Interpretable Parametric Neighbor Embedding

Edouard Couplet[1], Pierre Lambert[1], John A. Lee[1,2], and Cyril de Bodt[3]

[1] UCLouvain - ICTEAM/ELEN, Place du Levant 3, 1348 Louvain-la-Neuve, Belgium
[2] UCLouvain - IREC/MIRO, Avenue Hippocrate 55, 1200 Brussels, Belgium
[3] UNAMUR - naXys Research Institute, Rue Grafé 2, 5000 Namur, Belgium

Abstract. Neighbor embedding methods effectively preserve local structures in low-dimensional spaces but are difficult to interpret due to their nonlinear nature. Post-hoc explanations offer only approximate insights. We propose an interpretable neighbor embedding approach that projects each point via a linear combination of shared basis directions, yielding exact explanations through global bases and point-specific coefficients. We demonstrate the method using a t-SNE loss on a single-cell dataset.

Keywords: Dimensionality Reduction · Neighbor Embedding · Interpretability.

1 Introduction

Visual exploration of high-dimensional (HD) data is often performed using neighbor embedding (NE) techniques such as t-SNE [7] and UMAP [8]. Despite their effectiveness, NE methods are nonlinear, making it difficult to relate LD coordinates to HD features [9]. Linear methods like PCA [5] provide interpretable projections but fail to capture complex structures, highlighting a trade-off between DR quality and interpretability. Post-hoc approaches attempt to explain nonlinear embeddings using surrogate models [3][2][6], but these only approximate the embedding and may be inconsistent across the LD space. Our approach addresses this problem by learning a structured locally-linear mapping that enables direct reasoning about the contribution of HD features to each LD coordinate while preserving the flexibility of nonlinear neighbor embeddings.

2 Structured Parametrization

In previous work [4], we explored point-wise local linear mappings and showed that such mappings can produce low-dimensional embeddings with minimal loss in DR performance. However, interpreting individual weight matrices remained challenging. To improve interpretability, we propose a more *structured parametrization*. Let $\xi_i \in \mathbb{R}^d$ denote the HD coordinates of point i and $\mathbf{x}_i \in \mathbb{R}^2$ its LD embedding. We define the point-specific linear map as $\mathbf{W}_i = \sum_{j=1}^{p} c_{ij}\mathbf{B}_j$ and the corresponding embedding as $\mathbf{x}_i = \mathbf{W}_i^\top \xi_i$, where $\mathbf{B}_j \in \mathbb{R}^{d\times 2}$ are *global basis projections*, $c_{ij} \in \mathbb{R}$ are *point-specific coefficients*, and $p \ll n$.

Interpretability arises from the bases $\mathbf{B}_j$ capturing global directions and the coefficients c_{ij} describing how each point combines these directions. To enable flexible out-of-sample embeddings, the coefficients $\mathbf{c}_i = [c_{i1}, \ldots, c_{ip}]^\top$ can be predicted from $\mathbf{x}_i$ using a simple parametric mapping: $\mathbf{c}_i = f_\theta(\xi_i)$. The model can then be trained end-to-end with a neighbor-embedding loss.

3 Implementation and Preliminary Results

We implemented the method using a t-SNE loss. The basis projections $\mathbf{B}_j$ and f_θ (a two-layer neural network with softmax outputs) were optimized jointly in an end-to-end fashion using `PyTorch`'s automatic differentiation. We illustrate the results on the Genomics 10x brain cells dataset [1]. The dataset was subsampled to 25,000 points and preprocessed as in [11]. PCA was applied to retain the first 50 components, with loadings $V \in \mathbb{R}^{d\times 50}$, and our method was applied in this PCA space. The proposed parametrization enables various interpretation options. For instance, we can assess the effect of removing a basis to highlight its contribution. We can also identify most represented genes in a basis by projecting it back to gene space as $\mathbf{G}_j = V\mathbf{B}_j \in \mathbb{R}^{d\times 2}$ and rank genes by their L2 norm across the two LD coordinates, i.e., $\text{score}_g = \|(\mathbf{G}_j)_{g,:}\|_2$. The top genes are then selected as $\arg\max_g(\text{score}_g)$. These examples are illustrated in Figure 1.

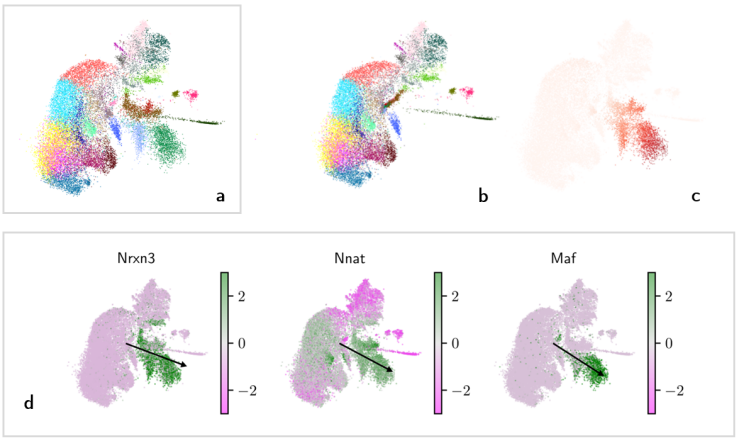

Fig. 1. **a** t-SNE embedding produced by our method. Cluster assignements and colors are from [10]. **b** Embedding obtained by removing the first basis (no retraining needed). **c** Difference between **a** and **b**. **d** Standardized gene expression for the 3 most represented genes in $\mathbf{B}_1$. Black arrows indicate the corresponding directions $(\mathbf{G}_1)_g$.

4 Conclusion and Further Work

Our method aims to combine the power of neighbor embeddings with direct interpretability through structured local mappings. We are currently conducting broader experiments and evaluation. Future work will explore adding structure to the coefficients $\mathbf{c}_i$, for example to enable hierarchical interpretations.

References

1. https://explore.data.humancellatlas.org/projects/74b6d569-3b11-42ef-b6b1-a0454522b4a0
2. Bibal, A., Clarinval, A., Dumas, B., Frénay, B.: Ixvc: An interactive pipeline for explaining visual clusters in dimensionality reduction visualizations with decision trees. Array **11**, 100080 (2021)
3. Bibal, A., Vu, V.M., Nanfack, G., Frénay, B.: Explaining t-sne embeddings locally by adapting lime. In: ESANN. pp. 393–398 (2020)
4. Couplet, E., Lambert, P., Verleysen, M., Mulders, D., Lee, J.A., De Bodt, C.: Natively interpretable t-sne. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases. pp. 107–123. Springer (2023)
5. Jolliffe, I.T.: Principal component analysis and factor analysis. In: Principal component analysis, pp. 115–128. Springer (1986)
6. Lambert, P., Marion, R., Albert, J., Jean, E., Corbugy, S., de Bodt, C.: Globally local and fast explanations of t-SNE-like nonlinear embeddings. In: AIMLAI (2022)
7. Maaten, L.v.d., Hinton, G.: Visualizing data using t-sne. Journal of machine learning research **9**(Nov), 2579–2605 (2008)
8. McInnes, L., Healy, J., Melville, J.: Umap: Uniform manifold approximation and projection for dimension reduction. arXiv preprint arXiv:1802.03426 (2018)
9. Wattenberg, M., Viégas, F., Johnson, I.: How to use t-sne effectively. Distill **1**(10), e2 (2016)
10. Wolf, F.A., Angerer, P., Theis, F.J.: Scanpy: large-scale single-cell gene expression data analysis. Genome biology **19**(1), 15 (2018)
11. Zheng, G.X., Terry, J.M., Belgrader, P., Ryvkin, P., Bent, Z.W., Wilson, R., Ziraldo, S.B., Wheeler, T.D., McDermott, G.P., Zhu, J., et al.: Massively parallel digital transcriptional profiling of single cells. Nature communications **8**(1), 14049 (2017)