bioRxiv preprint doi: https://doi.org/10.1101/545798; this version posted February 10, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Differential Principal Components Reveal Patterns of Differentiation in Case/Control Studies

Benjamin J. Lengerich[1,2] **and Eric P. Xing**[1,2,3]

[1]Computer Science Department, Carnegie Mellon University, Pittsburgh, PA 15213
[2]Machine Learning Department, Carnegie Mellon University, Pittsburgh, PA 15213
[3]Petuum Inc., Pittsburgh, PA 15222

Dimensionality reduction is an important task in bioinformatics studies. Common unsupervised methods like principal components analysis (PCA) extract axes of variation that are high-variance but do not necessarily differentiate experimental conditions. Methods of supervised discriminant analysis such as partial least squares (PLS-DA) effectively separate conditions, but are hamstrung by inflexibility and overfit to sample labels. We would like a simple method which repurposes the rich literature of component estimation for supervised dimensionality reduction.
We propose to address this problem by estimating principal components from a set of difference vectors rather than from the samples. Our method directly utilizes the PCA algorithm as a module, so we can incorporate any PCA variant for improved components estimation. Specifically, Robust PCA, which ameliorates the deleterious effects of noisy samples, improves recovery of components in this framework. We name the resulting method *Differential Robust PCA* (drPCA). We apply drPCA to several cancer gene expression datasets and find that it more accurately summarizes oncogenic processes than do standard methods such as PCA and PLS-DA. A Python implementation of drPCA and Jupyter notebooks to reproduce experimental results are available at `www.github.com/blengerich/drPCA`.

Correspondence: *blengeri@cs.cmu.edu*

Introduction

Many bioinformatics datasets contain more features than samples, often because -omic assays profile many biomarkers but collecting data from a large number of individuals is costly. This reduces the statistical power of machine learning algorithms to distinguish signal from noise, a problem known as the curse of dimensionality (1). One way to alleviate this problem is to reduce the number of features in the dataset.

Principal components analysis (PCA) is the most popular way to summarize high-dimensional datasets. PCA projects datapoints onto the axes of major variation (2). While PCA minimizes reconstruction error of the training data as measured in Euclidean distance, the selected axes (and the resulting data representations) are not guaranteed to be biologically meaningful. For example, in gene expression studies, the axes of major variation often correspond to technical artifacts or biological processes which are not tightly regulated (3). These high-variance process are selected by PCA in order to reduce recovery error but they may not efficiently characterize the phenomenon of interest. Projecting data onto the top principal components can thus discard valuable information about tightly-regulated biological processes.

We propose to learn the principal components which summarize the *differences* between groups rather than optimize reconstruction error. If the low-dimensional representations succinctly capture the variation between groups, they may be more useful for understanding the processes of differentiation.

To estimate components of differentiation, we apply PCA-based methods to a set of vectors that define the difference between case and control groups. We call this framework *Differential PCA* (dPCA) and find that *Differential Robust PCA* (drPCA) compares favorably to supervised dimensionality reduction techniques while maintaining simplicity, modularity, and extensibility. Beyond the improved performance on the datasets presented in this article, we are excited about the possibility of the framework to be expanded by incorporating other techniques of dimensionality reduction techniques that have been developed for biological datsets.

A Python implementation of drPCA, as well as Jupyter notebooks to reproduce experimental results, can be downloaded from [1].

Motivating Example

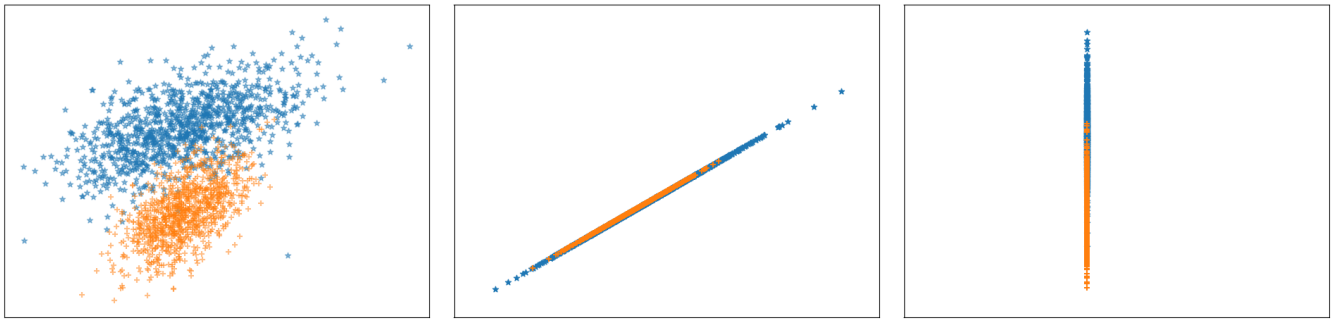
Shown in Figure 1 is an illustration of various dimensionality reduction methods on a toy dataset. Fig. 1a depicts the two-dimensional datapoints we are interested in compressing. This dataset has two different clusters: a background dataset (orange) generated by $X_{bg}^{(i)} \sim N\left((0,0), \begin{bmatrix} 1.5 & 0.5 \\ 0.5 & 0.5 \end{bmatrix}\right)$, and a foreground dataset (orange) generated by the sample-specific process $X_{fg}^{(i)} \sim N\left(X_{bg}^{(i)} + (0,2), \begin{bmatrix} 3.5 & 0.5 \\ 0.5 & 0.1 \end{bmatrix}\right)$. Each foreground datapoint is thus vertically offset from a background setting and perturbed by Gaussian noise.

When analyzing this dataset by components analysis, there are many experimental questions that may be asked. To understand the differences between foreground and background data, we may want to identify the vertical axis as the component of differentiation.

PCA projects the datapoints onto the axis of major variance (Fig. 1b), but this axis does not distinguish foreground and background samples. As a result, the clusters are completely

[1] `www.github.com/blengerich/drpca`

bioRxiv preprint doi: https://doi.org/10.1101/545798; this version posted February 10, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



(a) Raw Data **(b)** PCA **(c)** drPCA (ours)

Fig. 1. A toy example of dimensionality reduction. (a) Foreground datapoints (blue) are vertically offset from the background datapoints (orange) and perturbed by zero-mean noise. (b) Projection of datasets onto the first component recovered by PCA. (c) Projection of datasets onto the first component recovered by drPCA. In this setting, drPCA successfully recovers the axis which differentiates the foreground and background data, while PCA projects both groups onto an axis which does not distinguish foreground samples from background samples.

overlapping. In contrast, drPCA identifies the axis which differentiates foregrond and background samples (Fig. 1c).

Differential PCA

A differential PCA method (dPCA) was first propsoed by (4) as a targeted solution for ChIP-Seq data in which the goal is to identify protein binding capacities which differentiate experimental settings. Because Ji *et al.* specifically formulated dPCA for ChIP-Seq experiments, they make the natural assumption that there are many different experimental settings and many replicates of each setting. As a result, their proposed method first computes the mean of the data samples in each group, and then performs PCA on the dataset of differences between means. This process leads to selection of axes which characterize the protein-binding process studied in CHiP-Seq data, and has been successfully applied in several studies (5, 6).

When applying this idea to settings beyond ChIP-seq data, we encounter a major problem: the number of principal components is strictly less than the number of datapoints. If each datapoint is defined as the mean of the replicates for an experimental setting, then we must always have more experimental settings than desired components. For the common task of case/control data, no principal components are defined, and only a single prinicpal component can be extracted if the dataset is "grounded" by adding a zero vector.

To alleviate this problem, we do not take the means of each cluster. Instead, our dataset of difference vectors is calculated directly from pairs of samples. Given a list of matched samples, we calculate the high-dimensional difference vectors between the samples in the foreground set and the samples in the background set. To make this simple methodological change explicit, we will refer to the method of Ji *et al.* as dPCA-Mean, and the non-averaged method as dPCA.

Running components analysis directly on this differential dataset would identify axes of variance in the differences, but we are seeking to summarize the differential vectors. To analyze the differences as vectors, we "ground" the differential dataset by adding pseudo-samples of zero. With the number of zeros equivalent to the number of difference vectors, the principal components of the difference dataset summarize axes of differentiation between the sample clusters. This procedure is summarized in Alg. 1.

This construction allows us to utilize the well-studied suite of denoising and structured variants of PCA by replacing the call to PCA in line 6 of Alg. 1 with a call to a variant of PCA.

Algorithm 1 Differential PCA

Input: foreground dataset X, background dataset Y, matched pairs M, number of components N. **Output**: reduced datasets $\tilde{X}, \tilde{Y}$, components C, singular values S

1: **procedure** DPCA(X, Y, M, N)
2: $D := \{\}$
3: **for** each $(i, j) \in M$ **do**
4: $D := D \cup \{X[i] - Y[j]\}$
5: $D := D \cup \{\mathbf{0}\}$
6: **end for**
7: $C, S := PCA(D)$
8: $\tilde{X} := XC^T[: N]$
9: $\tilde{Y} := YC^T[: N]$
10: **return** $\tilde{X}, \tilde{Y}, C, S$
11: **end procedure**

Differential Robust PCA. As described below, Robust PCA (rPCA) decomposes the data into the sum of a low-rank component and a sparse component to increase stability in the presence of noise (7). We can incorporate rPCA in the dPCA framework by replacing the function call to PCA in line 6 of Alg. 1 with a call to $rPCA$. We call this method Differential Robust PCA (drPCA). In our experiments, we see that drPCA performs extremely well in differential datasets, even when case and control samples are not from matched sources. This highlights the benefit of the dPCA framework to reuse the rich literature of methods for components analysis.

Unmatched Samples. If samples are not taken from matched individuals, it becomes necessary to construct a new dataset of matched pairs. In this case, we can generate a matched dataset by uniformly selecting datapoints from each condition to be

bioRxiv preprint doi: https://doi.org/10.1101/545798; this version posted February 10, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

matched. This process produces a noisy differential dataset which is handled by the denoising aspects of drPCA.

Related Work

The most popular methods for dimensionality reduction are unsupervised linear methods, which find a linear transformation that projects the high-dimensional data points onto a nearby low-dimensional subspace. Of these, the most widely-known method is principal component analysis (PCA), which learns a set of linearly orthogonal features that represent the directions of maximal variance in the original data (8). Since PCA was first introduced, many variants have been developed. For example, Sparse PCA (9) uses an elastic net penalty to encourage element-wise sparsity in the projection matrix, while Independent Component Analysis (10) (ICA) recovers statistically independent components to separate signal sources. Robust PCA (rPCA) learns to decompose the data into the sum of a low-rank component and a sparse component, leading to increased stability in the presence of noise (7, 11).

There are also approaches that use richer models for the underlying latent representation of the data. These include methods that perform simultaneous dimensionality reduction and feature selection (12) or non-linear dimensionality reduction methods. Among these deep models, unsupervised methods such as variational (13) and denoising (14) autoencoders seek to learn latent features by optimizing data reconstruction in a bottlenecked architecture. These methods may also be extended to the supervised case (15); however, the deep architecture often requires more samples than are available for high-dimensional genomic assays. Additionally, it can be difficult to analyze the non-linear compression functions in a sample-agnostic way. For these reasons, we consider only linear dimensionality reduction techniques in the remainder of this paper.

Supervised Dimensionality Reduction. Supervised methods use sample labels in order to produce more meaningful data representations. Here, we describe several supervised methods frequently used in bioinformatics analyses.

Linear Discriminant Analysis. Linear Discriminant Analysis (LDA) (16) seeks to separate datapoints according to sample labels. To do so, LDA analytically maximizes cluster separation under a linear model. As a result, LDA produces clusters which are extremely well-separated. However, because LDA uses a closed-form solution, it can be difficult to extend the framework to recover desired structure in the components and can make it challenging to recover biologically interpretable components.

Supervised PCA. Supervised PCA (17) modifies traditional PCA by considering only the subset of explanatory variables which have sufficient correlation with the sample labels. In this way, Supervised PCA estimates sparse components which contain only features that may be predictive of the phenomenon of interest. However, as combinations of these features, the components may not describe the differences between sample groups.

Partial Least Squares and Canonical Correlation Analysis. Partial least squares (PLS) (18) and Canonical Correlation Analysis (CCA) (19) are bilinear factor models which fit linear projections for both outcomes and regressors. If categorical outcomes are used, as in sample labels, PLS is called PLS-Discriminant Analysis (PLS-DA) (20). While PLS-DA has many pleasing qualities, including handling high dimensions and multicolinearity well (for which CCA struggles), the method can be difficult to extend. Specifically, we would like to incorporate biological knowledge into our component estimation procedure, but it is not immediately clear how to modify PLS-DA to achieve this without implementing a new optimization procedure. This motivates us to consider the simple framework of differential PCA, in which well-tuned PCA variants may be used interchangably.

Contrastive PCA. A recently developed method for case/control data is Contrastive PCA (cPCA) (21). cPCA identifies axes that have large variance in the foreground samples but small variance in the background samples. In this way, cPCA can identify patterns of differentiation in the diseased state, potentially implicating dysregulated pathways. As shown in Section , cPCA and drPCA tend to select different components from cancer gene expression data; this suggests that the patterns which most differentiate cancers from one another are not the oncogenic processes which caused the cancers.

Differential Expression. For gene expression data, a closely related framework is differential expression (DE). In DE analyses, statistical tests are used to assess the probability that the means of the expression amounts for each gene are the same in both experimental settings. While both DE and dPCA seek to analyze the differences between two experimental conditions, there are stark differences. Firstly, dPCA performs subspace mapping rather than feature selection. In addition, we can induce structure in the recovered components via Bayesian methods, but this would be difficult in DE studies due to the univariate testing nature.

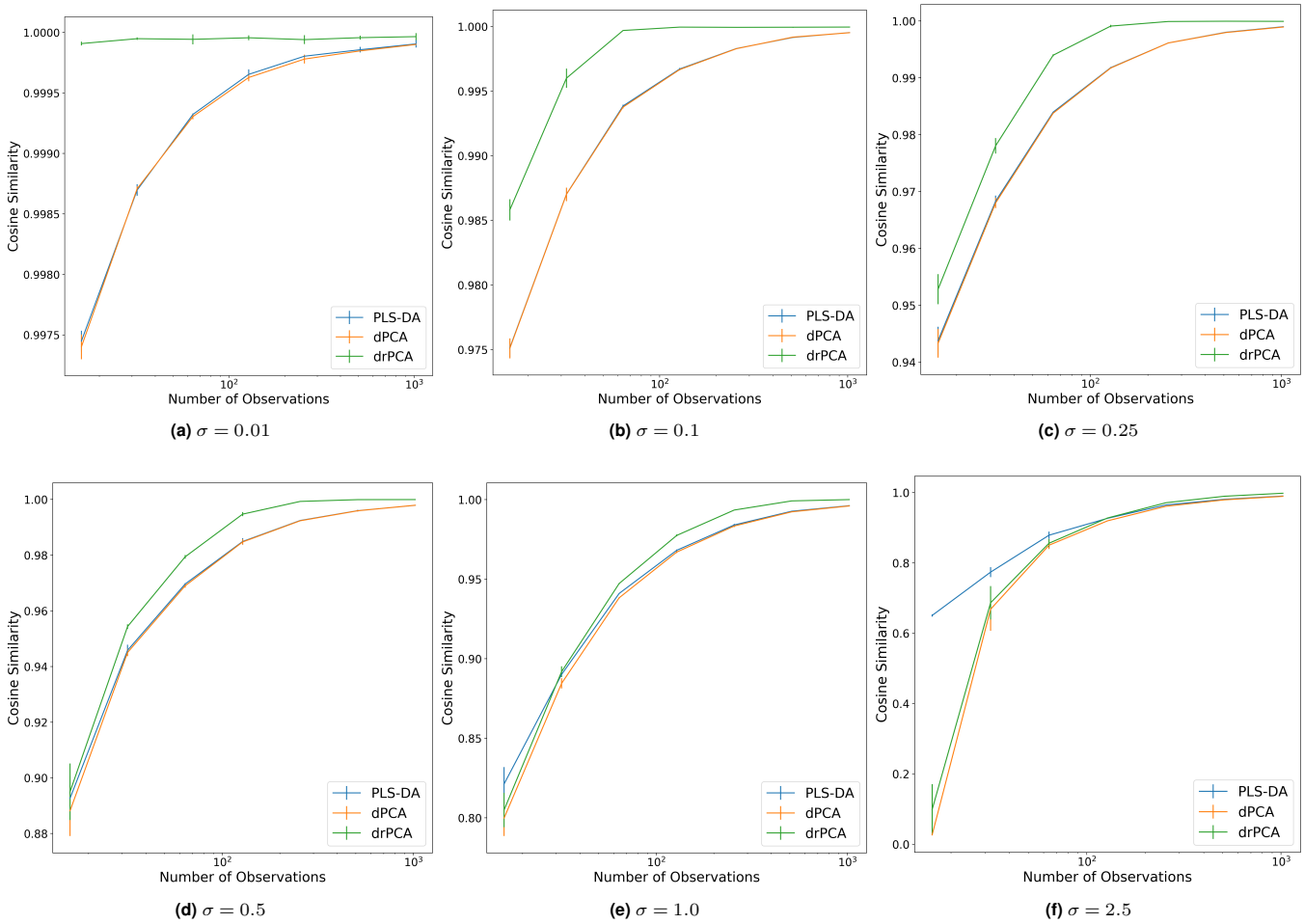
Discriminant Analysis of Principal Components. For SNP data, Discriminant Analysis of Principal Components (DAPC) is popular for identifying and describing clusters of genetically related individuals (22). However, DAPC uses population structure data that is specific to SNP assays. In contrast, our proposed method is relevant to any case/control study.

Experiments

To understand the behavior of these dimensionality reduction methods, we perform several experiments. First, we use simulated data to quantify how well drPCA recovers axes of differentiation from data. Next, we turn to cancer gene expression analysis to inspect the biological processes which differentiate tumor samples from healthy controls.

Simulated Data. We simulate data according to a mixture of two clusters. The background dataset is generated by $X_{bg}^{(i)} \sim N\Big((0,0), \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}\Big)$, and a foreground dataset is generated by

bioRxiv preprint doi: https://doi.org/10.1101/545798; this version posted February 10, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



(a) $\sigma = 0.01$ **(b)** $\sigma = 0.1$ **(c)** $\sigma = 0.25$

(d) $\sigma = 0.5$ **(e)** $\sigma = 1.0$ **(f)** $\sigma = 2.5$

Fig. 2. Recovery of axes of differentiation from simulated data for varying levels of noise governed by σ. On the y-axis, we show the cosine similarity between the true axis of differentiation and each method's first component. Results are averaged over 5 experimental settings, with standard deviations depicted by error bars. Results from other baseline methods are omitted because they all have cosine similarity below 0.8. In settings with low noise and moderate sample sizes, drPCA outperforms PLS-DA. In settings with large noise or very few samples, PLS-DA outperforms drPCA.

the sample-specific process $X_{fg}^{(i)} \sim N\left(X_{bg}^{(i)} + \mu, \sigma \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\right)$, creating an offset of μ between the foreground and background clusters. In this experiment, each datapoint consists of 1000 dimensions. To simulate biological data in which many features are unrelated to the process under study, we make the offset μ sparse, with only 5 non-zero entries. We generate n foreground and background datapoints according to this structure, and measure the cosine similarity between the estimated axes of differentiation and μ. Results for various n are shown in Figure 2.

As shown in Figure 2, drPCA outperforms the baselines at recovering μ, the axis of differentiation, in settings with low noise and moderate sample sizes. In settings with large noise or very few samples, PLS-DA outperforms drPCA. Other baseline methods have extremely poor performance on this task.

Cancer Gene Expression Studies. We investigate several RNA-seq gene expression datasets from The Cancer Genome Atlas[2]. These datasets profile cancer patients and contain tumor samples with some matched healthy tissues. We inspect three different cancer types: Breast Invasive Carcinoma (BRCA), Lung Adenocarcinoma (LUAD), and Glioblastoma Multiforme (GBM). These datasets contain 1102, 533, and 312 samples from cancer tissues, respectively. In addition, they contain 113, 59, and 10 samples from control tissues, respectively. The datasets are extremely high-dimensional; the datasets contain 15584, 14533, and 30584 distinct transcripts after thresholding features for a minimum standard deviation. In addition to these disease-specific datasets, we also evaluate dimensionality reduction on the combination of the three datasets (Combined). For the GBM dataset, we supplement the matched differences with 750 unmatched differences produced by randomly matching case and control samples. To evaluate the performance of the estimated components, we hold out 40% of the patients from each dataset for downstream tasks.

[2]cancergenome.nih.gov

Differential Components Separate Case and Control Samples. We first visually inspect the clusters induced in the top two components of each method. The projected data for the Combined dataset are shown in Fig. 3; similar results for the BRCA, LUAD, and GBM datasets are available at `github.com/blengerich/drpca`.

Standard PCA (Fig. 3a) and unsupervised variants (Fig. 3b,

bioRxiv preprint doi: https://doi.org/10.1101/545798; this version posted February 10, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

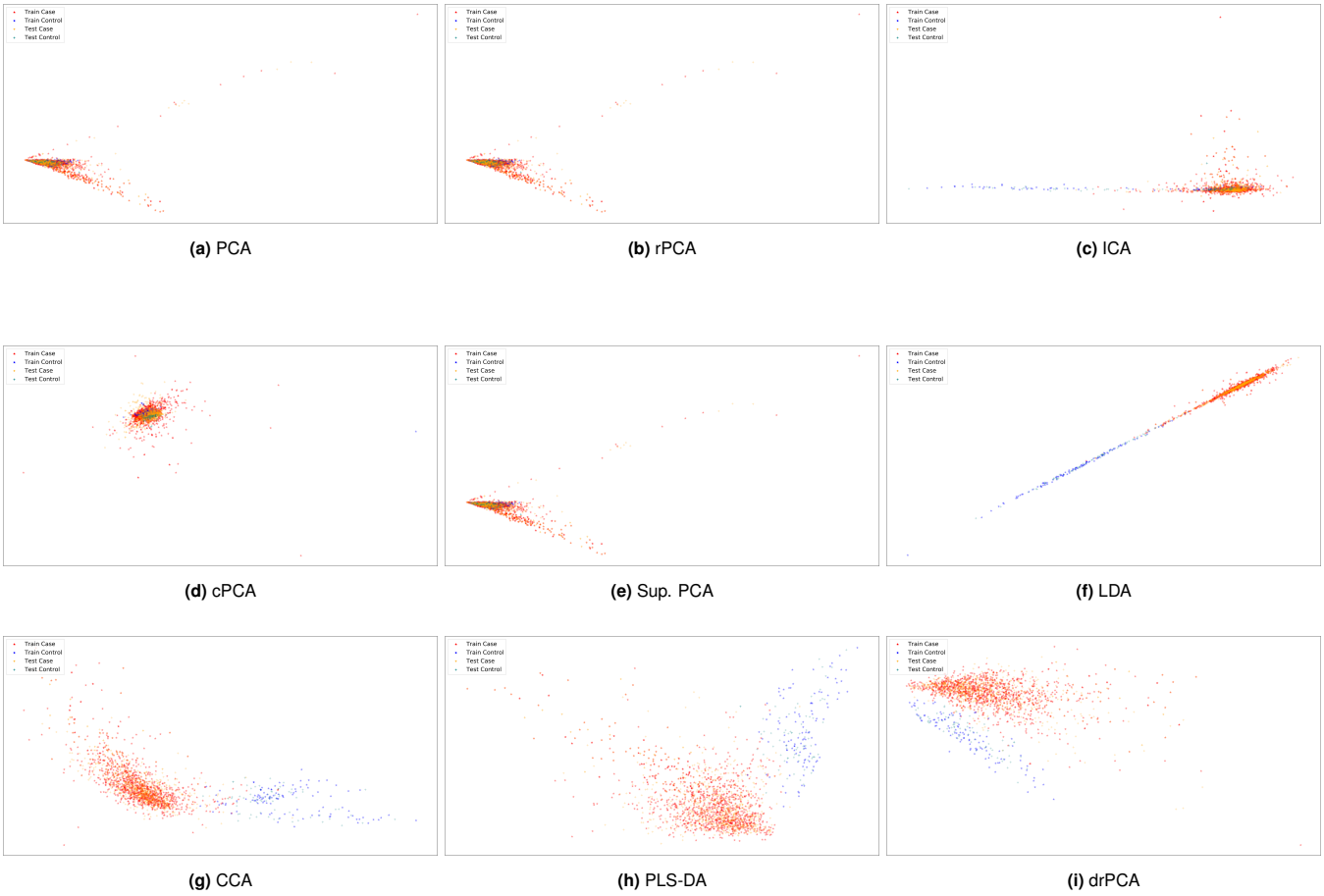


Fig. 3. Low-dimensional representations of samples from the combined cancer dataset. The supervised methods LDA (f), CCA (g), PLS-DA (h), and drPCA (i) all separate case and control samples.

Fig. 3c) project the data onto axes which do not differentiate case and control samples. While these axes may be useful for characterization of the samples, they are unlikely to correspond to processes which are causal for the tumors. In addition, we see that Contrastive PCA (Fig. 3d) effectively identifies processes which have high variance in the cancer samples but low variance in the control samples. These components may correspond to processes which are dysregulated in tumors but not oncogenic, a hypothesis supported by inspection of the component loadings (Tab. 1).

In contrast to the unsupervised dimensionality reduction methods, the supervised methods (Figs. 3e,3f,3g,3h,3i) all effectively separate case and control samples, with separability transferring betwen the training and test sets. Of these, drPCA and PLS-DA produce the most visually distinctive clusters, with mean silhouette scores greater than 0.6.

Differential Representations are Useful for Predictive Tasks. To test the presence of biological signal in these low-dimensional representations, we measure the performance of random forest (RF) classifiers for two tasks. First, we train the RFs to label case and control samples. After learning components from the training set, we optimize a RF classifier to predict case/control labels by cross-validation on the same training set. Plotted in Fig. 4 are the performances of the classifiers on the held-out test set projected into a given number of components. As expected, the supervised methods CCA, PLS-DA, LDA, dPCA, and drPCA all significantly outperform the unsupervised methods because they use the sample labels. However, the scientific utility of the representations produced by PLS-DA, LDA, and CCA are questionable; changing the task severely degrades predictive performance.

After reducing dimensionality based on case/control labels, we train another RF to predict the tissue of origin for each sample (recall that the combined dataset is composed of samples from three different tissue types). As shown in Fig. 5, the LDA representations contain very little information that is predictive of this task. As a result, the AUROCs using this method does not surpass 0.6. The representations from CCA and PLS-DA also struggle on this simple task. In contrast, the representations from dPCA and drPCA are among the best-performing representations for this task. This suggests that the differential components are biologically meaningful while the baseline supervised methods overfit to the labels.

Differential Components Summarize Oncogenic Patterns. Do the components which separate tumor samples from control samples give high weight to oncogenic processes? To answer this question, we sort the variables according to the magnitude of the weight in the first component of each method.

bioRxiv preprint doi: https://doi.org/10.1101/545798; this version posted February 10, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

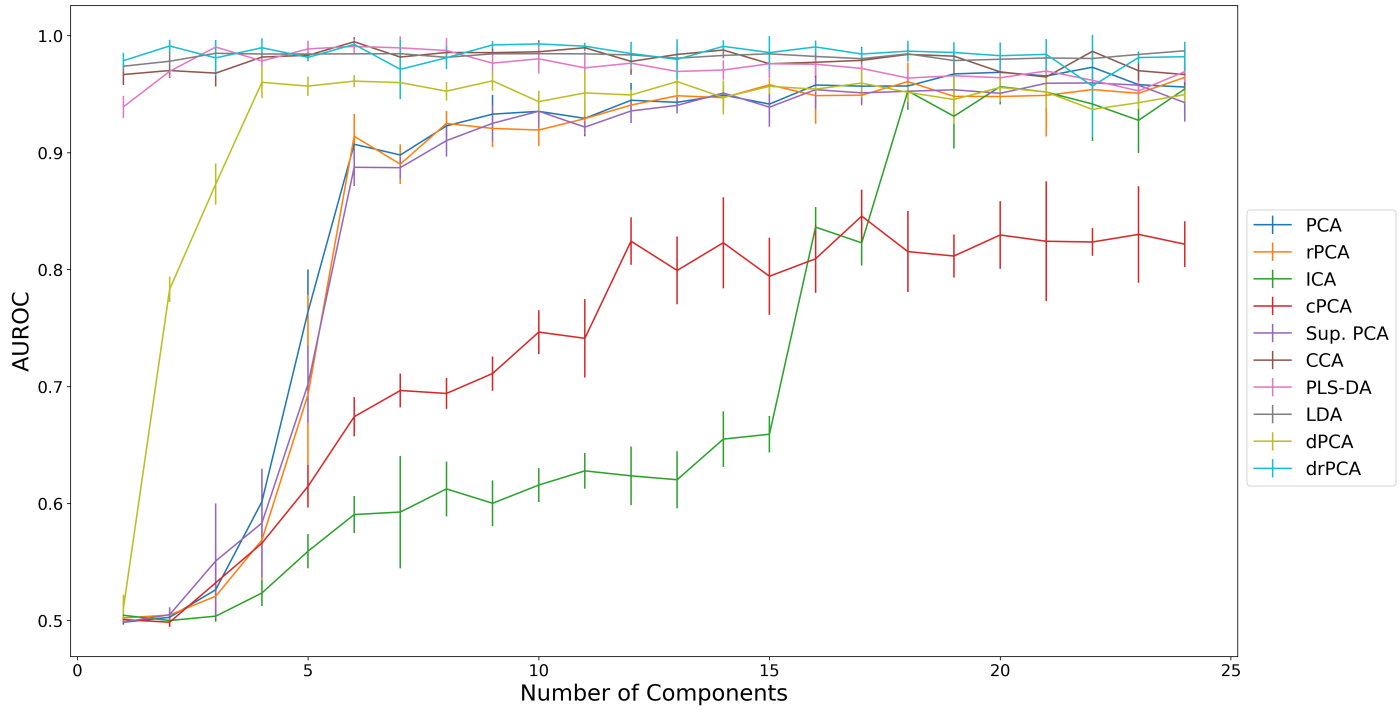


Fig. 4. Prediction of case/control status from the Combined dataset. The y-axis is the areas under the Receiver Operating Characteristic curves (AUROCs) of the prediction, with x-axis indicating the number of components used in the representations. Errorbars indicate the standard deviation over 3 train/test splits.

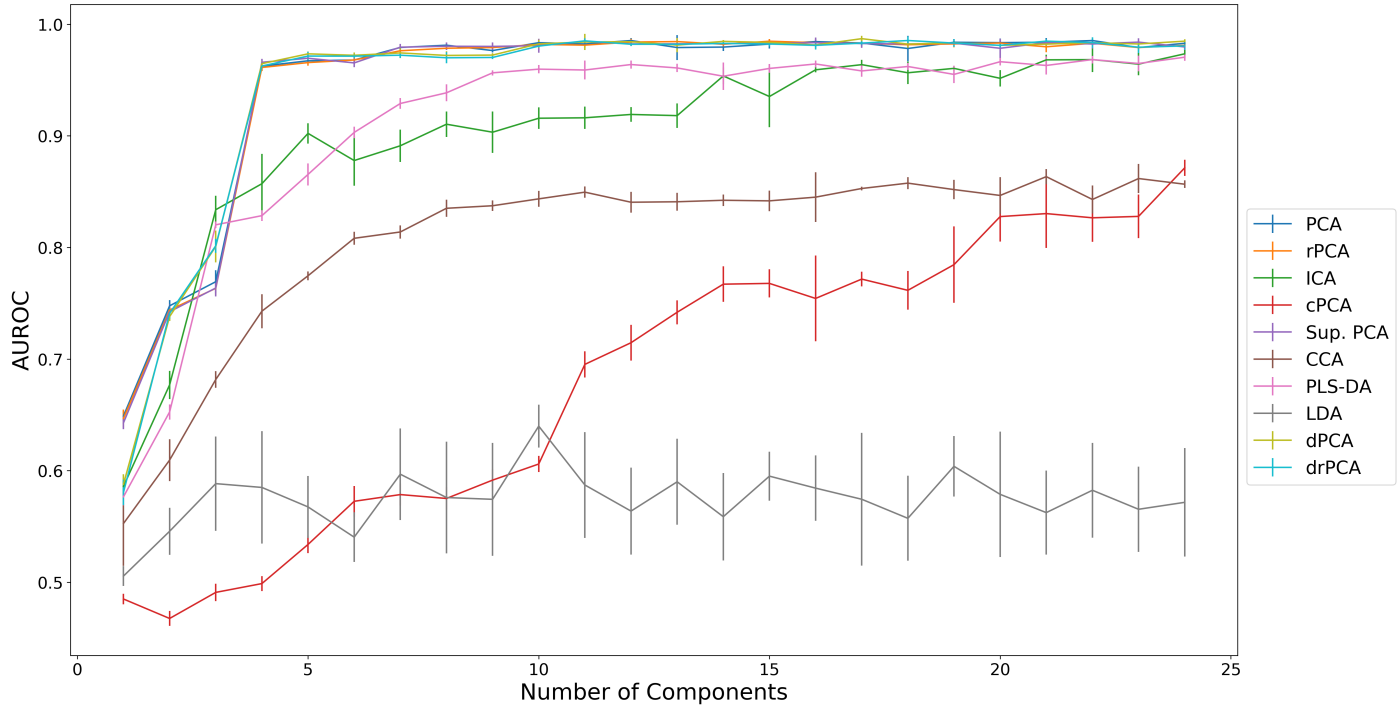


Fig. 5. Prediction of tissue of origin from the Combined dataset, with representations learned using case/control labels. The y-axis is the areas under the Receiver Operating Characteristic curves (AUROCs) of the prediction, with x-axis indicating the number of components used in the representations. Errorbars indicate the standard deviation over 3 train/test splits. The representations from LDA, CCA, and PLS-DA underperform most unsupervised methods, demonstrating that the components are not biologically meaningful. In contrast, the drPCA representations are among the best-performing.

From this ranked list, we count the number of selections annotated as oncogenes or tumor suppresor genes (TSG) in COSMIC (24) at each rank. As shown in Figure 6, the differential components give the highest weight to these oncogenic processes.

For a finer-grained anlaysis, we inspect the loadings of the components and variable selection patterns of each of the methods. Shown in Table 1 are the 5 highest weighted genes in the top component for each method; in addition to the methods for components analysis, we also compare to the results

bioRxiv preprint doi: https://doi.org/10.1101/545798; this version posted February 10, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

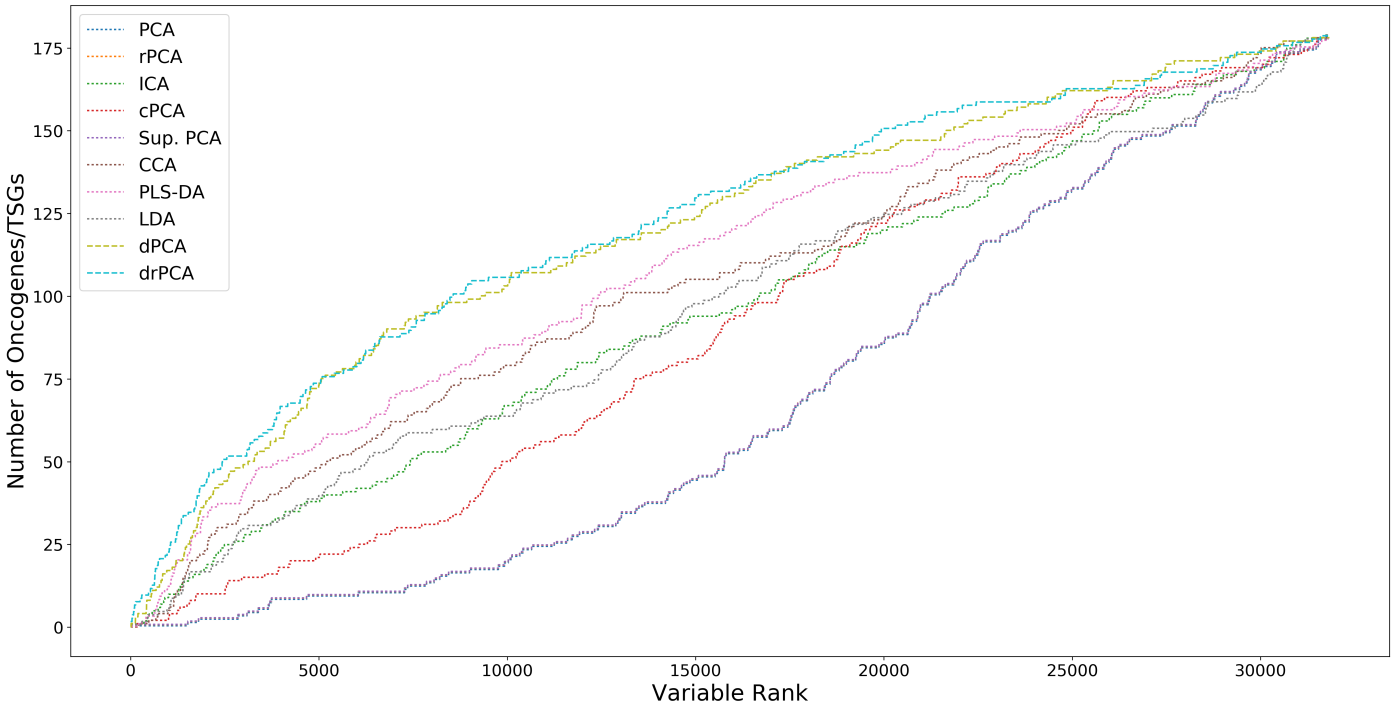


Fig. 6. Oncogene/Tumor Suppresor Gene (TSG) selection according to weight in the first component estimated from the Combined cancer dataset. Differential methods give the highest weights to cancer-associated genes.

LIMMA	PCA	ICA	cPCA
KIAA0101*	RP1-102G20.2	TBC1D9*	MYH1*
PAFAH1B3*	RNY1P15	FOXA1*	ACTC1*
C4	OFD1P17	RAB30*	AC005616.2
F2R*	RP1-102G20.5	RP11-102N12.3	DBET
ARHGAP6*	RP1-102G20.4	DACH1*	MYOG*

Sup. PCA	CCA	PLS-DA
RP1-102G20.2	VEGFD	AOC3*
RNY1P15	RP11-257I8.2	RP11-736K20.4
OFD1P17	CLEC3B*	VEGFD
RP1-102G20.5	RP11-193H18.3	CLEC3B*
RP1-102G20.4	CA4*	NPR1*

LDA	dPCA	drPCA
GLYAT*	ZSCAN25*	ZSCAN25*
CIDEC*	C2orf42*	C2orf42*
ADIPOQ*	FBXO42*	FBXO42*
TRHDE-AS1	FAM133B*	FAM133B*
C14orf180*	C2orf49*	RBM33*

Table 1. The most highly weighted genes in the first component that each method recovers from the Combined cancer dataset. Genes associated with tumor driver mutations (as annotated in DriverDB(23)) are indicated by a * symbol. The drPCA method gives highest weight to putative driver genes, while many baseline methods pay most attention to high-variance processes, often ribosomal proteins.

of the LIMMA differential enrichment test (25), as compiled by GEPIA (26). Traditional components analyses tend to give higher weight to genes which are associated with high-variance processes, such as cell cycle or ribosomal proteins. In contrast, the differential components select variables more directly related to tumor generation. Each of the top 5 genes selected by drPCA has been implicated as a driver mutation according to DriverDB (23). The components which differentiate tumor samples from control samples are indeed mostly composed of genes known to be associated with cancer.

Discussion

Summarizing the ways in which samples differ between experimental conditions is a central task in scientific inquiry, but made difficult by the large dimensionality of modern bioinformatics datasets. Common methods of unsupervised dimensionality reduction produce components which do not distinguish between experimental groups. This problem is worsened by the selective pressure of the observations in gene expression studies; axes of major variation in the data often correspond to unregulated, and largely noncritical, processes. In this paper, we have presented a way to extract the processes which differentiate experimental conditions by adapting unsupervised techniques to the supervised setting. Our framework can outperform methods designed specifically for supervised data when denoising methods, such as in drPCA, are used.

We are interested to see the variety of unsupervised dimensionality reduction techniques that can be repurposed in this supervised setting. For instance, we may want to estimate components which correspond to genetic pathways, similarly to (27, 28). This can be accomplished under the differential PCA framework by using a Bayesian PCA method with a prior that links genes according to pathway annotations. Even without these additional biological information, we have shown that drPCA outperforms common supervised dimensionality reduction methods at producing biologically-meaningful components.

bioRxiv preprint doi: https://doi.org/10.1101/545798; this version posted February 10, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Conclusions

In this paper, we have considered whether components analysis can compress samples from case/control studies onto biologically meaningful axes. We found that PCA and unsupervised variants do not separate case and control samples due to overemphasis on high-variance processes such as cell cycle and ribosomal processes. Supervised methods of dimensionality reduction separate case and control samples, but the resulting components have questionable biological utility and are overfit to sample labels. To address this problem, we have presented differential PCA: a method which applies PCA on the set of difference vectors between the samples. Our new method Differential Robust PCA (drPCA) effectively identifies axes of differentiation, outperforming standard supervised methods such as PLS-DA even under noisy conditions. When applied to gene expression data of cancer patients, drPCA produces components which summarize oncogenic processes. In future work, we are interested to incorporate prior biological knowledge to extract pathway-level components.

Acknowledgements

Thanks to Bryon Aragam, Haohan Wang, Michael Kleyman, and Ziv Bar-Joseph for helpful discussion about these ideas.

Funding

This work is supported by the National Institutes of Health grants R01-GM093156 and P30-DA035778.

Bibliography

1. Gordon Hughes. On the mean accuracy of statistical pattern recognizers. *IEEE transactions on information theory*, 14(1):55–63, 1968.
2. Karl Pearson. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572, 1901. doi: 10.1080/14786440109462720.
3. J Luo, M Schumacher, A Scherer, Despoina Sanoudou, D Megherbi, T Davison, T Shi, W Tong, L Shi, H Hong, et al. A comparison of batch effect removal methods for enhancement of prediction performance using maqc-ii microarray gene expression data. *The pharmacogenomics journal*, 10(4):278, 2010.
4. Hongkai Ji, Xia Li, Qian-fei Wang, and Yang Ning. Differential principal component analysis of chip-seq. *Proceedings of the National Academy of Sciences*, 110:201204398, 2013.
5. Shaun Mahony, Matthew D Edwards, Esteban O Mazzoni, Richard I Sherwood, Akshay Kakumanu, Carolyn A Morrison, Hynek Wichterle, and David K Gifford. An integrated model of multiple-condition chip-seq data reveals predeterminants of cdx2 binding. *PLoS computational biology*, 10(3):e1003501, 2014.
6. Angela Yen and Manolis Kellis. Systematic chromatin state comparison of epigenomes associated with diverse properties including sex and tissue type. *Nature communications*, 6:7973, 2015.
7. Emmanuel J Candès, Xiaodong Li, Yi Ma, and John Wright. Robust principal component analysis? *Journal of the ACM (JACM)*, 58(3):11, 2011.
8. I. T. Jolliffe. *Graphical Representation of Data Using Principal Components*, pages 64–91. Springer New York, New York, NY, 1986. ISBN 978-1-4757-1904-8. doi: 10.1007/978-1-4757-1904-8_5.
9. Hui Zou, Trevor Hastie, and Robert Tibshirani. Sparse principal component analysis. *Journal of computational and graphical statistics*, 15(2):265–286, 2006.
10. Te-Won Lee. Independent component analysis. In *Independent component analysis*, pages 27–66. Springer, 1998.
11. Huan Xu, Constantine Caramanis, and Sujay Sanghavi. Robust pca via outlier pursuit. In *Advances in Neural Information Processing Systems*, pages 2496–2504, 2010.
12. Micol Marchetti-Bowick, Benjamin J Lengerich, Ankur P Parikh, and Eric P Xing. Hybrid subspace learning for high-dimensional data. *arXiv preprint arXiv:1808.01687*, 2018.
13. Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. *arXiv preprint arXiv:1401.4082*, 2014.
14. Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, pages 1096–1103, New York, NY, USA, 2008. ACM. ISBN 978-1-60558-205-4. doi: 10.1145/1390156.1390294.
15. Elina Parviainen. Deep bottleneck classifiers in supervised dimension reduction. In Konstantinos Diamantaras, Wlodek Duch, and Lazaros S. Iliadis, editors, *Artificial Neural Networks – ICANN 2010*, pages 1–10, Berlin, Heidelberg, 2010. Springer Berlin Heidelberg. ISBN 978-3-642-15825-4.
16. Ronald A Fisher. The use of multiple measurements in taxonomic problems. *Annals of eugenics*, 7(2):179–188, 1936.
17. Eric Bair, Trevor Hastie, Debashis Paul, and Robert Tibshirani. Prediction by supervised principal components. *Journal of the American Statistical Association*, 101(473):119–137, 2006. doi: 10.1198/016214505000000628.
18. Svante Wold, Michael Sjöström, and Lennart Eriksson. Pls-regression: a basic tool of chemometrics. *Chemometrics and intelligent laboratory systems*, 58(2):109–130, 2001.
19. Bruce Thompson. *Canonical Correlation Analysis*. American Cancer Society, 2005. ISBN 9780470013199. doi: 10.1002/0470013192.bsa068.
20. Michael Sjöström, Svante Wold, and Bengt Söderström. Pls discriminant plots. In *Pattern Recognition in Practice, Volume II*, pages 461–470. Elsevier, 1986.
21. Abubakar Abid, Vivek K Bagaria, Martin J Zhang, and James Zou. Contrastive principal component analysis. *arXiv preprint arXiv:1709.06716*, 2017.
22. Thibaut Jombart, Sébastien Devillard, and François Balloux. Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. *BMC genetics*, 11(1):94, 2010.
23. Wei-Chung Cheng, I-Fang Chung, Chen-Yang Chen, Hsing-Jen Sun, Jun-Jeng Fen, Wei-Chun Tang, Ting-Yu Chang, Tai-Tong Wong, Hsei-Wei Wang, et al. Driverdb: an exome sequencing database for cancer driver gene identification. *Nucleic acids research*, 42(D1):D1048–D1054, 2014.
24. Simon A Forbes, Nidhi Bindal, Sally Bamford, Charlotte Cole, Chai Yin Kok, David Beare, Mingming Jia, Rebecca Shepherd, Kenric Leung, Andrew Menzies, et al. Cosmic: mining complete cancer genomes in the catalogue of somatic mutations in cancer. *Nucleic acids research*, 39 (suppl_1):D945–D950, 2010.
25. Matthew E Ritchie, Belinda Phipson, Di Wu, Yifang Hu, Charity W Law, Wei Shi, and Gordon K Smyth. limma powers differential expression analyses for rna-sequencing and microarray studies. *Nucleic acids research*, 43(7):e47–e47, 2015.
26. Zefang Tang, Chenwei Li, Boxi Kang, Ge Gao, Cheng Li, and Zemin Zhang. Gepia: a web server for cancer and normal gene expression profiling and interactive analyses. *Nucleic acids research*, 45(W1):W98–W102, 2017.
27. Weiguang Mao, Boris Harmann, Stuart C Sealfon, Elena Zaslavsky, and Maria Chikina. Pathway-level information extractor (plier) for gene expression data. *bioRxiv*, 2017. doi: 10.1101/116061.
28. Jaclyn N Taroni, Peter C Grayson, Qiwen Hu, Sean Eddy, Matthias Kretzler, Peter A Merkel, and Casey S Greene. Multiplier: a transfer learning framework for transcriptomics reveals systemic features of rare disease. *bioRxiv*, 2019. doi: 10.1101/395947.