Deconfounded Dimension Reduction via Partial Embeddings

Andrew A. Chen[a,b,*], Kelly Clark[a], Blake Dewey[c], Anna DuVal[c], Nicole Pellegrini[c], Govind Nair[d], Youmna Jalkh[e], Samar Khalil[e], Jon Zurawski[e], Peter Calabresi[c], Daniel Reich[d], Rohit Bakshi[e,f], Haochang Shou[a,b,1], Russell T. Shinohara[a,b,1], for the Alzheimer's Disease Neuroimaging Initiative[c], and for the North American Imaging in Multiple Sclerosis Cooperative

[a] *Penn Statistics in Imaging and Visualization Center, Department of Biostatistics, Epidemiology, and Informatics, University of Pennsylvania, Philadelphia, PA*

[b] *Center for Biomedical Image Computing and Analytics, University of Pennsylvania, Philadelphia, PA*

[c] *Department of Neurology, Johns Hopkins University School of Medicine, Baltimore, MD*

[d] *Translational Neuroradiology Section, National Institute of Neurological Disorders and Stroke, National Institutes of Health, Bethesda, MD*

[e] *Department of Neurology, Brigham and Women's Hospital, Harvard Medical School, Boston, MA*

[f] *Department of Radiology, Brigham and Women's Hospital, Harvard Medical School, Boston, MA*

[c] *Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf*

[1] *Equal contribution*

***Correspondence: Andrew A. Chen**

andrewac@pennmedicine.upenn.edu

Abstract

Dimension reduction tools preserving similarity and graph structure such as t-SNE and UMAP can capture complex biological patterns in high-dimensional data. However, these tools typically are not designed to separate effects of interest from unwanted effects due to confounders. We introduce the partial embedding (PARE) framework, which enables removal of confounders from any distance-based dimension reduction method. We then develop partial t-SNE and partial UMAP and apply these methods to genomic and neuroimaging data. Our results show that the PARE framework can remove batch effects in single-cell sequencing data as well as separate clinical and technical variability in neuroimaging measures. We demonstrate that the PARE framework extends dimension reduction methods to highlight biological patterns of interest while effectively removing confounding effects.

Keywords

Dimension reduction; embeddings; confounding effects; genomics; neuroimaging

1 Introduction

Dimension reduction tools such as principal coordinates analysis (PCoA), t-distributed stochastic neighbor embedding (t-SNE), and uniform manifold approximation and projection (UMAP) are widely employed for exploration of high-dimensional data. These methods all identify lower-dimensional embeddings in Euclidean space that preserve information in the original space. These methods has been demonstrated to reveal complex patterns including cell lineages in single-cell RNA sequencing (scRNA-seq) data (**?**) and neurodevelopmental changes in brain volumetric data (**?**). However, in their current form, these methods do not account for covariates and are known to be substantially influenced by confounders such as batch (Hicks et al., 2018).

Researchers have developed several extensions of dimension reduction tools that are designed for removal of confounding effects. For principal component analysis (PCA), researchers developed PCA with adjustment for confounding variation (**?**). Adjusted PCoA

(aPCoA) examines residuals from a linear model on principal coordinates, which are orthogonal to specified confounding variables (Shi et al., 2020). Projected t-SNE orthogonalizes the embeddings at each iteration of the t-SNE optimization to adjust for batch effects (Aliverti et al., 2020). Another method addresses batch effects by using t-SNE to construct a reference embedding based on one batch and then projects observations from other batches onto the reference (Poličar et al., 2021). To date, adjustment for confounders in distance-based dimension reduction methods has required modification of each framework to address this specific problem. Furthermore, many methods including UMAP have not been extended to address confounding.

We develop the partial embedding (PARE) as a generalizable framework for removing nuisance effects from any distance-based dimension reduction method. We achieve this by using the covariate-adjusted dissimilarities from aPCoA as inputs into dimension reduction methods. When the original distances are Euclidean, we can achieve identical results by treating adjusted principal coordinates as input data (see Methods). We refer to these covariate-adjusted dimension reduction results as partial embeddings (PAREs). These PAREs preserve pairwise distances from the original space while removing confounding effects. PAREs can be produced from a broad class of dimension reduction methods including t-SNE (van der Maaten and Hinton, 2008), UMAP (McInnes et al., 2020), Laplacian Eigenmaps (Belkin and Niyogi, 2003), diffusion map embeddings (Coifman et al., 2005), LargeVis (Tang et al., 2016), TriMap (Amid and Warmuth, 2022), ForceAtlas2 (Jacomy et al., 2014) and others. Specifically, we apply the PARE framework to t-SNE and UMAP to develop partial t-SNE (p-t-SNE) and partial UMAP (p-UMAP).

2 Methods

2.1 Adjusted principal coordinates analysis

Let $y_1, y_2, \ldots, y_n$ be multivariate observations from samples $i = 1, 2, \ldots, n$, which can be features from genomics, neuroimaging, or any type of multivariate data. Let $D = (d_{ij})_{n \times n}$ denote the sample dissimilarity matrix computed on these observations y_i, where $d_{ij} =$

$d(y_i, y_j)$ and d is a chosen dissimilarity function. Define the doubly-centered dissimilarity matrix $G = (I - \mathbf{1}\mathbf{1}^T)A(I - \mathbf{1}\mathbf{1}^T)$ where $A = (-\frac{1}{2}d_{ij}^2)_{n\times n}$. Principal coordinates analysis (PCoA) finds coordinates in Euclidean space that optimally preserve dissimiliarites from the original space. The classical solution finds these coordinates via eigendecomposition of G (Gower, 1966). Decomposing $G = U\Lambda U^T$, these coordinates are given by $Z = U\Lambda^{1/2}$. Under Euclidean dissimilarities, the principal coordinates Z preserve the exact distances from the original space. If the original dissimilarities are non-Euclidean, Z may contain imaginary coordinates. Adding a constant to every pairwise dissimilarity can produce coordinates in Euclidean space (Cailliez, 1983).

In adjusted principal coordinates analysis (aPCoA), a linear model is used to remove the effect of nuisance covariates from the principal coordinates (Shi et al., 2020). Let X be an $n \times p$ design matrix of nuisance covariates with corresponding projection matrix $H = X(X^TX)^{-1}X^T$. Covariate-adjusted coordinates are given by $E = (I - H)Z$ with adjusted dissimilarity matrix $\Delta = EE^T = (I - H)G(I - H)$. The first two adjusted coordinates are typically used for visualization.

2.2 Partial embeddings

We develop the partial embedding (PARE) framework by leveraging aPCoA to remove the effect of confounders from pairwise dissimilarities in the original space. The covariate-adjusted dissimilarity matrix $\Delta = (\delta_{ij})_{n\times n} = (I - H)G(I - H)$ is used as an input into dimension reduction methods. For example, UMAP defines affinities based on dissimilarity metrics d as $v_{ij} = v_{j|i} + v_{i|j} - v_{j|i}v_{i|j}$ where

$$v_{j|i} = \exp[(-d(y_i, y_j) - \rho_i)/\tau_i]$$

and ρ_i are the dissimilarity to the nearest neighbor of y_i and τ_i are normalizing factors computed based on dissimilarities among a chosen number of nearest neighbors. A PARE for UMAP can be formulated via the adjusted affinities

$$v_{j|i}^{\text{PARE}} = \exp[(-\delta_{ij} - \rho_i)/\tau_i].$$

For Euclidean distances, dimension reduction methods can instead take the principal coordinates as inputs. As examples, we highlight how t-SNE and UMAP can be equivalently

formulated in terms of principal coordinates. t-SNE measures similarity in the original space as affinities $p_{ij} = \frac{p_{j|i}+p_{i|j}}{2n}$ under a Gaussian kernel where

$$p_{j|i} = \frac{\exp(-\|y_i - y_j\|^2/2\sigma_i^2)}{\sum_{k\neq i}\exp(-\|y_i - y_k\|^2/2\sigma_i^2)},$$

$\|\cdot\|$ is the Euclidean norm, and σ_i are chosen to yield a specified perplexity value for each observation. UMAP using Euclidean distances defines similarities using a locally adaptive exponential kernel as $v_{ij} = v_{j|i} + v_{i|j} - v_{j|i}v_{i|j}$ where

$$v_{j|i} = \exp[(-\|y_i - y_j\| - \rho_i)/\tau_i].$$

Let z_i denote the principal coordinate vector for observation i. For Euclidean distances in the original space, the principal coordinates have identical pairwise distances such that $\|z_i - z_j\| = \|y_i - y_j\|$. Then the t-SNE and UMAP affinities can be written in terms of principal coordinates as

$$p_{j|i} = \frac{\exp(-\|z_i - z_j\|^2/2\sigma_i^2)}{\sum_{k\neq i}\exp(-\|z_i - z_k\|^2/2\sigma_i^2)},$$
$$v_{j|i} = \exp[(-\|z_i - z_j\| - \rho_i)/\tau_i].$$

We develop PAREs for any dimension reduction method based on Euclidean distances by instead taking adjusted principal coordinates $e_i = (I - H)z_i$ as input data. These adjusted coordinates preserve dissimilarities while removing unwanted effects due to the nuisance covariates X. We outline the steps in constructing PAREs using Euclidean distances below:

1. Obtain principal coordinates Z from the original data from the Euclidean distance matrix D as described in subsection 2.1.

2. Using a linear model, residualize Z with respect to nuisance covariates X to obtain adjusted coordinates $E = (I - H)Z$, where $H = X(X^TX)^{-1}X^T$.

3. Input the adjusted coordinates E to any dimension reduction method based on Euclidean distances.

Obtaining these adjusted coordinates only requires eigendecomposition of the original dissimilarity matrix followed by residualization using a linear model. Both steps are implemented via multiple packages in R, Python, MATLAB, and other programming languages.

For our investigation, we apply our PARE framework to t-SNE and UMAP using Euclidean distances to develop p-t-SNE and p-UMAP. We use R (version 4.1.1) implementations for t-SNE and UMAP in the packages Rtsne (version 0.15) and umap (version 0.2.7.0). Throughout our applications, we choose the perplexity as 10 for t-SNE and the number of nearest neighbors as 15 for UMAP.

2.3 Human pancreatic cell scRNA-seq data

We apply PAREs to human pancreatic cell scRNA-seq data to remove batch and donor effects from data collected across four separate studies with varying number of cells and RNA-seq protocol. We include RNA-seq data from Baron et al. (2016) (8569 cells, inDrop protocol), Lawlor et al. (2017) (1050 cells, SMARTer), Muraro et al. (2016) (2122 cells, CEL-Seq2), and Segerstolpe et al. (2016) (2133 cells, SMART-Seq2). We treat each study as a separate batch and treat each donor as distinct across studies. We follow a pre-processing pipeline proposed in Lun et al. (2016). First, we use Scran (release 3.15) in R to perform log-normalization and selection of highly variable genes (HVGs) using the counts data from each study. Genes that are not present in all four studies were removed from further evaluation. We then perform normalization by computing size factors across pools of cells, then obtaining factors for each cell via a deconvolution approach (Lun et al., 2016). Within each study, locally weighted scatterplot smoothing (LOESS) is applied to model the mean-variance relationship among genes. We then use a weighted arithmetic mean of mean and variance statistics across studies to select 2,000 HVGs.

We remove cells labelled as "unclear", "none", "unclassified" or "co-expression". After pre-processing, our human pancreatic cell dataset is comprised of 13,369 cells with four batches, 26 donors, and 13 cell types. We apply p-t-SNE and p-UMAP to remove batch effect or donor effects in the embeddings. For comparison, we also apply projected t-SNE for batch correction (BC-t-SNE, Aliverti et al., 2020) with perplexity of 10. We compare our methods visually and numerically using the local inverse Simpson's index (LISI, Korsunsky et al., 2019). For measuring integration of cells across batches, we compute LISI for batch (bLISI), which captures the effective number of batches in a local neighborhood around each cell. We also examine LISI computed for cell type (cLISI), which captures the number of neighboring

cell types and decreases as the separation between cell types increases. We compute bLISI and cLISI across a range of perplexity values, which capture different neighborhood sizes.

2.4 ADNI cortical thickness dataset

We apply PAREs to brain cortical thickness data from the Alzheimer's Disease Neuroimaging Initiative (ADNI) to distinguish technical and biological variability. The data for this study consist of baseline scans which are processed using the ANTs longitudinal single-subject template pipeline (Tustison et al., 2019) with code available on GitHub (https://github.com/ntustison/CrossLong). The ADNI study obtained informed consent from all participants. Institutional review boards approved the study at all of the contributing institutions. Further details for this preprocessing pipeline can be found in Beer et al. (2020).

The full sample consists of 505 subjects, 213 of whom are imaged on scanners manufactured by Siemens, 70 by Philips, and 222 by GE. The sample has a mean age of 75.3 (SD 6.70) and is comprised of 278 (55%) males, 115 (22.8%) Alzheimer's disease (AD) patients, 239 (47.3%) late mild cognitive impairment (LMCI), and 151 (29.9%) cognitively normal (CN) individuals. We apply p-t-SNE and p-UMAP to separate effects of diagnosis and scanner.

2.5 NAIMS traveling subjects study

To examine if PAREs can identify technical variability not visible in t-SNE and UMAP embeddings, we apply our PAREs to a study of patients with multiple sclerosis (MS) with multiple scan-rescan images across four different sites in the North American Imaging in Multiple Sclerosis (NAIMS) Cooperative. These sites include the University of Pennsylvania (Penn), the Brigham and Women's Hospital (BWH), the National Institutes of Health (NIH), and the Johns Hopkins University (Hopkins). Nine of the eleven participants are scanned at all four study centers. The mean age of our 11 participants (4 male, 7 female) at time of enrollment was 38 (range 29-47). We received informed consent from all participants, which was approved by the University of Pennsylvania's institutional review board (IRB).

A standardized high-resoution 3-tesla (3T) MRI brain scan protocol developed by the

NAIMS Cooperative was performed at each site (Wattjes et al., 2021). Images were acquired on Siemens Skyra (BWH, NIH), Siemens Prisma (Penn), and Philips Achieva (Hopkins) scanners. Each participant had two scans acquired on the same day at each visit to the study center.

Prior to automated segmentation, images undergo bias correction via nonuniform intensity normalization (N4ITK, Tustison et al., 2010) and FLAIR images are rigidly aligned to the corresponding T1-weighted image within a given scan session. Brain extraction is performed using Multi-Atlas Skull Stripping (MASS, Doshi et al., 2013) and intensity normalization is performed using WhiteStripe (Shinohara et al., 2014). White matter and gray matter volumes are estimated using Joint Label Fusion (JLF, Wang and Yushkevich, 2013), a segmentation method that leverages information from several atlases via weighted voting. These JLF volumes are used as inputs into t-SNE, UMAP, and our PARE methods. We use PAREs to identify scanner effects independently of within-subject similarities.

3 Results

3.1 Case study 1: Human pancreatic cells

We first apply PAREs to analyze scRNA-seq data from human pancreatic cells across four published studies, treated as separate batches (Baron et al., 2016; Lawlor et al., 2017; Muraro et al., 2016; Segerstolpe et al., 2016). We observe clear batch effects in the original t-SNE and UMAP visualizations along with a lack of integration among several cell types (Fig. 1). Applying p-t-SNE and p-UMAP to remove batch effects considerably reduces separation by batch both visually and numerically, as measured by increases in the local inverse Simpson's index for batch (bLISI, Supplementary Fig. 1). Remaining batch differences can partially be explained by donor effects, since PAREs with respect to donor show greater visual integration across batches and higher bLISI. PAREs also achieve greater distinction of cell types as measured by decreases in cell type LISI (cLISI, Supplementary Fig. 1). Comparing p-t-SNE to the existing projected t-SNE for batch correction (BC-t-SNE, Aliverti et al., 2020), we find that BC-t-SNE achieves greater batch integration but obscures important biological

patterns that separate cell types (median cLISI increases from 1.22 in t-SNE to 1.32 in BC-t-SNE). We also show in **Supplementary Fig. 2** that effective results can be achieved by computing a subset of principal coordinates, which is less computationally intensive. Using scRNA-seq data, we demonstrate that PAREs can uniquely isolate biological variability from unwanted sampling effects in scRNA-seq data.

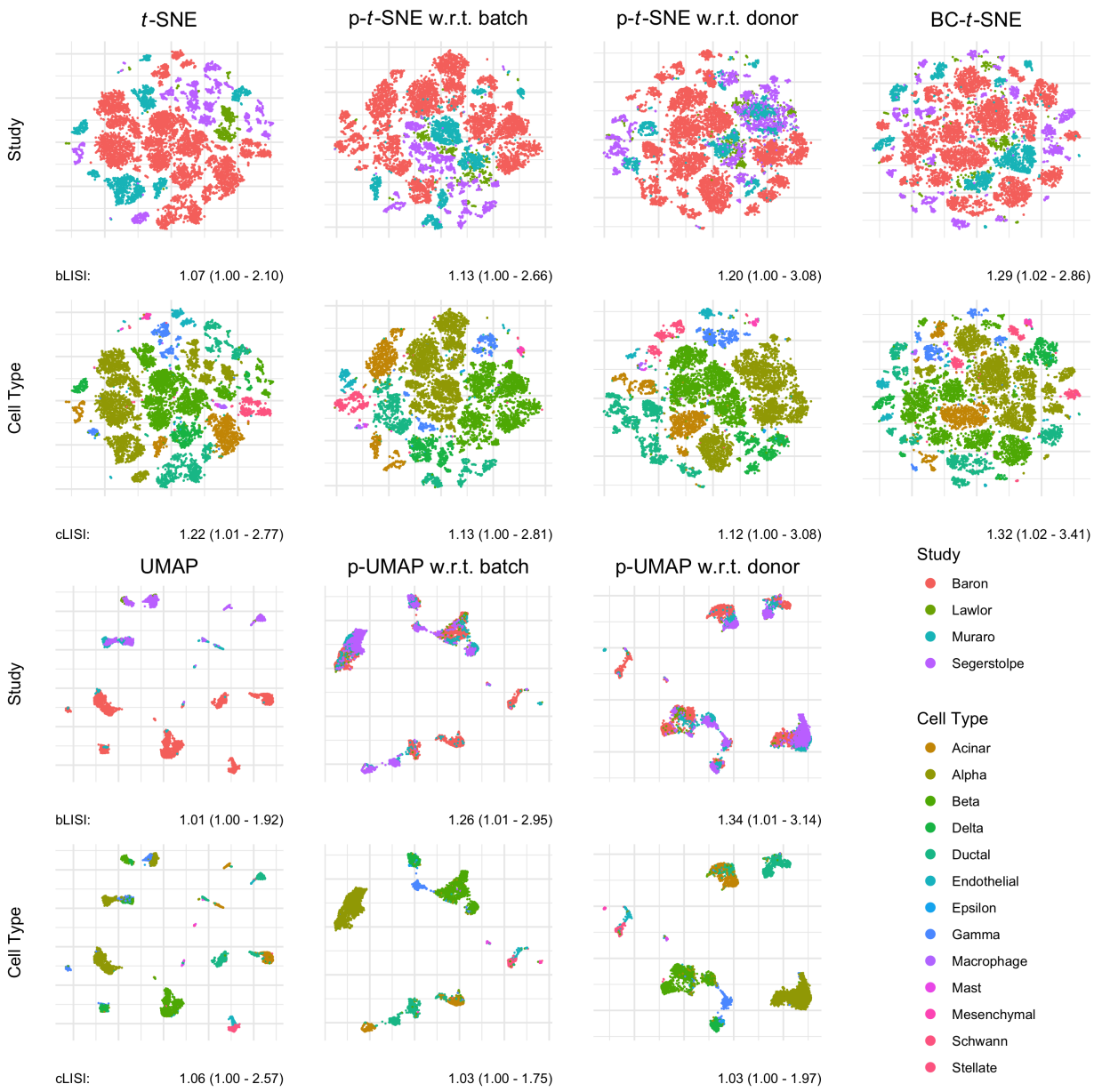

Figure 1: **Embeddings and partial embeddings of single-cell RNA-sequencing measurements from 13,369 human pancreatic cells across four studies.** The original counts data is log-normalized and reduced to 2,000 highly variable genes. Local Simpson's index is computed for each cell for batch (bLISI) and cell type (cLISI) with the median, 2.5% quantile, and 97.5% quantile shown. Higher bLISI indicates greater integration across batches and lower cLISI indicates greater separation between cell types. Partial t-SNE (p-t-SNE) and partial UMAP (p-UMAP) adjust for either batch or donor effects. We compare our new methodology to the existing projected t-SNE for batch correction (BC-t-SNE). All t-SNE embeddings have a perplexity of 10 and UMAP embeddings use 15 nearest neighbors.

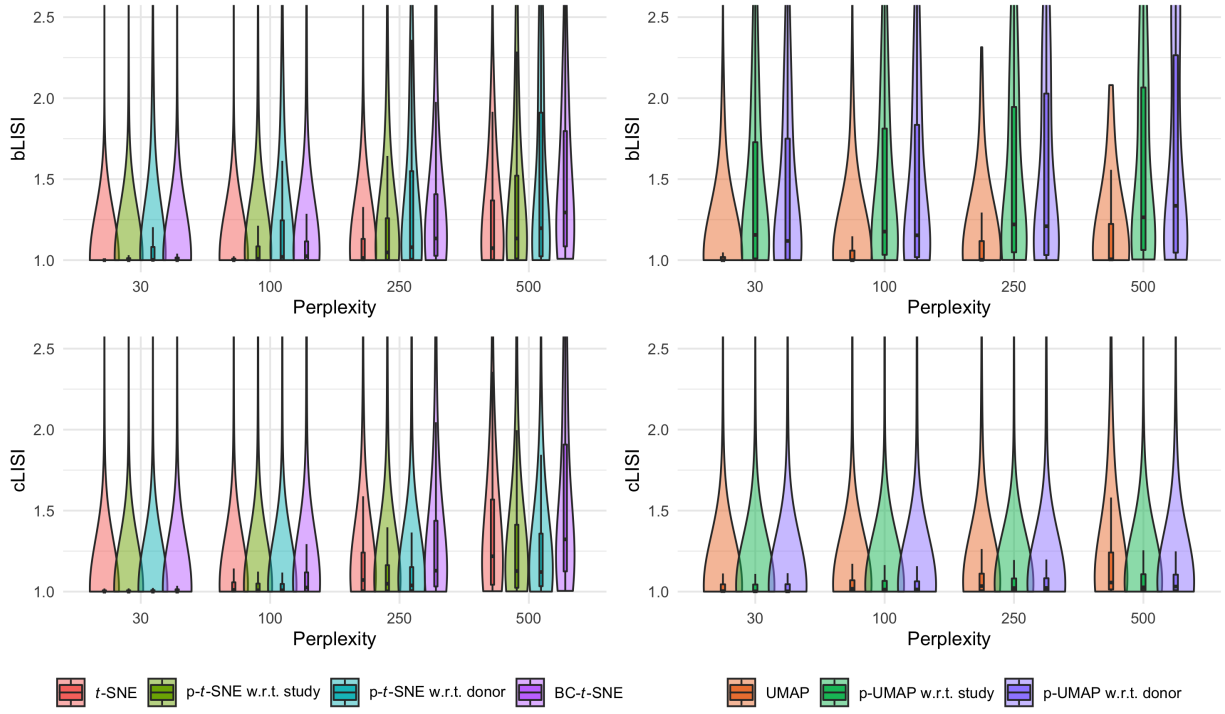

Figure 2: **Local Simpson's index for batch (bLISI) and cell type (cLISI) across multiple perplexity values.** LISI is computed using distances from the embeddings. The original embeddings and partial embeddings are compared across perplexity values, which capture different neighborhood sizes around each cell.

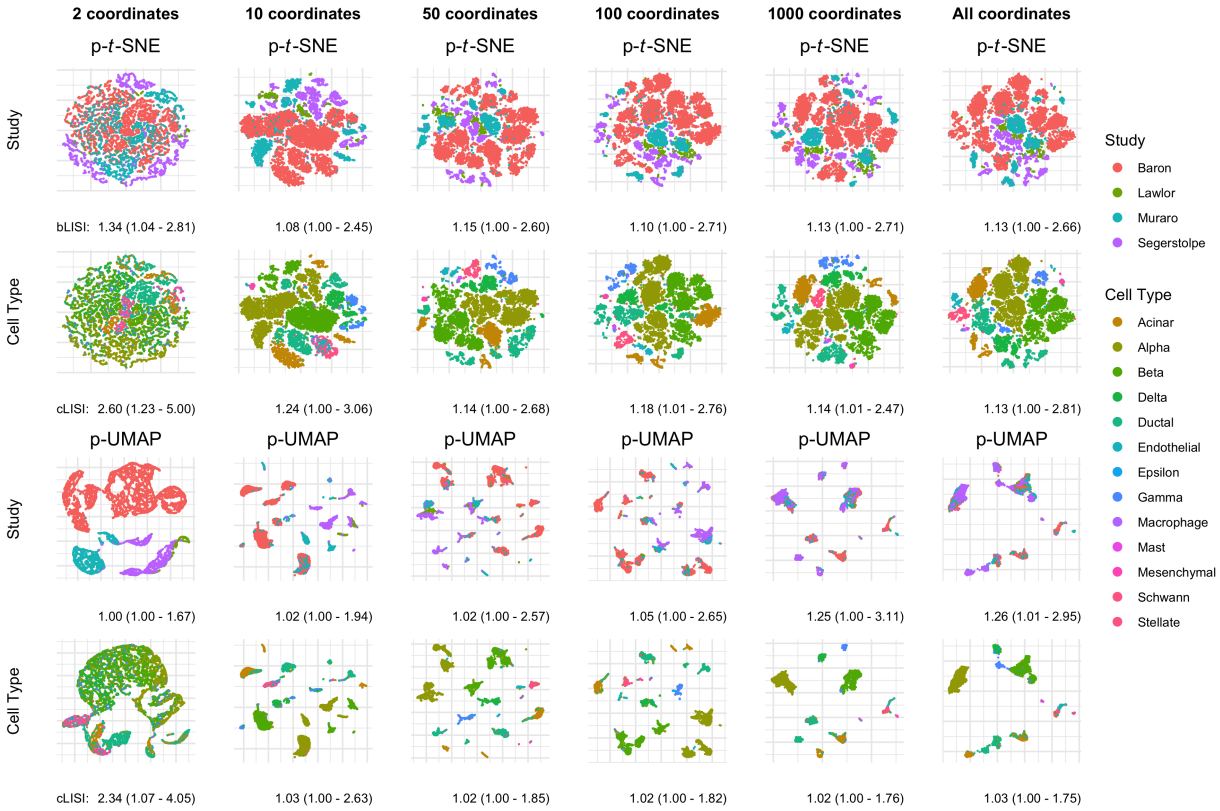

Figure 3: **Single-cell RNA-sequencing data embeddings and partial embeddings with respect to batch across varying numbers of principal coordinates.** Each dimension reduction method takes a subset of adjusted or unadjusted principal coordinates. The dimension of this subset is varied across figure columns. Partial t-SNE (p-t-SNE) and partial UMAP (p-UMAP) adjust for batches.

3.2 Case study 2: Brain cortical thickness

We next apply PAREs to brain cortical structure measurements to separate biological effects from scanner-related artifacts in the Alzheimer's Disease Neuroimaging Initiative (ADNI) study. In the ADNI study, researchers previously identified diagnosis-related atrophy in cortical structure and notable batch effects due to differences in scanner properties across study sites (Querbes et al., 2009; Beer et al., 2020). In Fig. 4a, we observe that the original embeddings display both of these effects. However, the confounding scanner effects result in the overlap among images acquired on a Siemens scanner and those from patients with an Alzheimer's disease (AD) diagnosis. To specifically investigate differences between people with and without AD, we demonstrate that PAREs adjusted for scanner manufacturer maintain diagnosis effects while obscuring scanner influence (Fig. 4a). PAREs are also used to examine scanner effects without the influence of diagnosis effects, highlighting known differences among scanners in the ADNI study.

3.3 Case study 3: Traveling subject brain volumetric data

Finally, as a proof of concept, we use PAREs to identify scanner effects in brain white matter and gray matter volumes collected as part of a multi-site traveling subjects study of multiple sclerosis (MS). The study involves eleven MS patients with multiple scans across four major imaging centers. We include Siemens images with distortion correction, which was designed to reduce differences with the Philips scanner at Johns Hopkins University (Hopkins). Original t-SNE and UMAP embeddings clearly separate white matter and gray matter volumetric measurements by subject regardless of site (Fig. 4b). However, these original embeddings do not capture other types of variability, including potential site effects. We apply p-t-SNE and p-UMAP to remove subject effects and discover deviation of images acquired on the Philips scanner at Hopkins from those acquired on Siemens scanners at other sites (Fig. 4b). Here, we show that PAREs can identify technical variability in neuroimaging measures that could not be detected in the original embeddings.

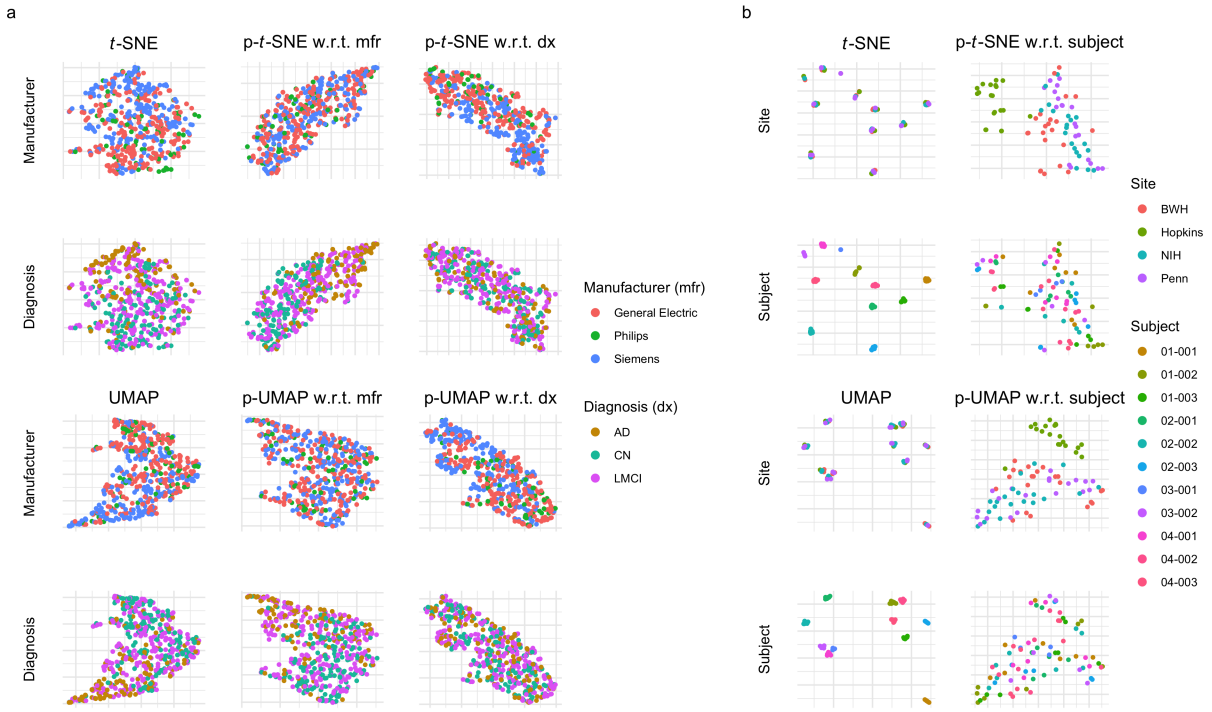

Figure 4: **Application of partial embeddings to brain cortical thickness measurements (a) and regional volumes (b) across two neuroimaging studies. (a)** visualizes cortical thickness data from the Alzheimer's Disease Neuroimaging Initiative, from which we include 505 participants. These participants are diagnosed as cognitively normal (CN), having late mild cognitive impairment (LMCI), or having Alzheimer's disease (AD). Participants are acquired across many scanners with three distinct manufacturers. **(b)** shows results from a traveling subjects study of eleven multiple sclerosis (MS) patients with multiple images across four study sites. The Hopkins site uses a Philips scanner while the three other sites use Siemens scanners.

4 Discussion

We propose the PARE framework, which extends any distance-based dimension reduction method to adjust for confounders. Our analyses demonstrate that PAREs can be used to target specific patterns in high-dimensional data by removal of confounders. We demonstrate that our proposed PAREs are able to remove batch effects in scRNA-seq exploration, emphasize diagnosis-related changes in brain cortical structure, and identify scanner effects in brain volumetric measurements. For dimension reduction based on Euclidean distances, our PARE framework relies solely on PCoA and linear regression, which are both widely available and computationally efficient. While we only investigate PAREs built on t-SNE and UMAP, this framework can be readily applied to a broad class of dimension reduction methods based on distances from the original space.

The PARE framework opens several new directions for methodological development. Future investigations can examine how the PARE framework performs for extensions of methods not considered in this article. PAREs are constructed from the residuals of a linear model, but other models including linear mixed models and general additive models could also be considered for longitudinal and non-linear effects. Extensions of this framework can readily incorporate multiple complex data types by integrating at the level of principal coordinates. Furthermore, PAREs can be extended to examine data types independently of one another by projection of dissimilarity matrices (Székely and Rizzo, 2014). In summary, the PARE framework is able to remove nuisance effects in any distance-based dimension reduction tool. Our framework enables discovery of notable patterns in complex high-dimensional data and introduces a foundation for future methodological research.

Acknowledgements

This work was supported by the National Institute of Neurological Disorders and Stroke (grant numbers R01 NS085211 and R01 NS060910 and the Intramural Research Program), the National Multiple Sclerosis Society (RG-1707-28586), the National Institute of Mental Health (R01 MH123550 and R01 MH112274), and a seed grant from the University of Penn-

sylvania Center for Biomedical Image Computing and Analytics (CBICA). The content is solely the responsibility of the authors and does not necessarily represent the official views of the funding agencies.

RB has received consulting fees from Bristol-Myers Squibb and EMD Serono and research support from Bristol-Myers Squibb, EMD Serono, and Novartis. DSR has received research funding from Abata Therapeutics, Sanofi-Genzyme, and Vertex Pharmaceuticals, all unrelated to the current study. RTS receives consulting income from Octave Bioscience and compensation for scientific reviewing from the American Medical Association.

References

Aliverti, E., Tilson, J. L., Filer, D. L., Babcock, B., Colaneri, A., Ocasio, J., Gershon, T. R., Wilhelmsen, K. C., and Dunson, D. B. (2020). Projected t-SNE for batch correction. *Bioinformatics (Oxford, England)*, 36(11):3522–3527.

Amid, E. and Warmuth, M. K. (2022). TriMap: Large-scale Dimensionality Reduction Using Triplets. *arXiv:1910.00204 [cs, stat]*.

Baron, M., Veres, A., Wolock, S. L., Faust, A. L., Gaujoux, R., Vetere, A., Ryu, J. H., Wagner, B. K., Shen-Orr, S. S., Klein, A. M., Melton, D. A., and Yanai, I. (2016). A Single-Cell Transcriptomic Map of the Human and Mouse Pancreas Reveals Inter- and Intra-cell Population Structure. *Cell Systems*, 3(4):346–360.e4.

Beer, J. C., Tustison, N. J., Cook, P. A., Davatzikos, C., Sheline, Y. I., Shinohara, R. T., and Linn, K. A. (2020). Longitudinal ComBat: A method for harmonizing longitudinal multi-scanner imaging data. *NeuroImage*, 220:117129.

Belkin, M. and Niyogi, P. (2003). Laplacian Eigenmaps for Dimensionality Reduction and Data Representation. *Neural Computation*, 15(6):1373–1396.

Cailliez, F. (1983). The analytical solution of the additive constant problem. *Psychometrika*, 48(2):305–308.

Coifman, R. R., Lafon, S., Lee, A. B., Maggioni, M., Nadler, B., Warner, F., and Zucker, S. W. (2005). Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps. *Proceedings of the National Academy of Sciences*, 102(21):7426–7431.

Doshi, J., Erus, G., Ou, Y., Gaonkar, B., and Davatzikos, C. (2013). Multi-Atlas Skull-Stripping. *Academic Radiology*, 20(12):1566–1576.

Gower, J. C. (1966). Some Distance Properties of Latent Root and Vector Methods Used in Multivariate Analysis. *Biometrika*, 53(3/4):325–338.

Hicks, S. C., Townes, F. W., Teng, M., and Irizarry, R. A. (2018). Missing data and technical variability in single-cell RNA-sequencing experiments. *Biostatistics*, 19(4):562–578.

Jacomy, M., Venturini, T., Heymann, S., and Bastian, M. (2014). ForceAtlas2, a Continuous Graph Layout Algorithm for Handy Network Visualization Designed for the Gephi Software. *PLOS ONE*, 9(6):e98679.

Korsunsky, I., Millard, N., Fan, J., Slowikowski, K., Zhang, F., Wei, K., Baglaenko, Y., Brenner, M., Loh, P.-r., and Raychaudhuri, S. (2019). Fast, sensitive and accurate integration of single-cell data with Harmony. *Nature Methods*, 16(12):1289–1296.

Lawlor, N., George, J., Bolisetty, M., Kursawe, R., Sun, L., Sivakamasundari, V., Kycia, I., Robson, P., and Stitzel, M. L. (2017). Single-cell transcriptomes identify human islet cell signatures and reveal cell-type-specific expression changes in type 2 diabetes. *Genome Research*, 27(2):208–222.

Lun, A. T. L., Bach, K., and Marioni, J. C. (2016). Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. *Genome Biology*, 17:75.

McInnes, L., Healy, J., and Melville, J. (2020). UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv:1802.03426 [cs, stat]*.

Muraro, M. J., Dharmadhikari, G., Grün, D., Groen, N., Dielen, T., Jansen, E., van Gurp, L., Engelse, M. A., Carlotti, F., de Koning, E. J. P., and van Oudenaarden, A. (2016). A Single-Cell Transcriptome Atlas of the Human Pancreas. *Cell Systems*, 3(4):385–394.e3.

Poličar, P. G., Stražar, M., and Zupan, B. (2021). Embedding to reference t-SNE space addresses batch effects in single-cell classification. *Machine Learning*.

Querbes, O., Aubry, F., Pariente, J., Lotterie, J.-A., Démonet, J.-F., Duret, V., Puel, M., Berry, I., Fort, J.-C., Celsis, P., and The Alzheimer's Disease Neuroimaging Initiative (2009). Early diagnosis of Alzheimer's disease using cortical thickness: Impact of cognitive reserve. *Brain*, 132(8):2036–2047.

Segerstolpe, Å., Palasantza, A., Eliasson, P., Andersson, E.-M., Andréasson, A.-C., Sun, X., Picelli, S., Sabirsh, A., Clausen, M., Bjursell, M. K., Smith, D. M., Kasper, M., Ämmälä, C., and Sandberg, R. (2016). Single-Cell Transcriptome Profiling of Human Pancreatic Islets in Health and Type 2 Diabetes. *Cell Metabolism*, 24(4):593–607.

Shi, Y., Zhang, L., Do, K.-A., Peterson, C. B., and Jenq, R. R. (2020). aPCoA: Covariate adjusted principal coordinates analysis. *Bioinformatics*, 36(13):4099–4101.

Shinohara, R. T., Sweeney, E. M., Goldsmith, J., Shiee, N., Mateen, F. J., Calabresi, P. A., Jarso, S., Pham, D. L., Reich, D. S., and Crainiceanu, C. M. (2014). Statistical normalization techniques for magnetic resonance imaging. *NeuroImage: Clinical*, 6:9–19.

Székely, G. J. and Rizzo, M. L. (2014). Partial distance correlation with methods for dissimilarities. *The Annals of Statistics*, 42(6):2382–2412.

Tang, J., Liu, J., Zhang, M., and Mei, Q. (2016). Visualizing Large-scale and High-dimensional Data. In *Proceedings of the 25th International Conference on World Wide Web*, WWW '16, pages 287–297, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Tustison, N. J., Avants, B. B., Cook, P. A., Zheng, Y., Egan, A., Yushkevich, P. A., and Gee, J. C. (2010). N4ITK: Improved N3 Bias Correction. *IEEE Transactions on Medical Imaging*, 29(6):1310–1320.

Tustison, N. J., Holbrook, A. J., Avants, B. B., Roberts, J. M., Cook, P. A., Reagh, Z. M., Duda, J. T., Stone, J. R., Gillen, D. L., Yassa, M. A., and Initiative, f. t. A. D. N. (2019). Longitudinal Mapping of Cortical Thickness Measurements: An Alzheimer's Disease Neuroimaging Initiative-Based Evaluation Study. *Journal of Alzheimer's Disease*, 71(1):165–183.

van der Maaten, L. and Hinton, G. (2008). Visualizing Data using t-SNE. *Journal of Machine Learning Research*, 9(Nov):2579–2605.

Wang, H. and Yushkevich, P. (2013). Multi-atlas segmentation with joint label fusion and corrective learning—an open source implementation. *Frontiers in Neuroinformatics*, 7.

Wattjes, M. P., Ciccarelli, O., Reich, D. S., Banwell, B., de Stefano, N., Enzinger, C., Fazekas, F., Filippi, M., Frederiksen, J., Gasperini, C., Hacohen, Y., Kappos, L., Li, D. K. B., Mankad, K., Montalban, X., Newsome, S. D., Oh, J., Palace, J., Rocca, M. A., Sastre-Garriga, J., Tintoré, M., Traboulsee, A., Vrenken, H., Yousry, T., Barkhof, F., Rovira, À., Magnetic Resonance Imaging in Multiple Sclerosis study group, Consortium of Multiple Sclerosis Centres, and North American Imaging in Multiple Sclerosis Cooperative MRI guidelines working group (2021). 2021 MAGNIMS-CMSC-NAIMS consensus recommendations on the use of MRI in patients with multiple sclerosis. *The Lancet. Neurology*, 20(8):653–670.