform Manifold Approximation and Projection (UMAP) techniques could provide improved topological preservation, particularly when working with highly complex, non-linear data structures.

Moving forward, further improvements should focus on the implementation of the suggested adjustments, particularly the Wasserstein distance metrics and UMAP integration. These additions could enable the methodology to better handle non-Euclidean data distributions and enhance the visualization of high-dimensional data in ways that traditional MDS methods might struggle with. It will also be critical to maintain the computational efficiency that was achieved in the current implementation. This can be done by leveraging optimized sampling strategies and GPU-based acceleration for distance calculations, ensuring that the approach remains scalable and efficient even as dataset sizes grow.

In summary, while the current implementation offers significant progress in solving the scalability issues of MDS, future work will focus on refining and extending this approach. The proposed future directions, especially incorporating advanced distribution metrics and manifold learning algorithms, are expected to make the method even more powerful and flexible in handling large-scale data analysis, particularly in fields like machine learning, computational biology, and data science, where high-dimensional datasets are prevalent.

References

[1] D'Agostino, R.B., & Stephens, M.A. (1986). *Goodness-of-fit Techniques*. Marcel Dekker.

[2] Clauset, A., Shalizi, C.R., & Newman, M.E. (2009). Power-law distributions in empirical data. *SIAM Review*, 51(4), 661-703.

[3] Cochran, W.G. (1977). *Sampling Techniques*, 3rd Edition. Wiley.

[4] Tillé, Y. (2006). *Sampling Algorithms*. Springer.

[5] Vitter, J.S. (1985). Random sampling with a reservoir. *ACM Transactions on Mathematical Software*, 11(1), 37-57.

[6] Massey Jr, F.J. (1951). The Kolmogorov-Smirnov test for goodness of fit. *Journal of the American Statistical Association*, 46(253), 68-78.

[7] Anderson, T.W., & Darling, D.A. (1954). A test of goodness of fit. *Journal of the American Statistical Association*, 49(268), 765-769.

[8] Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B*, 57(1), 289-300.

[9] Gibbons, J.D., & Chakraborti, S. (2011). *Nonparametric Statistical Inference*, 5th Edition. Chapman and Hall/CRC.

[10] Cormen, T.H., Leiserson, C.E., Rivest, R.L., & Stein, C. (2009). *Introduction to Algorithms*, 3rd Edition. MIT Press.

[11] Drineas, P., Kannan, R., & Mahoney, M.W. (2006). Fast Monte Carlo algorithms for matrices I: Approximating matrix multiplication. *SIAM Journal on Computing*, 36(1), 132-157.

[12] Cassella, G., & Berger, R.L. (2002). *Statistical Inference*, 2nd Edition. Duxbury.