Exploring patterns enriched in a dataset with contrastive principal component analysis

Abid et al., Nature Communications 2018

Outline

1. Motivation
2. Quick review of PCA
3. Intro to contrastive PCA
4. Results
5. Algorithmic details
6. Discussion

Outline

1. Motivation
2. Quick review of PCA
3. Intro to contrastive PCA
4. Results
5. Algorithmic details
6. Discussion

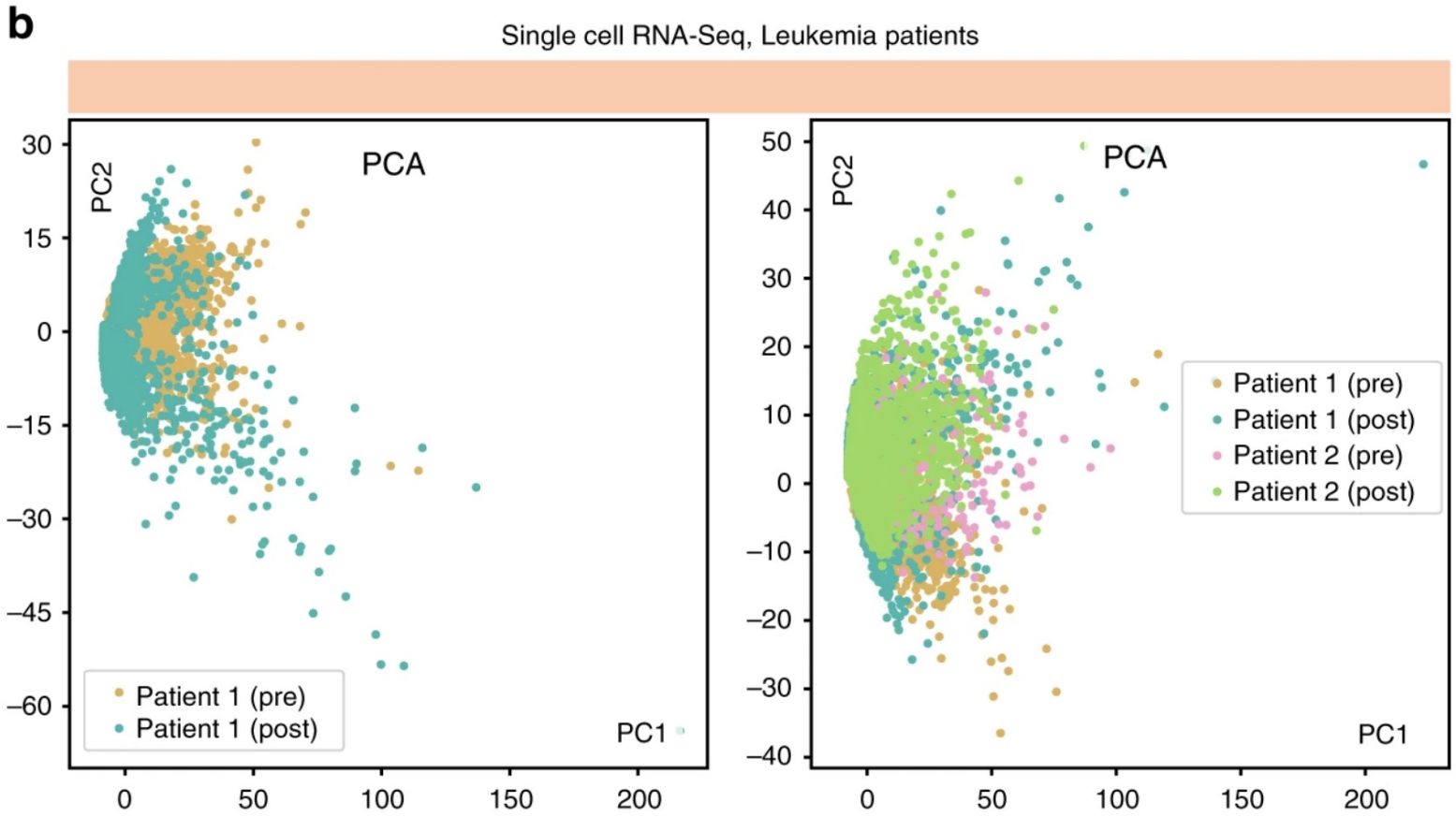
Motivation

- Computational biologists love visualizing new datasets
- Many widely-used approaches for summarizing data/visualizing it in 2D
- Commonly used methods include PCA/t-SNE/UMAP

Problems with PCA

- Oftentimes (especially in biology) we want to explore variations *enriched* in one (target) dataset compared to another (background) dataset
- Data from sick vs. healthy patients, treatment vs. control, etc.
- Unfortunately, enriched variations may be subtle compared to overall variations

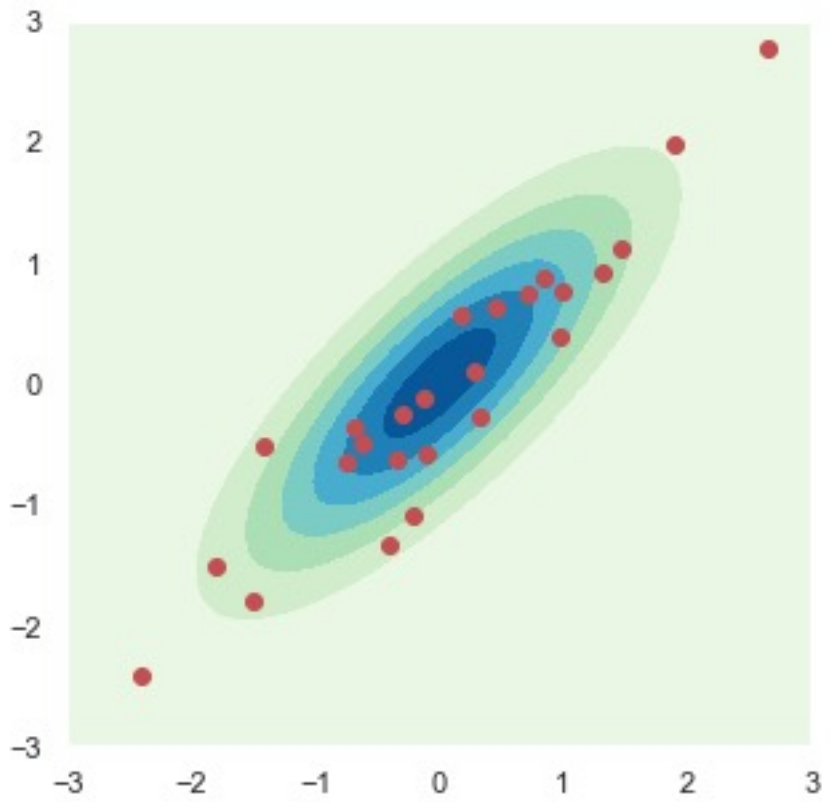
Example


b
Single cell RNA-Seq, Leukemia patients
PCA
PC2
PC1
30
15
0
−15
−30
−45
−60
0
50
100
150
200
Patient 1 (pre)
Patient 1 (post)
PCA
PC2
PC1
50
40
30
20
10
0
−10
−20
−30
−40
0
50
100
150
200
Patient 1 (pre)
Patient 1 (post)
Patient 2 (pre)
Patient 2 (post)

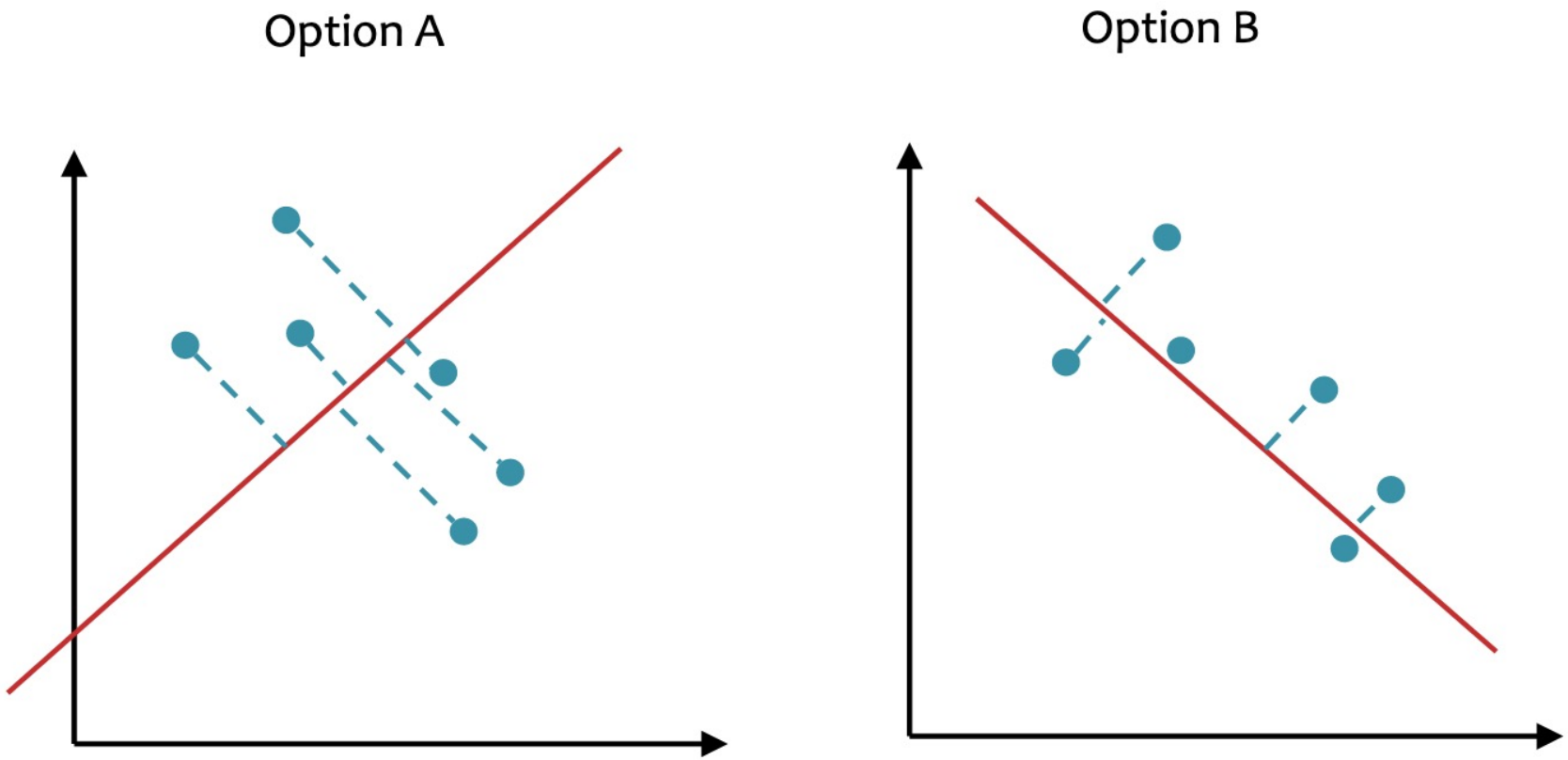
Outline

1. Motivation
2. Quick review of PCA
3. Intro to contrastive PCA
4. Results
5. Algorithmic details
6. Discussion

PCA review



PCA review



PCA review

- Direction that minimizes residuals also maximizes variance of projections
- Is an eigenvector of the covariance matrix of the data
  - In particular, the eigenvector with largest eigenvalue
- Unfortunately, if covariance matrix dominated by background variations, the direction we pick won’t reflect enriched variations in target

Outline

1. Motivation
2. Quick review of PCA
3. Intro to contrastive PCA
4. Results
5. Algorithmic details
6. Discussion

Contrastive PCA

- Idea: enriched variations are (by definition) found only in target points, not in background
- Find directions that have high variance in target dataset, low variance in background dataset

Contrastive PCA

- Target dataset: $\boldsymbol{x}_i \in \mathbb{R}^d$
- Background dataset: $\boldsymbol{y}_j \in \mathbb{R}^d$
- Target/background covariance matrices: $C_{\boldsymbol{x}}$ / $C_{\boldsymbol{y}}$
- Target Variance: $\lambda_X(\boldsymbol{v}) = \boldsymbol{v}^T C_X \boldsymbol{v}$

- Background Variance: $\lambda_Y(\boldsymbol{v}) = \boldsymbol{v}^T C_Y \boldsymbol{v}$

Contrastive PCA

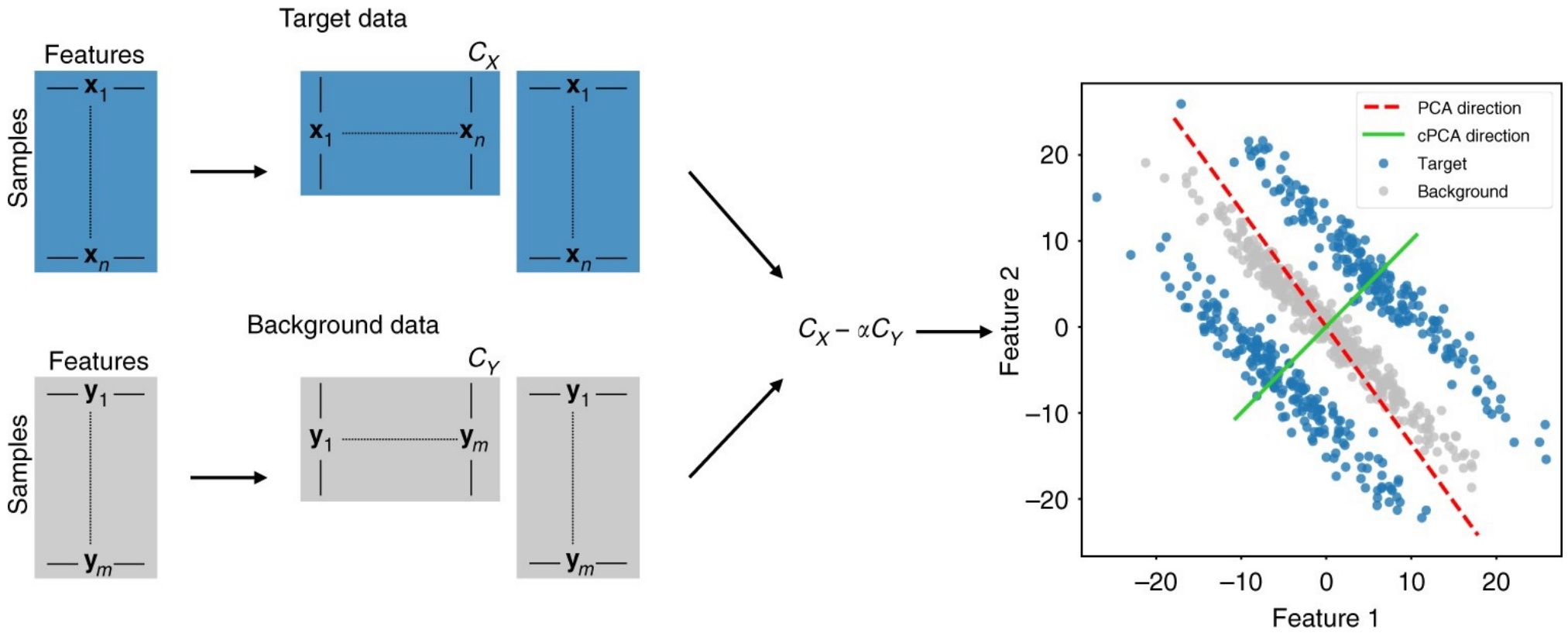
$$\boldsymbol{v}^* = \text{argmax}_{\boldsymbol{v}}\ \lambda_X(\boldsymbol{v}) - \alpha\lambda_Y(\boldsymbol{v})$$

How to choose?

$$\boldsymbol{v}^* = \text{argmax}_{\boldsymbol{v}}\ \boldsymbol{v}^T(C_{\boldsymbol{x}} - \alpha C_{\boldsymbol{y}})\boldsymbol{v}$$

Eigenvector

Concept Figure/Simulated Data Results



Contrastiveness

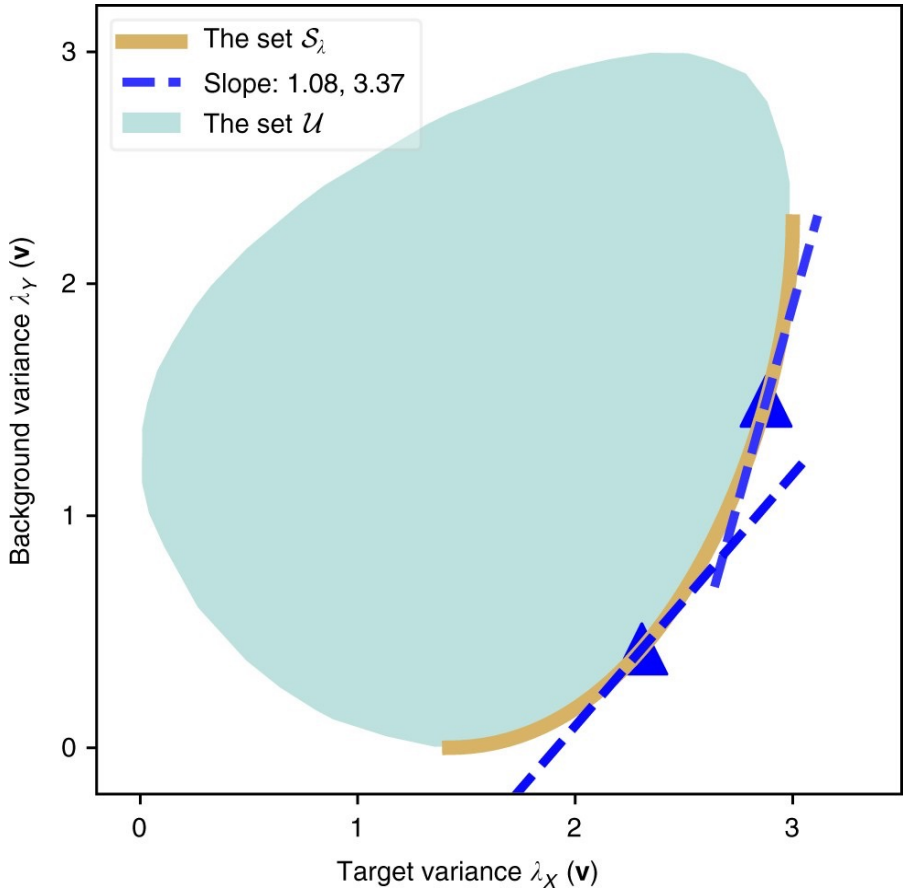
Definition 1. *(Contrastiveness) For two directions* $\mathbf{v}_1, \mathbf{v}_2 \in \mathbb{R}^d_{unit}$, $\mathbf{v}_1$ *is more contrastive than* $\mathbf{v}_2$ *with respect to the target and the background covariance matrices* C_X *and* C_Y, *written as* $\mathbf{v}_1 \succ \mathbf{v}_2$, *if one of the following is true:*

$$(1) \quad \lambda_X(\mathbf{v}_1) \geq \lambda_X(\mathbf{v}_2), \text{ and } \lambda_Y(\mathbf{v}_1) < \lambda_Y(\mathbf{v}_2)$$

$$(2) \quad \lambda_X(\mathbf{v}_1) > \lambda_X(\mathbf{v}_2), \text{ and } \lambda_Y(\mathbf{v}_1) \leq \lambda_Y(\mathbf{v}_2).$$

Contrastiveness and cPCA

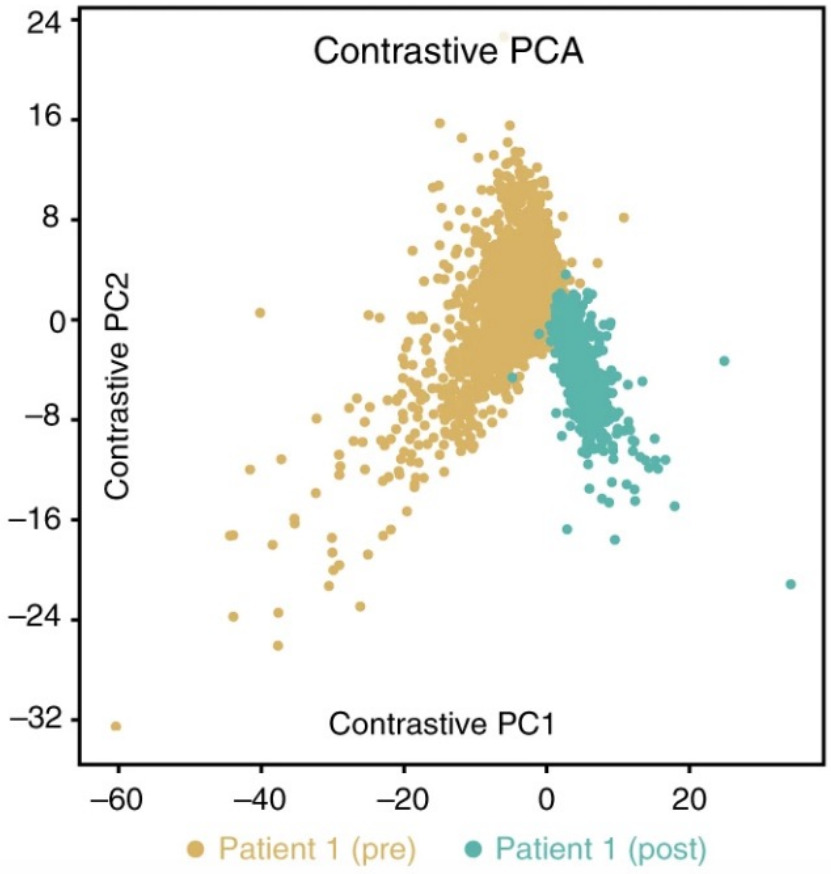
- Can show that cPCA directions are always “most contrastive” (i.e., no directions are more contrastive)

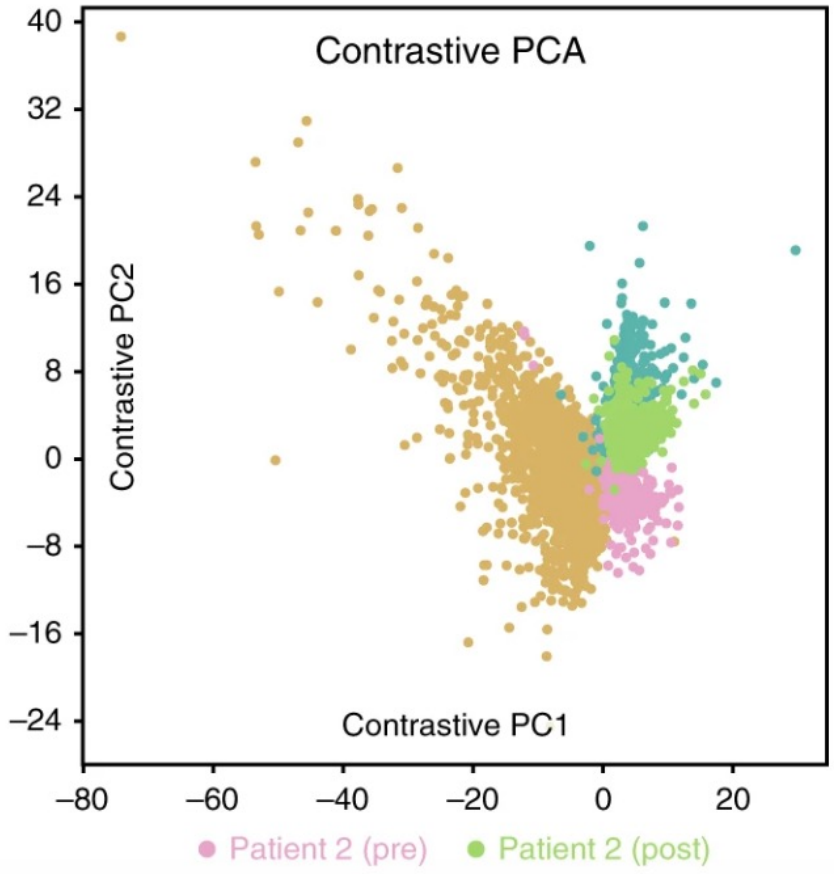


Outline

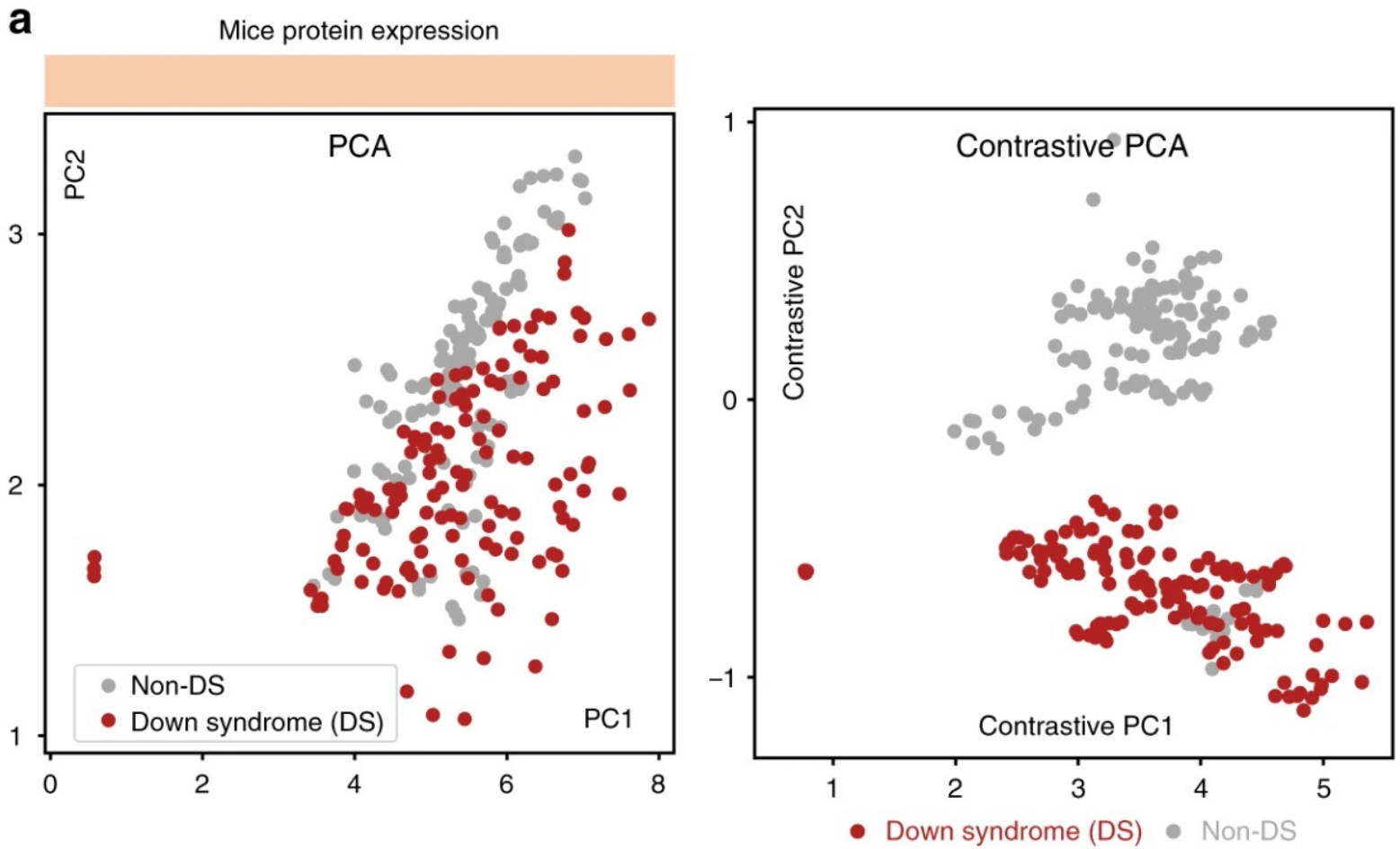
1. Motivation
2. Quick review of PCA
3. Intro to contrastive PCA
4. Results
5. Algorithmic details
6. Discussion

AML scRNA-seq

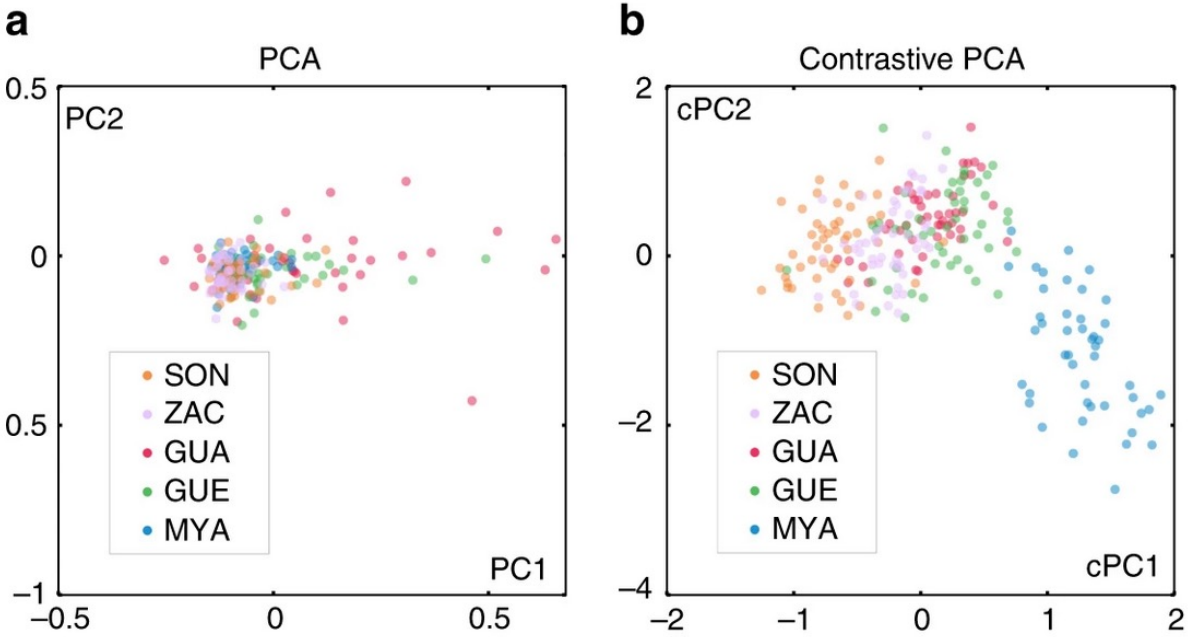




Mice Protein



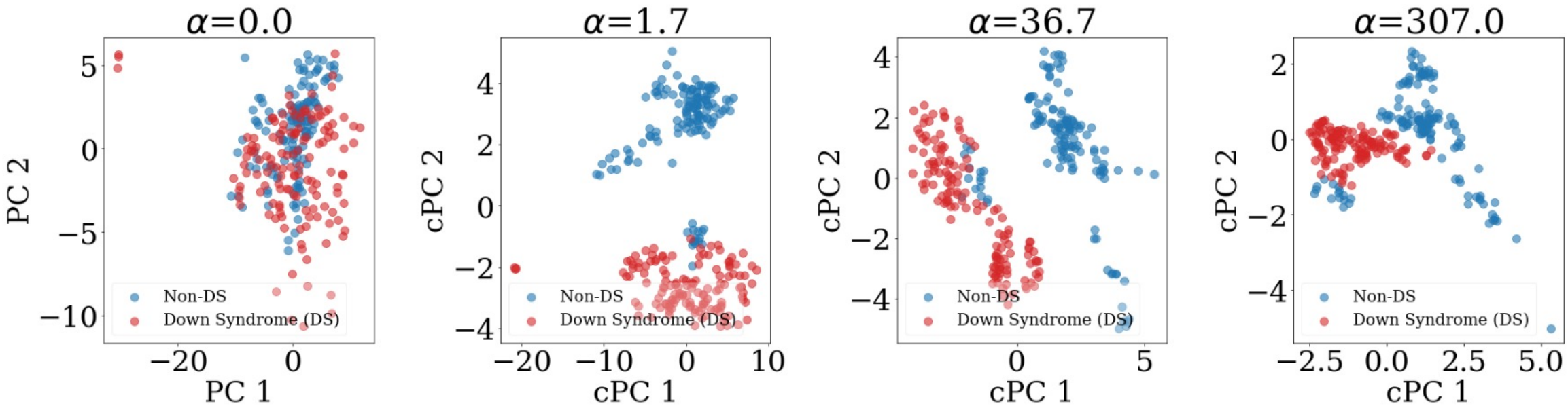
Mexican Ancestry



Outline

1. Motivation
2. Quick review of PCA
3. Intro to contrastive PCA
4. Results
5. Algorithmic details
6. Discussion

The role of alpha



How to choose alpha?

Algorithm 2 cPCA with Auto Selection of α

Inputs: target data $\{\mathbf{x}_i\}_{i=1}^n$; background data $\{\mathbf{y}_i\}_{i=1}^m$; list of possible contrastive parameters $\{\alpha_i\}$; the number of components k; the number of α's to present p.

for each α_i **do**
  Compute the subspace V_i using Algorithm 1 with the contrast parameter set to α_i.
end for

How to choose alpha?

for each pair V_i, V_j **do**
  Compute the principal angles $\theta_1 \ldots \theta_k$ between V_i, V_j
  Define the affinity $d(V_i, V_j) = \prod_{h=1}^{k} \cos\theta_h$
end for
With $D_{ij} = d(V_i, V_j)$ as an affinity matrix between subspaces, do spectral clustering on D to produce p clusters.

How to choose alpha?

for each cluster of subspaces $\{c_i\}_{i=1}^{p}$ **do**
Compute its medoid, V_i^* the subspace defined as

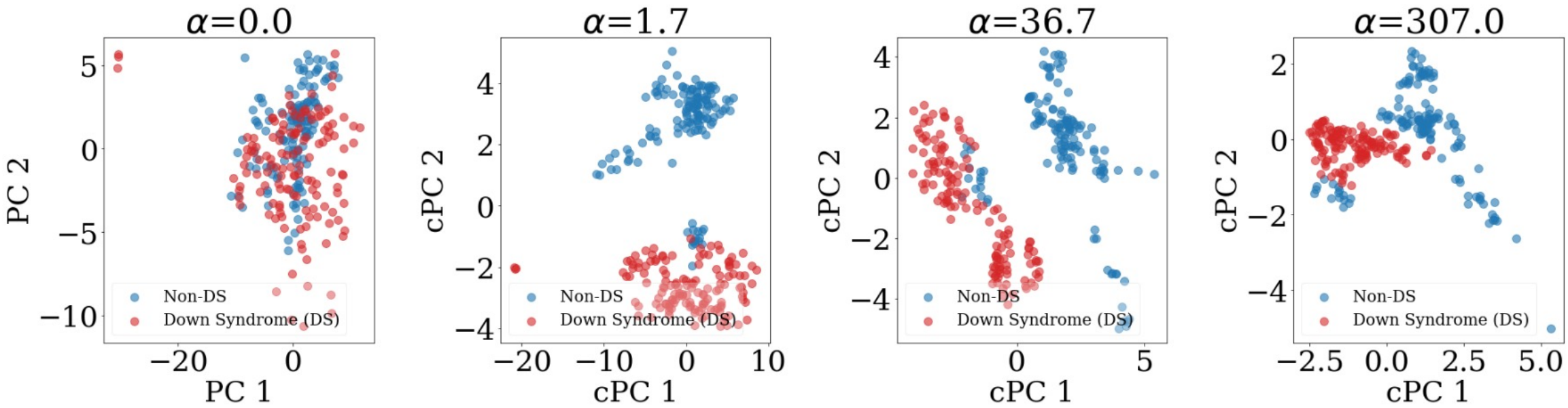
$$V_i^* \stackrel{\text{def}}{=} \arg\min V \in c_i \sum_{V' \in c_i} d(V, V')$$

Let α_i^* be the contrast parameter corresponding to V_i^*.
end for
Return: $\alpha_1^* \cdots \alpha_p^*$ and the subspaces $V_1^* \cdots V_p^*$.

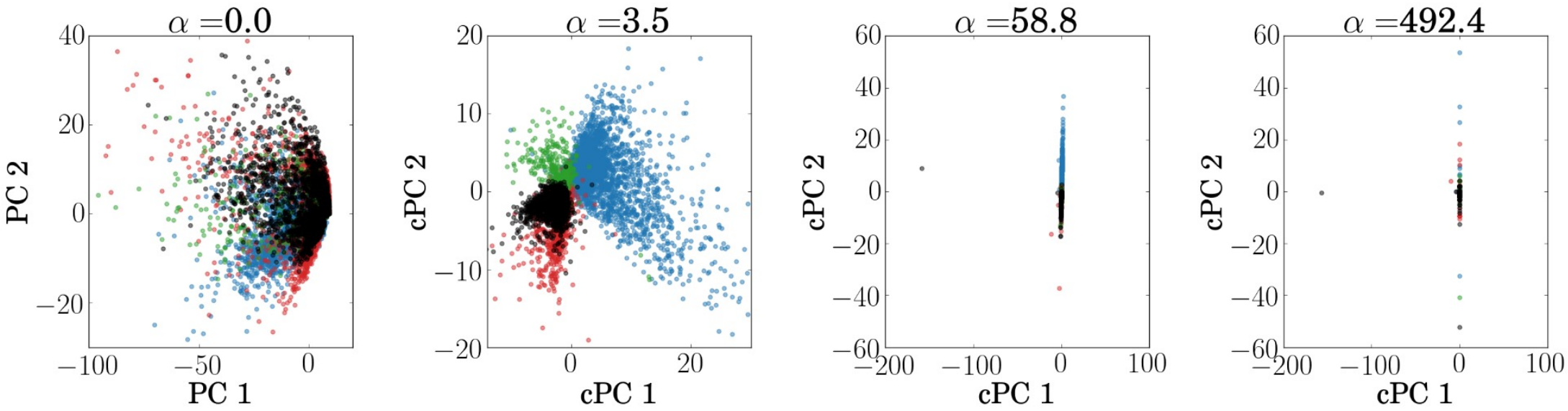
Summary of alpha choosing procedure

- Try a bunch of alphas and get the resulting subspaces
- Measure how similar each pair of subspaces is
- Cluster subspaces based on their similarities
- Pick a representative subspace/alpha from these clusters
- Return list of potential alphas

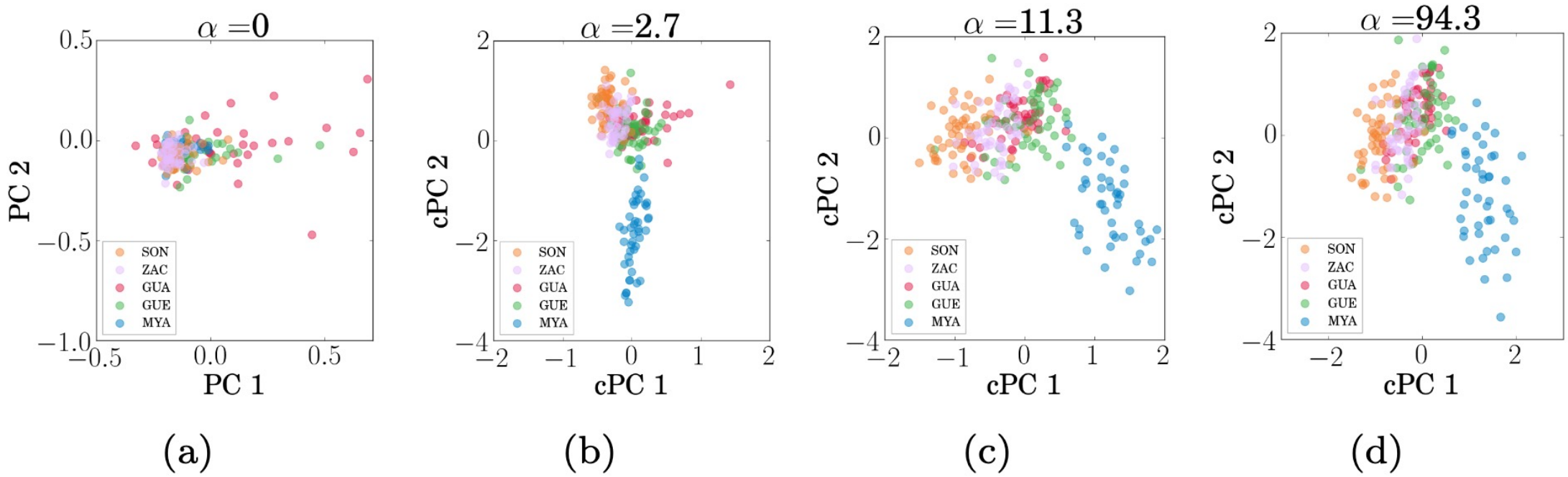
Different alpha values for mice protein



Different alpha values for scRNA-seq



Different alpha values for Mexican heritage



Outline

1. Motivation
2. Quick review of PCA
3. Intro to contrastive PCA
4. Results
5. Algorithmic details
6. Discussion

Discussion Questions

- What do people think of the definition of “contrastiveness”?
- Are people ok with the selection process for alpha?
- How useful would cPCA be when labels aren’t available?
- Could cPCA be extended to other cases?