Exploring High-Dimensional Biological Data with Sparse Contrastive Principal Component Analysis

A Preprint

Philippe Boileau
Graduate Group in Biostatistics,
University of California, Berkeley
philippe_boileau@berkeley.edu

Nima S. Hejazi
Graduate Group in Biostatistics and
Center for Computational Biology,
University of California, Berkeley
nhejazi@berkeley.edu

Sandrine Dudoit
Department of Statistics,
Division of Biostatistics, and
Center for Computational Biology,
University of California, Berkeley
sandrine@stat.berkeley.edu

November 12, 2019

Abstract

Motivation: Statistical analyses of high-throughput sequencing data have re-shaped the biological sciences. In spite of myriad advances, recovering interpretable biological signal from data corrupted by technical noise remains a prevalent open problem. Several classes of procedures, among them classical dimensionality reduction techniques, and others incorporating subject-matter knowledge, have provided effective advances; however, no procedure currently satisfies the dual objectives of recovering stable and relevant features simultaneously.

Results: Inspired by recent proposals for making use of control data in the removal of unwanted variation, we propose a variant of principal component analysis that extracts sparse, stable, interpretable, and relevant biological signal. The new methodology is compared to competing dimensionality reduction approaches through a simulation study as well as via analyses of several publicly available protein expression, microarray gene expression, and single-cell transcriptome sequencing datasets.

Availability: A free and open-source software implementation of the methodology, the `scPCA R` package, is made available via the Bioconductor Project. Code for all analyses presented in the paper are also made available via GitHub.

***K*eywords** dimensionality reduction, principal component analysis, high-dimensional inference, sparsity, stability, unwanted variation, single-cell, genomics, computational biology

1 Introduction

Principal component analysis (PCA) is a well-known dimensionality reduction technique, widely used for data pre-processing and exploratory data analysis (EDA). Although popular for the interpretability of its results and ease of implementation, PCA's ability to extract signal from high-dimensional data is demonstrably unstable [28, 14], in that its recovered results can vary widely with perturbations of the data [30]. What is more, PCA is often unable to reduce the dimensionality of the data in a contextually meaningful manner [25, 1]. Consequently, variants of PCA have been developed in attempts to remedy these severe issues, including, among many others, sparse PCA (SPCA) [37], which increases the interpretability and stability

of the principal components in high dimensions by sparsifying the loadings, and contrastive PCA (cPCA) [1], which captures relevant information in the data by eliminating technical effects through comparison to a so-called background dataset. While SPCA and cPCA have both individually proven useful in resolving distinct shortcomings of PCA, neither is capable of simultaneously tackling the issues of interpretability, stability, and relevance. We propose a combination of these techniques, *sparse constrastive PCA* (scPCA), which draws on cPCA to remove technical effects and on SPCA for sparsification of the loadings, thereby extracting interpretable, stable, and uncontaminated signal from high-dimensional biological data.

1.1 Motivation

A longstanding problem in genomics and related disciplines centers on teasing out important biological signal from technical noise, i.e., removing *unwanted* variation corresponding to experimental artifacts (e.g., batch effects). A common preliminary approach for accomplishing such a task involves the application of classical PCA to capture and deflate technical noise, followed by traditional statistical inference techniques (e.g., clustering cells, testing for differences in mean gene expression levels between populations of cells) [23]. Such an approach operates under the assumption that meaningful biological signal is not present in the leading principal components (PCs), and that the removal of the variance contained therein allows recovery of the signal previously masked by technical noise. Should these assumptions prove unmet, relevant biological signal may be unintentionally discarded, or worse, technical noise may be significantly amplified.

Several more sophisticated approaches have been proposed, including the use of control genes [7, 26] and control samples [8] whose behavior is known *a priori*. Unfortunately, access to such controls may be severely limited in many settings (e.g., as with prohibitively expensive assays). Alternative approaches, for use in settings where control genes or control samples are unavailable, such as surrogate variable analysis [18], reconstruct sources of unwanted variation that may subsequently be controlled for via covariate adjustment in a typical regression modeling framework. In the context of single-cell RNA-seq data, a class of data that has garnered much interest due to the granularity of biological information it encodes, related approaches have been combined as part of the ZINB-WaVE methodology [27], which uses a strategy based on factor analysis to remove unwanted variation.

Although such approaches have proven useful, model-based techniques rely on assumptions about the data-generating process to target biological signal, warranting that much caution be taken in their use. Additionally, owing to the diversity of experimental settings in high-dimensional biology, such techniques are often targeted to specific experimental paradigms (e.g., bulk RNA-seq but *not* single-cell RNA-seq). Violations of the assumptions embedded in these techniques may often be difficult — impossible, even — to diagnose, leading to a lack of overlap in findings between such model-based approaches when applied to the same datasets. Accordingly, Zhang *et al.* [33] have shown the lack of consensus among model-based differential expression techniques on RNA-seq datasets, demonstrating that their use gives rise to subjective analyses. By contrast, we propose a wholly data-driven approach to removing unwanted variation, harnessing the information contained in control samples, pre-treatment groups, or other signal-free observations, all while enhancing the interpretability and stability of findings by inducing sparsity.

The remainder of the present manuscript is organized as follows. In Section 1.2, contrastive PCA, sparse PCA, and other popular dimensionality reduction techniques are briefly surveyed. Next, in Section 2, scPCA is formally defined and its desirable properties are detailed. A simulation study and several analyses of publicly available microarray gene expression, and single-cell transcriptome sequencing (scRNA-seq) data are presented in Section 3, and the analysis of a protein expression dataset is detailed in Section S6, providing a rich comparison of the proposed methodology to other popular techniques currently relied upon for the exploration of high-dimensional biological data. Finally, we conclude by reviewing the effectiveness of scPCA based on the results of these experiments and discussing paths for further investigation.

1.2 Background

1.2.1 Contrastive PCA

The development of contrastive PCA was motivated by the need to detect and visualize variation in the data deemed most relevant to the scientific question of interest. Given a target dataset believed to contain biological signal(s) of interest and a similar background dataset believed to comprise only noise (i.e., unwanted variation), the cPCA algorithm returns a subspace of the target data that contains (a portion of) the variation absent from the background data [1]. cPCA aims to identify the pertinent variation in the target data by *contrasting* the covariance matrix of its features with that of the background data. For example, consider

a scRNA-seq dataset whose samples are contaminated by a batch effect. Provided a collection of control samples subjected to the same batch effect, cPCA may be used to remove this unwanted technical noise (see Section 3.1).

Algorithmically, cPCA is very similar to PCA. Consider a column-centered target dataset $\mathbf{X}_{n\times p}$ and a column-centered background dataset $\mathbf{Y}_{m\times p}$, where n and m denote, respectively, the number of target and background observations (e.g., cells) and p denotes the number of features (e.g., genes). Define their empirical covariance matrices as $\mathbf{C_X}$ and $\mathbf{C_Y}$, and let $\mathbb{R}^p_{\text{unit}} = \{\mathbf{v} \in \mathbb{R}^p : \|\mathbf{v}\|_2 = 1\}$ be the set of unit vectors of length p. The variances along direction $\mathbf{v} \in \mathbb{R}^p_{\text{unit}}$ in the target and background datasets are represented by $\lambda_{\mathbf{X}}(\mathbf{v}) = \mathbf{v}^\top \mathbf{C_X} \mathbf{v}$ and $\lambda_{\mathbf{Y}}(\mathbf{v}) = \mathbf{v}^\top \mathbf{C_Y} \mathbf{v}$, respectively. The most contrastive direction $\mathbf{v}^\star_\gamma$ for some fixed contrastive parameter $\gamma \in \mathbb{R}^+$ is found by solving

$$\begin{aligned} \mathbf{v}^\star_\gamma &= \underset{\mathbf{v}\in\mathbb{R}^p_{\text{unit}}}{\text{argmax}}\, \lambda_{\mathbf{X}}(\mathbf{v}) - \gamma\lambda_{\mathbf{Y}}(\mathbf{v}) \\ &= \underset{\mathbf{v}\in\mathbb{R}^p_{\text{unit}}}{\text{argmax}}\, \mathbf{v}^\top(\mathbf{C_X} - \gamma\mathbf{C_Y})\mathbf{v} \\ &= \underset{\mathbf{v}\in\mathbb{R}^p_{\text{unit}}}{\text{argmax}}\, \mathbf{v}^\top \mathbf{C}_\gamma \mathbf{v}, \end{aligned} \tag{1}$$

where $\mathbf{C}_\gamma = \mathbf{C_X} - \gamma\mathbf{C_Y}$ is the contrastive covariance matrix [1]. cPCA can therefore be performed by computing the eigenvalue decomposition of $\mathbf{C}_\gamma$. The eigenvectors of $\mathbf{C}_\gamma$ are then used to map the target data to the *contrastive* principal components (cPCs).

The contrastive parameter γ quantifies the trade-off between each feature's variances in the target and background datasets. When $\gamma = 0$, cPCA reduces to PCA — hence, the variance along $\lambda_{\mathbf{X}}(\mathbf{v})$ is maximized. On the other hand, as $\gamma \to \infty$, the variance in the background data dominates the variance in the target data such that only directions spanned by the background dataset are captured. This is akin to projecting the target dataset into the space spanned by directions in the background data and then performing PCA on this projection [1]. The effect of the contrastive parameter is illustrated in fig. S1 for simulated data similar to those used by Abid *et al.* [1].

Although cPCA offers a novel approach for the removal of unwanted variation, it possesses some drawbacks. In particular, no rigorous framework exists for selecting the contrastive parameter γ in order to achieve the optimal amount of contrast between the target and background data. Indeed, Abid *et al.* [1]'s approach to selecting an appropriate γ relies on visual inspection. Additionally, as with PCA, loading vectors may be highly variable and difficult to interpret in high dimensions since they represent linear combinations of all variables in the dataset. Relatedly, cPCs are not certifiably free of unwanted technical and biological effects, potentially obscuring relevant biological signal. This issue is only exacerbated as the dimension of the subspace orthogonal to the background data increases, jeopardizing the stability of the cPCs and enfeebling conclusions drawn from them.

1.2.2 Sparse PCA

In addition to being difficult to interpret, the PCs generated by applying PCA to high-dimensional data are generally unstable; that is they are subject to major changes under minor perturbations of the data (we refer to Johnstone and Paul [15] for a recent review). Luckily, an abundance of techniques for sparsifying PCA loadings have been developed to mitigate these issues; we direct the interested reader to Zou and Xue [36] for a recent review. Here, we consider the SPCA technique developed by Zou *et al.* [37]. In contrast to standard PCA, SPCA generates interpretable and stable loadings in high dimensions, with most entries of the matrix being zero.

SPCA was born from the geometric interpretation of PCA, which reframes PCA as a regression problem. Given a matrix whose columns form an orthonormal basis $\mathbf{V}_{p\times k}$, the objective is to find the projection $\mathbf{P}_k = \mathbf{V}_{p\times k}\mathbf{V}^\top_{p\times k}$ producing the best linear manifold approximation of the data $\mathbf{X}_{n\times p}$. This is accomplished by minimizing the mean squared error:

$$\mathbf{V}^\star_{p\times k} = \underset{\mathbf{V}_{p\times k}}{\text{argmin}} \sum_{i=1}^{n} \|x_i - \mathbf{V}_{p\times k}\mathbf{V}^\top_{p\times k} x_i\|_2^2, \tag{2}$$

where x_i is the i^{th} row of $\mathbf{X}$ and $\mathbf{V}^\star_{p\times k}$ is exactly the loadings matrix of the first k PCs [36]. A sparse loadings matrix can be obtained by imposing an elastic net constraint on a modification of this objective function.

Zou *et al.* [37] show that optimizing the following criterion provides loadings of the first k sparse PCs of $\mathbf{X}$:

$$(\mathbf{A}^{\star}_{p\times k}, \mathbf{B}^{\star}_{p\times k}) = \underset{\mathbf{A}_{p\times k}, \mathbf{B}_{p\times k}}{\operatorname{argmin}} \sum_{i=1}^{n} \|x_i - \mathbf{A}_{p\times k}\mathbf{B}^{\top}_{p\times k} x_i\|_2^2 + \lambda_0 \sum_{j=1}^{k} \beta_j^2 + \sum_{j=1}^{k} \lambda_{1,j}|\beta_j| \quad \text{subject to } \mathbf{A}^{\top}_{p\times k}\mathbf{A}_{p\times k} = \mathbf{I}_{k\times k}, \tag{3}$$

where β_j is the j^{th} column of $\mathbf{B}_{p\times k}$ and whereλ_0 and $\lambda_{1,j}$ are, respectively, the ℓ_2 and ℓ_1 penalty parameters for the non-normalized j^{th} loading β_j; as in the original SPCA manuscript the ℓ_1 penalty is allowed to be loading-specific [37]. Then, the sparse loadings are the normalized versions of $\beta_j^{\star}$, i.e., the vectors $\beta_j^{\star}/\|\beta_j^{\star}\|_2$. Zou *et al.* [37] also show that the full dataset need not be used to optimize the criterion; indeed, only the Gram matrix $\mathbf{X}^{\top}_{n\times p}\mathbf{X}_{n\times p}$ is required [37].

Although SPCA provides a transparent and efficient method for the sparsification of PCA's loading matrices, and hence the generation of stable principal components in high dimensions, its development stopped short of providing means by which to identify the most relevant directions of variation in the data, presenting an obstacle to its efficacious use in biological data exploration and analysis. This motivates the development of exploratory methods that build upon the strengths of both SPCA and cPCA.

1.2.3 Other Competing Methods

Other general methods frequently employed to reduce the dimensionality of high-dimensional biological data include t-distributed stochastic neighbor embedding (t-SNE) [29] and uniform manifold approximation and projection (UMAP) [22] (e.g., [2, 3]). Unlike PCA, SPCA, and cPCA, both are nonlinear dimensionality reduction techniques — that is, they do not enforce linear relationships between features. Such a relaxation permits the capturing of local nonlinear structures in the data that would otherwise go unnoticed, though neither approach guarantees that their low-dimensional embeddings reflect the global structure of the data. Becht *et al.* [3] demonstrated the extreme computational efficiency exhibited by these techniques in their application to large datasets while Amir *et al.* [2] and Becht *et al.* [3] illustrated the stability of their findings, further increasing their popularity as methods of choice for EDA in computational biology. Yet, the flexibility and speed of t-SNE and UMAP come at a cost: these techniques are not endowed with the ease of interpretability of factor analysis methods. In lacking an interpretable link between the data's features and low-dimensional representation, their use as hypothesis-generating tools is restricted. Furthermore, like PCA, neither t-SNE nor UMAP have the ability to explicitly remove unwanted technical effects.

Though the dimensionality reduction methods discussed thus far can be applied to various kinds of high-dimensional biological data, still many others have been developed expressly for use with specific high-throughput assay biotechnologies. One such method, ZINB-WaVE, relies on a zero-inflated negative binomial model to better account for the count nature, zero inflation, and over-dispersion of scRNA-seq data, and has been shown to outperform less tailored techniques such as t-SNE [27]. Unlike more general factor analysis methods (e.g., PCA), ZINB-WaVE takes advantage of the rich annotation metadata that are often available with scRNA-seq datasets to remove sources of unwanted variation, while preserving global biological signal [27]. Analogous to PCA, the latent factors produced by ZINB-WaVE are not sparse. Technical and biological noise may remain after taking into account known and unknown sources of unwanted variation [27], potentially blurring any meaningful interpretation of latent factors. Other successful methods for reducing the dimensionality of scRNA-seq data, such as scVI [20], have relied on the variational autoencoder framework to learn nonlinear structures in the data at hand and thereby infer values of latent variables. Further discussion of such techniques lies outside the scope of the present work on account of the dissimilarity to methods inspired by factor analysis, like SPCA, cPCA, and ZINB-WaVE, which are our focus.

2 Methodology

2.1 Sparse Contrastive PCA

Given a pair of target and background datasets as defined in Section 1.2.1, the scPCA procedure applies SPCA with minimal modifications to their contrastive covariance matrix $\mathbf{C}_{\gamma}$. The numerical solution to the SPCA criterion of Equation (3) is obtained by the following alternating algorithm until convergence of the sparse loadings [37]:

For fixed A: For each j, let $Y_j \coloneqq \mathbf{C}_\gamma^{\frac{1}{2}} \alpha_j$, where α_j is the j^{th} column of $\mathbf{A}$. Then, the elastic net solution for the j^{th} loading vector is

$$\beta_j^\star = \underset{\beta_j}{\operatorname{argmin}} \|Y_j - \mathbf{C}_\gamma^{\frac{1}{2}} \beta_j\|_2^2 + \lambda_0 \|\beta_j\|_2^2 + \lambda_{1,j} \|\beta_j\|_1.$$

Generally, for ease of computation, $\lambda_{1,j} = \lambda_1$, for $j = 1, \ldots, k$. The entries of the loadings matrix $\mathbf{B}$ are independent of the choice for the ℓ_2 penalty (ridge) parameter λ_0 [37], existing only to ensure the reconstruction of the sparse principal components. The ridge penalty is set to zero when $\mathbf{C}_\gamma^{\frac{1}{2}}$ is full rank; otherwise, a small constant value is used to remedy issues of indeterminacy that arise when fitting the elastic net. In fact, this was the original motivation behind the development of ridge regression [11].

For fixed B: Only the first term of the SPCA criterion of Equation (3) must be minimized with respect to $\mathbf{A}$. The solution is given by the reduced rank form of the Procrustes rotation, computed as $\mathbf{A}^\star = \mathbf{U}\mathbf{V}^\top$ [37]. The matrices of left and right singular vectors are obtained from the following singular value decomposition:

$$\mathbf{C}_\gamma \mathbf{B} = \mathbf{U}\mathbf{D}\mathbf{V}^\top.$$

Generally, $\mathbf{C}_\gamma$ is not positive-semidefinite and its square root is undefined. Instead, a positive-semidefinite matrix $\widetilde{\mathbf{C}}_\gamma$, approximating $\mathbf{C}_\gamma$, is used. $\widetilde{\mathbf{C}}_\gamma$ is obtained by replacing the diagonal matrix in the eigendecomposition of $\mathbf{C}_\gamma$ by a diagonal matrix in which negative eigenvalues are replaced by zeros [35]:

$$\mathbf{C}_\gamma = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^\top$$
$$\widetilde{\mathbf{C}}_\gamma = \mathbf{V}\mathbf{D}\mathbf{V}^\top$$

$$\text{where } \mathrm{D}_{ii} = \begin{cases} \Lambda_{ii}, & \text{if } \Lambda_{ii} > 0 \\ 0, & \text{otherwise} \end{cases}, \qquad \text{for } i = 1, \ldots, p.$$

Thus, the directions of variation given by the negative eigenvalues of $\mathbf{C}_\gamma$ are discarded, as they correspond to those which are dominated by the variance in the background dataset. This procedure can be viewed as a preliminary thresholding of the eigenvectors of $\mathbf{C}_\gamma$, where the cutoff is an additional hyperparameter corresponding to a non-negative real number. Explicitly defining a small positive threshold may prove useful for datasets that possess many eigenvalues near zero, which correspond to sources of technical and biological noise remaining after the contrastive step. Empirically, however, providing a wide range of contrastive parameters γ in the hyperparameter space has been found to have a similar effect as using multiple cutoff values — that is, larger values of γ naturally produce sparser matrices $\widetilde{\mathbf{C}}_\gamma$.

For the purpose of contrastive analysis, a direction's importance is characterized by its target-background variance coupling; higher target variance and lower background variance pairs produce the best directions [1] and correspond to the largest positive eigenvalues. The elimination of directions with negative eigenvalues therefore guarantees that the sparse contrastive PCs (scPCs) are rotations of the target data relying on the sparse directions most variable in the target data but least variable in the background data, making a cutoff of zero a natural choice for the thresholding operation.

2.2 Framework for Hyperparameter Tuning

The scPCA algorithm relies on two hyperparameters: the contrastive parameter γ and the ℓ_1 penalty parameter λ_1. To select the optimal combination of γ and λ_1 from a grid of *a priori* specified values, we propose to cluster the n observations of the target dataset based on their first k scPCs, selecting as optimal the combination $\{\gamma, \lambda_1\}$ producing the "strongest" cluster assignments. This framework casts the selection of $\{\gamma, \lambda_1\}$ in terms of a choice of clustering algorithm, distance metric (based on $\widetilde{\mathbf{C}}_\gamma$), and clustering strength criterion. For ease of application, we propose to select $\{\gamma, \lambda_1\}$ by maximization of the average silhouette width over clusterings of the reduced-dimension representation of the target data. This procedure implicitly requires the choice of a clustering algorithm, such as k-means [19] or partitioning around medoids [16], to be applied to the representation of the data in the first k scPCs. Such methods require an appropriate choice for the number of clusters, which we contend will generally not be a limiting factor in the use of scPCA. Indeed, reasonable choices for the number of clusters can often be inferred in *omics* settings from sample annotation variables accompanying the data or from previously available biological knowledge. In Section 3, we empirically demonstrate that the results of the algorithm are robust to the choice of the number

of clusters. Additionally, scPCA has no particular dependence on average silhouette width as a criterion — that is, alternative criteria for assessing clustering strength could be used when appropriate. Naturally, this proposed hyperparameter tuning approach can be applied to cPCA by setting λ_1 to 0.

To address concerns of overfitting and to avoid discovering non-generalizable patterns from the data, we propose the use of cross-validation. For a grid of *a priori* specified contrastive parameters γ and ℓ_1 penalty parameters λ_1, V-fold cross-validation may be performed as follows:

1. Partition each of the target and background datasets into V roughly equally-sized subsets.
2. Randomly pair each of the target dataset's V subsets with one of the background's; these pairs form the fold-specific data for cross-validation.
3. Iteratively perform scPCA over the observations of the target and background data not contained in the holdout set (i.e., the training set) for each pair of contrastive parameters and ℓ_1 penalty parameters in the hyperparameter grid.
4. Project the holdout target data onto the low-dimensional space using the loadings matrices obtained from the previous step.
5. Compute a clustering strength criterion (e.g., average silhouette width) for a clustering of the target holdout data with the *a priori* specified number of clusters.
6. Finally, compute the cross-validated average of the clustering strength criteria (e.g., cross-validated average of average silhouette width) across the holdout sets for each pair of hyperparameters, selecting the pairing that maximizes the value of the criterion.

2.3 Algorithm and Software Implementation

The implementation of the scPCA algorithm is summarized in Algorithm 1. A free and open-source software implementation of scPCA is available in the `scPCA` package for the `R` language and environment for statistical computing [24]. The `scPCA` package is available as of a recent release of the Bioconductor Project [10, 9, 13] (`https://bioconductor.org/packages/scPCA`).

For ease of notation, Algorithm 1 introduces scPCA without the application of cross-validation to the target and background datasets. Algorithm 1, in the supplementary materials, details the cross-validated variant.

The code and data used to generate this manuscript are publicly available on GitHub (`https://github.com/PhilBoileau/EHDBDscPCA`).

Algorithm 1: scPCA

Result: Produces a sparse low-dimensional representation of the target data, $\mathbf{X}_{n\times p}$, by contrasting the variation of $\mathbf{X}_{n\times p}$ and some background data, $\mathbf{Y}_{m\times p}$, while applying an ℓ_1 penalty to the loadings generated by cPCA.

Input :

target dataset: $\mathbf{X}$

background dataset: $\mathbf{Y}$

binary variable indicating whether to column-scale the data: `scale`

vector of possible contrastive parameters: $\gamma = (\gamma_1, \ldots, \gamma_s)$

vector of possible ℓ_1 penalty parameters: $\lambda_1 = (\lambda_{1,1}, \ldots, \lambda_{1,d})$

number of sparse contrastive principal components to compute: k

clustering method: `cluster_meth`

number of clusters: `ncluster`

Center (and `scale` if so desired) the columns of $\mathbf{X}$, $\mathbf{Y}$
Calculate the empirical covariance matrices: $\mathbf{C}_{\mathbf{X}p\times p} \coloneqq \frac{1}{n}\mathbf{X}^\top\mathbf{X}$, $\mathbf{C}_{\mathbf{Y}p\times p} \coloneqq \frac{1}{m}\mathbf{Y}^\top\mathbf{Y}$
for *each* $\gamma_i \in \gamma$ **do**
  for *each* $\lambda_{1,j} \in \lambda_1$ **do**
    Compute the contrastive covariance matrix $\mathbf{C}_{\gamma_i} = \mathbf{C}_{\mathbf{X}} - \gamma_i\mathbf{C}_{\mathbf{Y}}$
    Compute the positive-semidefinite approximation of $\mathbf{C}_{\gamma_i}$, $\widetilde{\mathbf{C}}_{\gamma_i}$
    Apply SPCA to $\widetilde{\mathbf{C}}_{\gamma_i}$ for k components with ℓ_1 penalty $\lambda_{1,j}$
    Generate a low-dimensional representation by projecting $\mathbf{X}_{n\times p}$ on the sparse loadings of SPCA
    Normalize the low-dimensional representation produced to be on the unit hypercube
    Cluster the normalized low-dimensional representation using `cluster_meth` with `ncluster`
    Compute and record the clustering strength criterion associated with $(\gamma_i, \lambda_{1,j})$
Identify the combination of hyperparameters maximizing the clustering strength criterion: $\gamma^\star$, $\lambda_1^\star$

Output: The low-dimensional representation of the target data given by $(\gamma^\star, \lambda_1^\star)$, an $n \times k$ matrix; the $p \times k$ matrix of loadings given by $(\gamma^\star, \lambda_1^\star)$; contrastive parameter $\gamma^\star$; ℓ_1 penalty parameter $\lambda_1^\star$

3 Results

In the sequel, we detail the application of scPCA to a number of simulated and publicly available datasets, comparing our proposal to several competing techniques. An additional analysis of protein expression data is presented in Section S6.

3.1 Simulated scRNA-seq Data

The scPCA technique was tested on a simulated scRNA-seq dataset generated with the *Splat* framework from the `Splatter R` package [31]. *Splat* simulates a scRNA-seq count matrix by way of a gamma-Poisson hierarchical model. This simulation framework mimics real scRNA-seq data by including hyperparameters to control the number of over- and under-expressed genes (using multiplicative factors for mean expression levels), zero inflation, batch effects, and other technical and biological factors unique to scRNA-seq data.

A simple dataset of 300 cells and 500 genes was simulated such that the cells were approximately evenly distributed among three biological groups: two groups making up a target dataset, and a third group corresponding to a background dataset. 5% of the genes are differentially expressed between the background dataset and each of the two target datasets but not between the two target datasets, 10% of the genes are differentially expressed between the background dataset and the first target dataset but not the second target dataset, and 10% of the genes are differentially expressed between the background dataset and the second target dataset but not the first target dataset. There is overlap between these three sets of genes and, in particular, a total of 98 genes are differentially expressed between the two target datasets. Based on these levels of differential expression, cells are more dissimilar between the two target datasets than between either of the target datasets and the background dataset. Therefore, the samples comprising the background dataset can be viewed as a set of controls for use by cPCA and scPCA. Additionally, a large batch effect was

included to confound the biological variation between groups, effectively dividing each biological group into two subgroups of equal size (fig. 1A).

PCA, t-SNE, UMAP, cPCA, and scPCA were applied to the log-transformed and column-centered target data to determine whether these methods could identify the biological signal of interest, i.e., the two groups in the target dataset (fig. 1B, n.b., PCA was not included due to the similarity of results to fig. 1A). We note that cPCA was not performed in the traditional manner of Abid *et al.* [1], but with automatic hyperparameter selection as described in Section 2.2. The number of *a priori* specified clusters for the cPCA and scPCA methods was set to 2, and the column-centered background data were used in their contrastive steps. While PCA, t-SNE, UMAP, and cPCA were incapable of completely eliminating the batch effect in their two-dimensional representations, scPCA successfully removed the unwanted variation while producing the tightest clusters, as indicated by the average silhouette widths (see also fig. S3), and generating sparse, interpretable loadings.

To compare the loadings produced by cPCA and scPCA, each of their loadings' vectors were standardized as follows. The i^{th} entry of the j^{th} standardized loadings vector is given by: $\frac{|V_{ij}|-\min_i|V_{ij}|}{\max_i|V_{ij}|-\min_i|V_{ij}|}$, where $\mathbf{V}$ is a $p \times k$ loadings matrix. Juxtaposing the relative absolute weights of the first loadings vectors produced by cPCA and scPCA, each of which linearly separate the target dataset's groups, we find scPCA's to be, as expected, much sparser (see fig. 1C). In fact, only 20 genes have non-zero values in scPCA's first loading compared to 500 in cPCA; moreover, these 20 genes correspond to those which have the largest absolute entries in cPCA's first loading vector. Furthermore, these genes are among the most differentially expressed in the target dataset, based on the values of their multiplicative differential expression factors (fig. S2).

scPCA's results were also compared to those of the two leading latent factors found by ZINB-WaVE, a method of choice for dimensionality reduction for scRNA-seq data, under conditions in which the batch factor is viewed as known and unknown (fig. 1B). In both cases, ZINB-WaVE was applied to the count matrix of the simulated target dataset with no gene-level covariates. When the batch factor was treated as unknown, no cell-level covariates were included in the model; however, when we treated the batch factor as known, a binary cell-level covariate was added to indicate each sample's batch membership. When the source of unwanted variation is not explicitly regressed out in the ZINB-WaVE model, we find the results to be virtually identical to those of PCA. Even when the batch effect is included in the model, the clusters of the biological groups are elongated and less dense than those produced by scPCA, and the first latent factor does not linearly separate the groups.

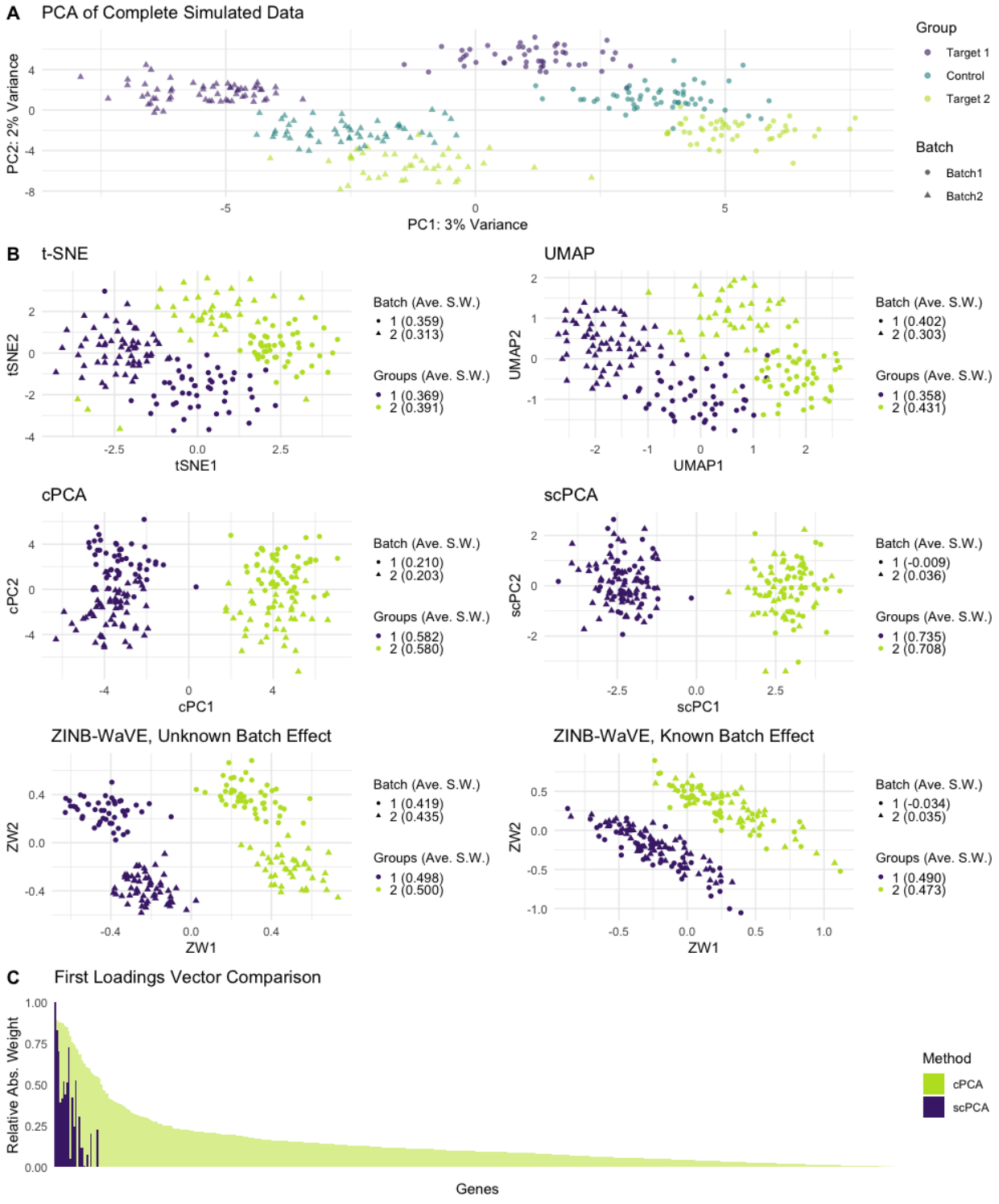

Figure 1: *Simulated scRNA-seq data.* **A** Plot of the first two principal components of the complete simulated dataset (i.e., the combination of the target and background datasets). The batch effect and the biological signal are responsible for approximately identical amounts of variance. **B** The two-dimensional representations of the target dataset by t-SNE, UMAP, cPCA, scPCA, and ZINB-WaVE, with accompanying average silhouette widths quantifying the strengths of the batch effect and the biological signal. Only scPCA fully removes the batch effect in two dimensions when batches are not adjusted for explicitly. **C** A gene-by-gene comparison of the relative absolute weights in the first loading vectors of cPCA and scPCA, in decreasing order with respect to the values produced by cPCA.

3.2 Dengue Microarray Data

Kwissa *et al.* [17] used gene expression microarrays to analyze the whole-blood transcriptome of 47 dengue patients hospitalized at the Siriraj Hospital in Bangkok and 9 local healthy controls. Of the affected patients, 18 were classified as having acute dengue fever (DF), 10 as having acute dengue hemorrhagic fever (DHF), and 19 as convalescent at least four weeks after discharge.

As part of data pre-processing, all but the 500 most variable genes were filtered out. The target dataset consists of the log-transformed microarray expression measures of 47 patients with some form of dengue, while the background dataset consist of the log-transformed microarray expression measures of the control samples. PCA, cPCA, scPCA, t-SNE, and UMAP were then applied to the column-centered matrix of the target data with the goal of discerning three unique clusters (fig. 2A), one for each sub-class of dengue (DF, DHF, and convalescent). cPCA and scPCA took as additional input the column-centered background dataset and specified three clusters *a priori*. t-SNE's embedding was found to be similar to UMAP's and is therefore only included in the supplementary materials (fig. S4).

Of the four dimensionality reduction methods, only cPCA and scPCA successfully fully separated the convalescent patients from those with DF and DHF in two dimensions. scPCA's low-dimensional representation was virtually identical to that of cPCA, producing very similar average silhouette widths among classes, though only a tenth of the genes have non-zero values in the first and second columns of the loadings matrix, and the most important genes identified by each methods' first loading differ substantially (fig. 2B). The genes found by scPCA include CD38, HLA-DQB1, and RSAD2 (Viperin), which have been previously associated to the susceptibility to, protection against, or presence of dengue [5, 4, 6]. For a full list of these genes, refer to Table S1 and Table S2.

No method successfully distinguished between the three sub-classes of dengue. In fact, previous research suggests that the transcriptome of patients with DF and DHF are virtually indistinct [17]. Instead, Kwissa *et al.* [17] found that DF and DHF patients may form distinct clusters based on viral load and concentration of the DENV NS-1 antigen in their plasma. This may explain the sub-clusters within the DF and DHF cases found by UMAP. Though the number of pre-specified clusters for each algorithm was set to three, cPCA's and scPCA's projections onto two dimensions contain two clusters. To test the sensitivity of these methods to this tuning parameter, both methods were reapplied to the data with varying numbers of pre-specified clusters. Each of cPCA's iteration produced virtually identical embeddings (fig. S5). However, scPCA's produced identical results to those of PCA when the number of clusters was set to four or higher (fig. S6). This may provide an empirical approach to selecting the appropriate number of clusters for scPCA, i.e., selecting the largest value before which the quality of the embedding deteriorates.

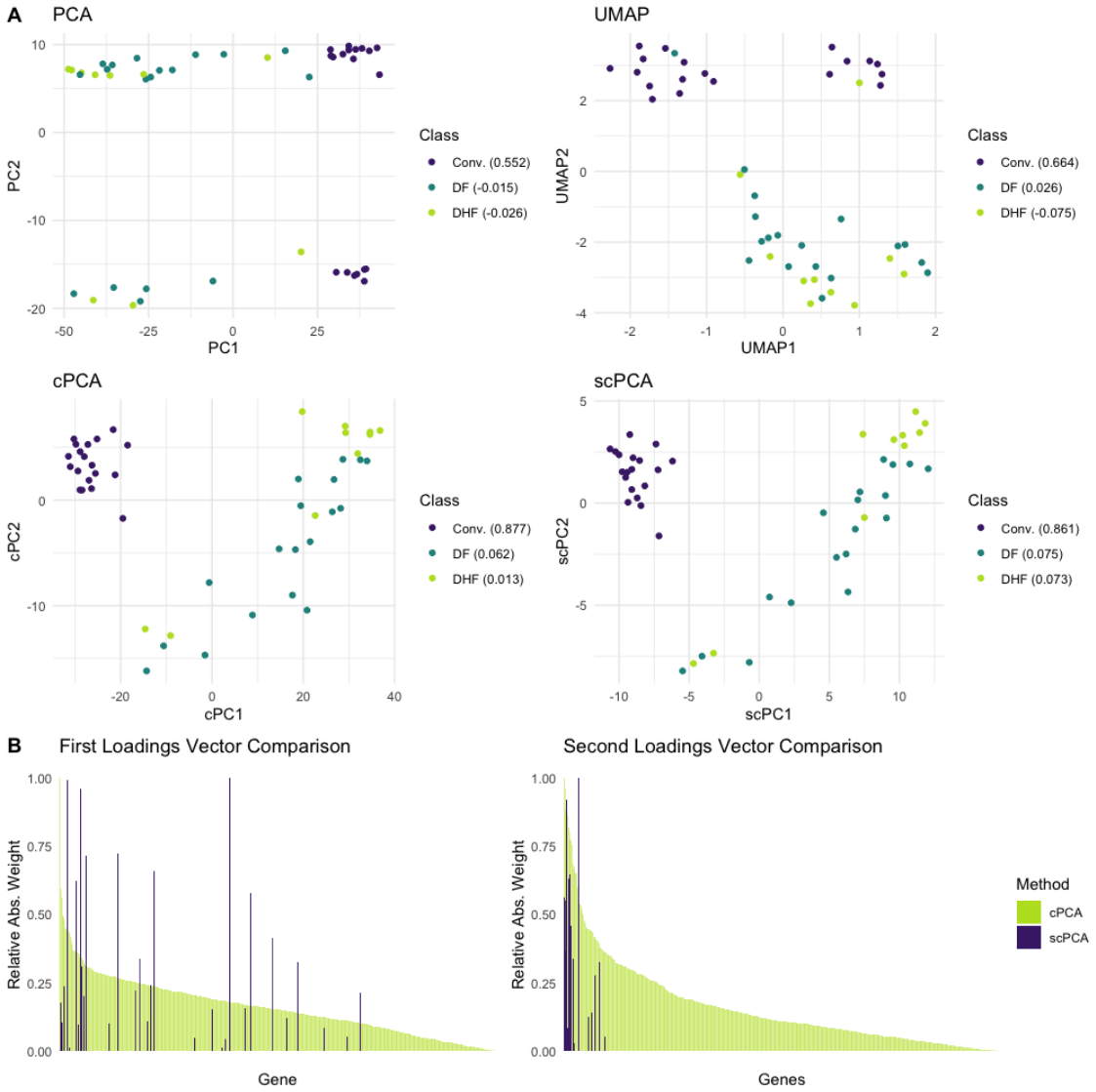

Figure 2: *Dengue microarray data.* **A** Two-dimensional representations of the target dataset by PCA, UMAP, cPCA, and scPCA, with accompanying average silhouette widths quantifying the strengths of the biological signal. cPCA and scPCA are the only methods that fully separate the convalescent patients from those with DF and DHF. The second PC of the PCA plot is dominated by some batch effect, and the low-dimensional representation produced by UMAP also appears to be affected by some source of unwanted variation. **B** The relative absolute weights in the two leading loading vectors of scPCA are much sparser than those of cPCA, though their two-dimensional embeddings are virtually identical. The genes are in decreasing order of cPCA's relative absolute weights, demonstrating that the genes with non-zero weights in scPCA's loadings generally correspond to the those genes with the largest absolute weights in cPCA's loadings. This is much more apparent in the second loadings vector where the distribution of cPCA's absolute weights has a thin tail, attributing increased importance to a small subset of genes.

3.3 Leukemia Patient scRNA-seq Data

Finally, we tested scPCA on scRNA-seq data from the cryopreserved bone marrow mononuclear cell (BMMC) samples of two acute myeloid leukemia (AML) patients (Patient 035: 4,501 cells; Patient 027: 7,898 cells), before and after undergoing allogeneic hematopoietic stem cell transplant treatment [34]. The BMMCs of two healthy individuals from the same publicly available dataset (Healthy 1: 1,985 cells; Healthy 2: 2,472 cells) were used to generate a control dataset. Following pre-processing, all but the 1,000 most variable genes measured across all 16,856 cells were removed. The scRNA-seq data from the AML patients were then split into separate target datasets since Zheng *et al.* [34] found evidence of distinct subpopulation membership following transplantation. Data belonging to the healthy controls were combined to create the background dataset. PCA, t-SNE, UMAP, ZINB-WaVE, cPCA, and scPCA were applied to the target datasets to explore differences in the AML patients' BMMCs's transcriptome engendered by the treatment (fig. 3, S7).

Of the six dimensionality reduction methods applied to Patient 035's data (fig. 3A), cPCA and scPCA best capture the biologically meaningful information relating to treatment status. Each produces linearly separable clusters corresponding to pre- and post-treatment cells; scPCA's projection yields a tighter cluster of pre-transplant cells when compared to that produced by cPCA, and the opposite is true regarding the clusters of post-transplant cells. Additionally, scPCA's projection required considerably less information even though its results are analogous to cPCA's: 176 genes and 17 genes have non-zero entries in the first and second columns of the loadings matrix produced by scPCA, respectively (fig. 3B). In general, the leading loadings of cPCA and scPCA place an increased importance on the same genes. Regarding the other methods' results, PCA and t-SNE fail to separate the pre- and post-transplant cells, and UMAP's and ZINB-WaVE's embeddings resemble a trajectory more closely than they do a set of clusters.

Similarly to Patient 035's results, the two-dimensional embeddings of Patient 027's data produced by PCA, t-SNE, UMAP, and ZINB-WaVE do not contain distinct clusters of pre- and post-transplant BMMCs (fig. S7); however, cPCA and scPCA generate low-dimensional representations of the data in which samples are clustered based on treatment status. Although cPCA's representation produces denser groupings, the first two columns of scPCA's loading matrix contain non-zero values in only three genes, STMN1, LINC00152, and PDLIM1, all of which have been previously linked to leukemia [21, 32, 12].

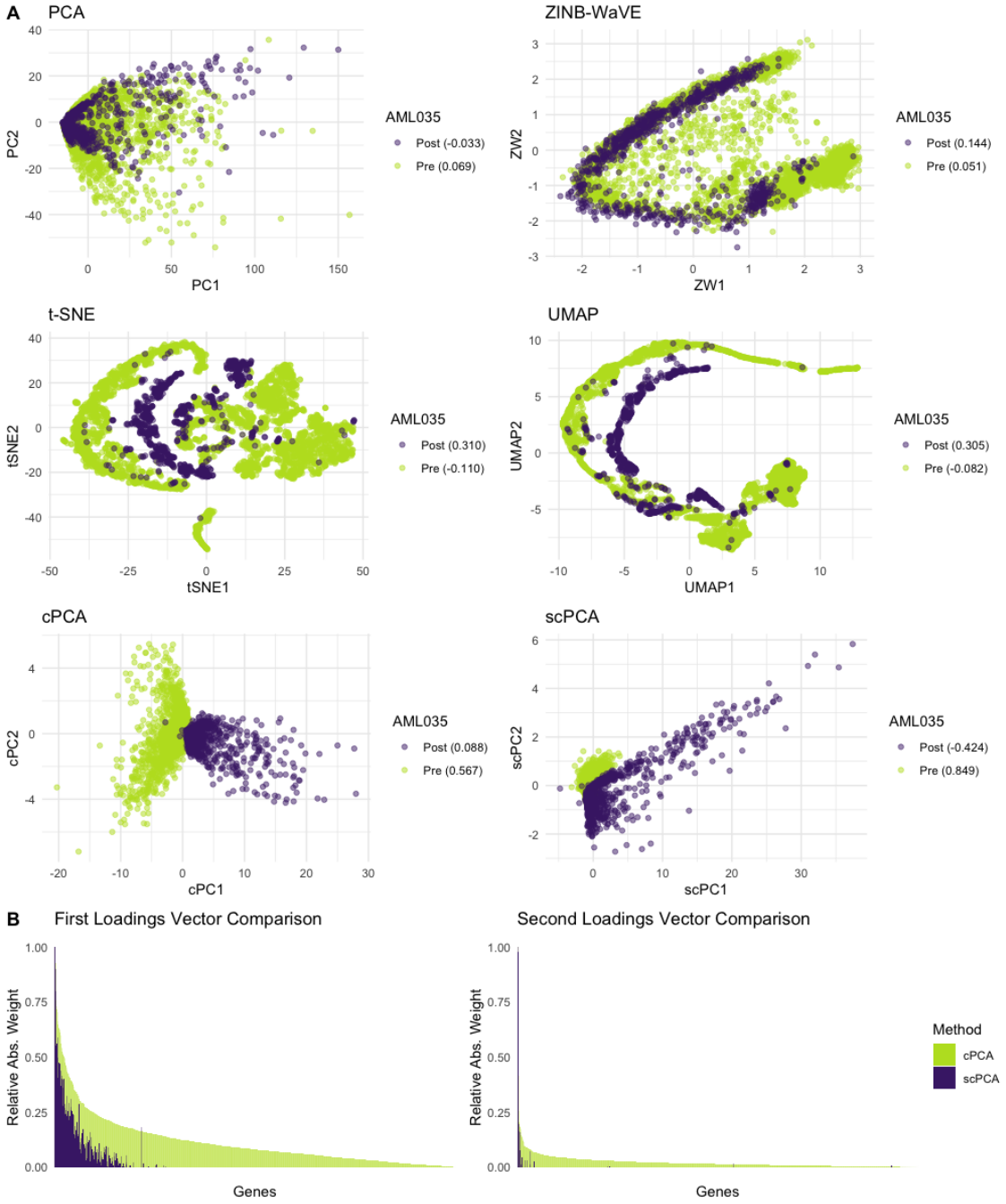

Figure 3: *AML Patient 035 scRNA-seq data.* **A** The two-dimensional embeddings of the patient's BMMC cells produced by PCA, ZINB-WaVE, t-SNE, UMAP, cPCA, and scPCA, with accompanying average silhouette widths quantifying the strengths of the biological signal. cPCA and scPCA produce representations of the data in which the pre- and post-transplant cells form discernible clusters. Based on visual inspection and average silhouette width, scPCA's grouping of pre-transplant cells is denser than that of cPCA's and the opposite is true of the post-transplant cells' cluster. **B** scPCA's embedding is much sparser, increasing interpretability of the exploratory analysis.

4 Discussion

We have proposed a novel dimensionality reduction technique for use with high-dimensional biological data: *sparse contrastive principal component analysis*. A central contribution of the proposed method is the incorporation of a penalization step that ensures both the sparsity and stability of the principal components generated by contrastive PCA. Our approach allows for high-dimensional biological datasets, such as those produced by the currently popular scRNA-seq experimental paradigm, to be examined in a manner that uncovers interpretable as well as relevant biological signal after the removal of unwanted technical variation, all the while placing only minimal assumptions on the data-generating process.

We also present a data-adaptive and algorithmic framework for applying contrastive dimensionality reduction techniques, like cPCA and scPCA, to high-dimensional biological data. Where the original proposal of cPCA relied upon visual inspection by the user in selecting the "best" contrastive parameter [1], a notoriously unreliable process, our extension formalizes the data-adaptive selection of tuning parameters. The automation of this step translates directly to significantly increased computational reproducibility. We have proposed the use of cross-validation to select tuning parameters from among a pre-specified set in a generalizable manner, using average silhouette width to assess clustering strength. Several other approaches to the selection of tuning parameters, including the choice of criterion for assessing the "goodness" of the dimensionality reduction (here, clustering strength as measured by the average silhouette width), may outperform our approach in practice and could be incorporated into the modular framework of scPCA — we leave the development of such approaches and assessments of their potential advantages as an avenue for future investigation.

We have demonstrated the utility of scPCA relative to competing approaches, including standard PCA, cPCA, t-SNE, UMAP, and ZINB-WaVE (where appropriate), using both a simulation study of single-cell RNA-seq data and the re-analysis of several publicly available datasets from a variety of high-dimensional biological assays. We have shown that scPCA recovers low-dimensional embeddings similar to cPCA, but with a more easily interpretable principal component structure, and in simulations, diminished technical noise. Further, our results indicate that scPCA generally produces denser, more relevant clusters than t-SNE, UMAP, and ZINB-WaVE. Moreover, we verify that clusters derived from scPCA correspond to biological signal of interest. Finally, as the cost of producing high-dimensional biological data with high-throughput experiments continues to decrease, we expect that the availability and utility of techniques like scPCA — for reliably extracting rich, sparse biological signals while data-adaptively removing technical artifacts — will strongly motivate the collection of control samples as a part of standard practice.

Funding

This work was supported by the Fonds de recherche du Québec - Nature et technologies [B1X to PB].

Acknowledgements

We thank Mark van der Laan and Hector Roux de Bézieux for helpful discussions and insights.

Conflicts of interest: None declared.

References

[1] Abid, A., Zhang, M. J., Bagaria, V. K., and Zou, J. (2018). Exploring patterns enriched in a dataset with contrastive principal component analysis. *Nature Communications*, **9**(1), 2134.

[2] Amir, E. A. D., Davis, K. L., Tadmor, M. D., Simonds, E. F., Levine, J. H., Bendall, S. C., Shenfeld, D. K., Krishnaswamy, S., Nolan, G. P., and Pe'Er, D. (2013). ViSNE enables visualization of high dimensional single-cell data and reveals phenotypic heterogeneity of leukemia. *Nature Biotechnology*, **31**(6), 545–552.

[3] Becht, E., McInnes, L., Healy, J., Dutertre, C. A., Kwok, I. W., Ng, L. G., Ginhoux, F., and Newell, E. W. (2019). Dimensionality reduction for visualizing single-cell data using UMAP. *Nature Biotechnology*, **37**(1), 38–47.

[4] Cardozo, D. M., Moliterno, R. A., Sell, A. M., Guelsin, G. A. S., Beltrame, L. M., Clementino, S. L., Reis, P. G., Alves, H. V., Mazini, P. S., and Visentainer, J. E. L. (2014). Evidence of HLA-DQB1 contribution to susceptibility of dengue serotype 3 in dengue patients in Southern Brazil. *Journal of Tropical Medicine*, **2014**.

[5] Castañeda, D. M., Salgado, D. M., and Narváez, C. F. (2016). B cells naturally induced during dengue virus infection release soluble CD27, the plasma level of which is associated with severe forms of pediatric dengue. *Virology*, **497**, 136–145.

[6] Fitzgerald, K. A. (2011). The interferon inducible gene: Viperin.

[7] Gagnon-Bartsch, J. A. and Speed, T. P. (2012). Using control genes to correct for unwanted variation in microarray data. *Biostatistics*, **13**(3), 539–552.

[8] Gagnon-Bartsch, J. A., Jacob, L., and Speed, T. P. (2013). Removing unwanted variation from high dimensional data with negative controls. *Berkeley: Tech Reports from Dep Stat Univ California*, pages 1–112.

[9] Gentleman, R., Carey, V., Huber, W., Irizarry, R., and Dudoit, S. (2006). *Bioinformatics and computational biology solutions using R and Bioconductor*. Springer Science & Business Media.

[10] Gentleman, R. C., Carey, V. J., Bates, D. M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., *et al.* (2004). Bioconductor: open software development for computational biology and bioinformatics. *Genome biology*, **5**(10), R80.

[11] Hoerl, A. E. and Kennard, R. W. (1970). Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics*, **12**(1), 55–67.

[12] Holleman, A., Cheok, M. H., den Boer, M. L., Yang, W., Veerman, A. J., Kazemier, K. M., Pei, D., Cheng, C., Pui, C.-H., Relling, M. V., Janka-Schaub, G. E., Pieters, R., and Evans, W. E. (2004). Gene-Expression Patterns in Drug-Resistant Acute Lymphoblastic Leukemia Cells and Response to Treatment. *New England Journal of Medicine*, **351**(6), 533–542.

[13] Huber, W., Carey, V. J., Gentleman, R., Anders, S., Carlson, M., Carvalho, B. S., Bravo, H. C., Davis, S., Gatto, L., Girke, T., *et al.* (2015). Orchestrating high-throughput genomic analysis with bioconductor. *Nature methods*, **12**(2), 115.

[14] Johnstone, I. M. and Lu, A. Y. (2009). On consistency and sparsity for principal components analysis in high dimensions. *Journal of the American Statistical Association*, **104**(486), 682–693. PMID: 20617121.

[15] Johnstone, I. M. and Paul, D. (2018). Pca in high dimensions: An orientation. *Proceedings of the IEEE*, **106**(8), 1277–1292.

[16] Kaufmann, L. (1987). Clustering by means of medoids. *Proc. Statistical Data Analysis Based on the L1 Norm Conference, Neuchatel, 1987*, pages 405–416.

[17] Kwissa, M., Nakaya, H. I., Onlamoon, N., Wrammert, J., Villinger, F., Perng, G. C., Yoksan, S., Pattanapanyasat, K., Chokephaibulkit, K., Ahmed, R., and Pulendran, B. (2014). Dengue virus infection induces expansion of a CD14(+)CD16(+) monocyte population that stimulates plasmablast differentiation. *Cell host & microbe*, **16**(1), 115–27.

[18] Leek, J. T. and Storey, J. D. (2007). Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS genetics*, **3**(9), e161.

[19] Lloyd, S. (1982). Least squares quantization in pcm. *IEEE Transactions on Information Theory*, **28**(2), 129–137.

[20] Lopez, R., Regier, J., Cole, M. B., Jordan, M. I., and Yosef, N. (2018). Deep generative modeling for single-cell transcriptomics. *Nature Methods*, **15**(12), 1053–1058.

[21] Machado-Neto, J. A., Saad, S. T. O., and Traina, F. (2014). Stathmin 1 in normal and malignant hematopoiesis. *BMB reports*, **47**(12), 660–5.

[22] McInnes, L., Healy, J., and Melville, J. (2018). UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction.

[23] Nguyen, L. H. and Holmes, S. (2019). Ten quick tips for effective dimensionality reduction. *PLOS Computational Biology*, **15**(6), 1–19.

[24] R Core Team (2019). R: A language and environment for statistical computing.

[25] Ringner, M. (2008). What is principal component analysis?. *Nature biotechnology*, (3), 303.

[26] Risso, D., Ngai, J., Speed, T. P., and Dudoit, S. (2014). Normalization of rna-seq data using factor analysis of control genes or samples. *Nature biotechnology*, **32**(9), 896.

[27] Risso, D., Perraudeau, F., Gribkova, S., Dudoit, S., and Vert, J.-P. (2017). Zinb-wave: A general and flexible method for signal extraction from single-cell rna-seq data. *BioRxiv*, page 125112.

[28] Shen, D., Shen, H., and Marron, J. (2013). Consistency of sparse pca in high dimension, low sample size contexts. *Journal of Multivariate Analysis*, **115**, 317 – 333.

[29] van der Maaten, L. and Hinton, G. (2008). Visualizing data using t-sne.

[30] Yu, B. (2013). Stability. *Bernoulli*, **19**(4), 1484–1500.

[31] Zappia, L., Phipson, B., and Oshlack, A. (2017). Splatter: simulation of single-cell RNA sequencing data. *Genome Biology*, **18**(1), 174.

[32] Zhang, X. and Tao, W. (2019). Long Noncoding RNA LINC00152 Facilitates the Leukemogenesis of Acute Myeloid Leukemia by Promoting CDK9 Through miR-193a. *DNA and Cell Biology*, **38**(3), 236–242.

[33] Zhang, Z. H., Jhaveri, D. J., Marshall, V. M., Bauer, D. C., Edson, J., Narayanan, R. K., Robinson, G. J., Lundberg, A. E., Bartlett, P. F., Wray, N. R., and Zhao, Q.-Y. (2014). A comparative study of techniques for differential expression analysis on rna-seq data. *PLOS ONE*, **9**(8), 1–11.

[34] Zheng, G. X. Y., Terry, J. M., Belgrader, P., Ryvkin, P., Bent, Z. W., Wilson, R., Ziraldo, S. B., Wheeler, T. D., McDermott, G. P., Zhu, J., Gregory, M. T., Shuga, J., Montesclaros, L., Underwood, J. G., Masquelier, D. A., Nishimura, S. Y., Schnall-Levin, M., Wyatt, P. W., Hindson, C. M., Bharadwaj, R., Wong, A., Ness, K. D., Beppu, L. W., Deeg, H. J., McFarland, C., Loeb, K. R., Valente, W. J., Ericson, N. G., Stevens, E. A., Radich, J. P., Mikkelsen, T. S., Hindson, B. J., and Bielas, J. H. (2017). Massively parallel digital transcriptional profiling of single cells. *Nature Communications*, **8**(1), 14049.

[35] Zou, H. and Hastie, T. (2018). *elasticnet: Elastic-Net for Sparse Estimation and Sparse PCA*. R package version 1.1.1.

[36] Zou, H. and Xue, L. (2018). A selective overview of sparse principal component analysis. *Proceedings of the IEEE*, **106**(8), 1311–1320.

[37] Zou, H., Hastie, T., and Tibshirani, R. (2006). Sparse principal component analysis. *Journal of Computational and Graphical Statistics*, **15**(2), 265–286.