

Selective inference and clustering

Lucy Gao

February 8 2023

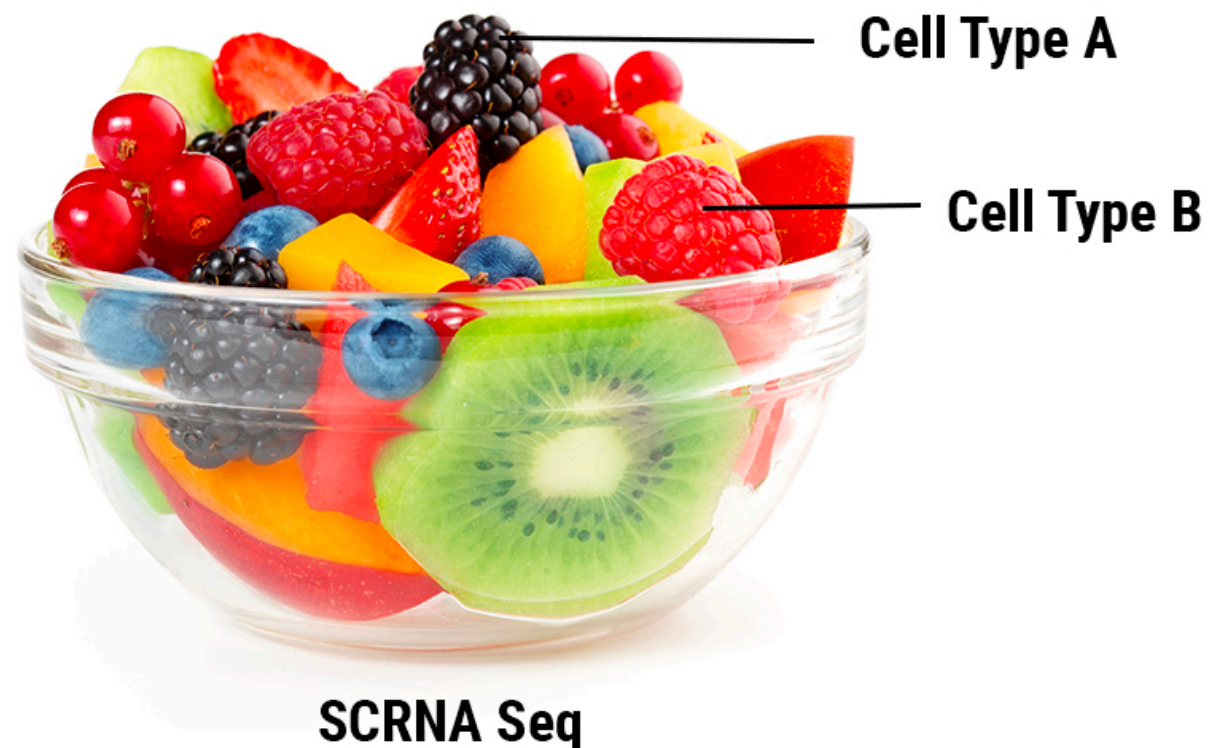
Bulk vs single cell RNA-seq

Bulk RNA sequencing measures gene expression levels averaged across cells in a tissue sample;

Single cell RNA sequencing measures expression of individual cells.

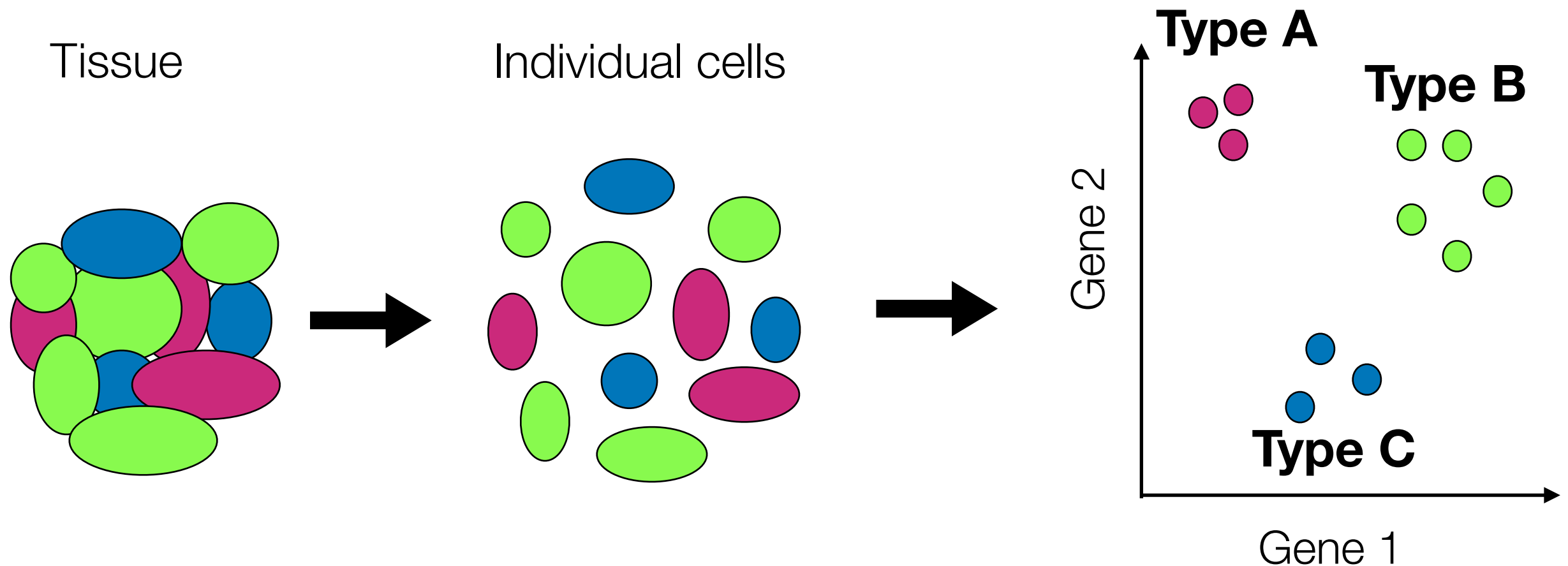


Bulk RNA Seq



SCRNA Seq

A common question in scRNA-seq



Are genes differentially expressed across cell types? Which genes?

What would we like to do?

Given a cell by gene matrix of expression levels $x \in \mathbb{R}^{n \times p}$, and a vector of cell type labels $L \in \{1, 2, \dots, K\}^n$:

For a pair of cell types (k, k') :

Test H_0 : expected expression for all genes are equal in cell type k and cell type k'

- Eg. Could apply a multivariate two-sample t -test to $\{x_i : L_i = k\}$ and $\{x_i : L_i = k'\}$ or equivalently fit multivariate linear model of X on L using x
- Eg. Or fit a multivariate Poisson GLM of X on L

Are genes differentially expressed across cell types?

What would we like to do?

Given a cell by gene matrix of expression levels $x \in \mathbb{R}^{n \times p}$, and a vector of cell type labels $L \in \{1, 2, \dots, K\}^n$:

For a pair of cell types (k, k') and for a gene j :

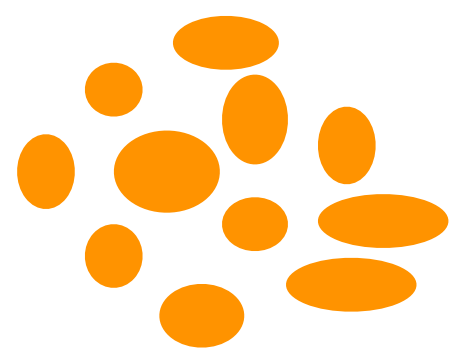
Test H_{0j} : expected expression for gene j are equal in cell type k and cell type k'

- Eg. Could apply a two-sample t -test to $\{x_{ij} : L_i = k\}$ and $\{x_{ij} : L_i = k'\}$ or equivalently fit linear model of X_j on L using x
- Eg. Or fit a Poisson GLM of X_j on L

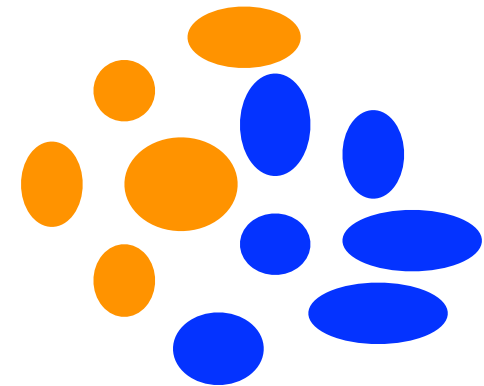
Which genes are differentially expressed across cell types?

How do we get cell types?

Cell type labels for our data are typically *unobserved*.



k-means clustering of
log-transformed data



$$\begin{bmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{bmatrix}$$

In practice, estimate cell types by grouping together cells with similar gene expression measurements via unsupervised learning.

Standard naive data analysis pipeline

Step 1 Estimate unknown cell types L via clustering

$$\text{Eg. } x = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} \xrightarrow[\text{clustering}]{\text{k-means}} \hat{L}(x) = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 1 \end{bmatrix}$$

Step 2 Test for differential expression across groups defined by $\hat{L}(x)$

Eg. Use x to fit a Poisson GLM of X_j on $\hat{L}(x)$, and report standard p-value for the slope coefficient.

The hypothesis is data dependent

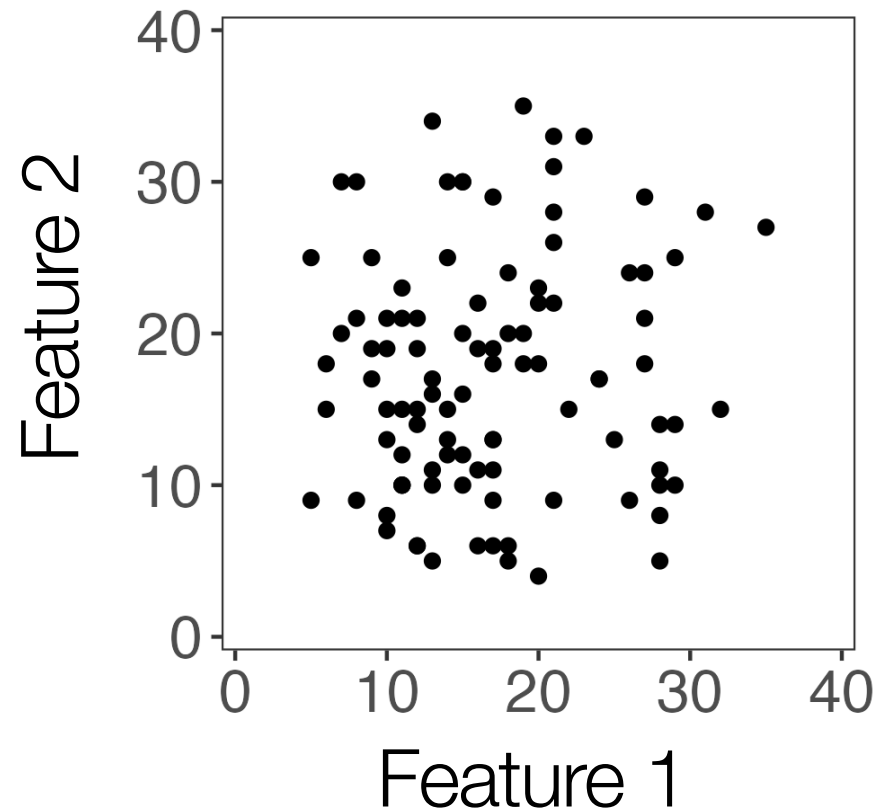
Suppose we observe x , a realization from $X \sim F_X$.

In step 1, we apply a clustering procedure to x to get $\hat{L}(x)$; this is a realization from random variable $\hat{L}(X)$

In step 2, we test $H_{0j}(\hat{L}(x))$: expected expression of gene j is the same across two levels of $\hat{L}(x)$

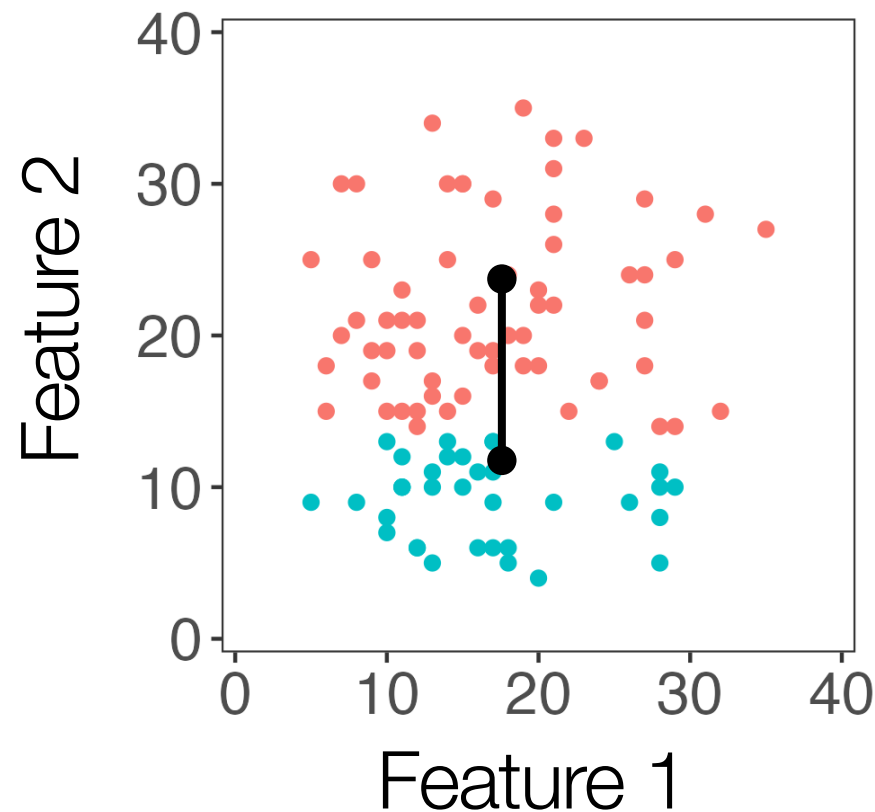
But $\hat{L}(x)$ literally groups cells together that have similar expression values in x ! Circularity gets us yet again ...

Naive p-values are too small



Step 1: Sample 100 cells with $\mathbb{E}[X_i] = \mu$ for all i

Naive p-values are too small



Step 1: Sample 100 cells with $\mathbb{E}[X_i] = \mu$ for each cell i

Step 2: Cluster the observations:

Select $H_{02}(\hat{L}(x))$: “the expected value of Feature 2 is the same between red observations and the blue observations.”

Step 3: Test $H_{02}(\hat{L}(x))$ using x with a two-sample t-test.

$$p < 10^{-10}$$



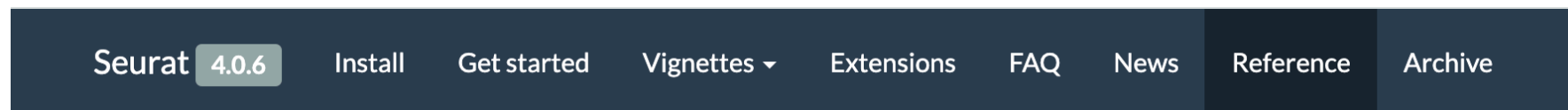
This naive strategy is pervasive

Seurat 4.1.0 Install Get started Vignettes ▾ Extensions FAQ News Reference Archive

TITLE	CITED BY	YEAR
Spatial reconstruction of single-cell gene expression data R Satija, JA Farrell, D Gennert, AF Schier, A Regev Nature biotechnology 33 (5), 495-502	3157	2015
Integrating single-cell transcriptomic data across different conditions, technologies, and species A Butler, P Hoffman, P Smibert, E Papalexi, R Satija Nature biotechnology 36 (5), 411-420	6429	2018
Comprehensive integration of single-cell data T Stuart, A Butler, P Hoffman, C Hafemeister, E Papalexi, WM Mauck, ... Cell 177 (7), 1888-1902. e21	6993	2019
Integrated analysis of multimodal single-cell data Y Hao, S Hao, E Andersen-Nissen, WM Mauck, S Zheng, A Butler, MJ Lee, ... Cell 184 (13), 3573-3587. e29	2432	2021

~ 20,000 citations!

... Even though it doesn't work



Gene expression markers of identity classes

Source: `R/generics.R`, `R/differential_expression.R`

Finds markers (differentially expressed genes) for identity classes

```
FindMarkers(object, ...)
```

Details

p-value adjustment is performed using bonferroni correction based on the total number of genes in the dataset. Other correction methods are not recommended, as Seurat pre-filters genes using the arguments above, reducing the number of tests performed. Lastly, as Aaron Lun has pointed out, p-values should be interpreted cautiously, as the genes used for clustering are the same genes tested for differential expression.

Awkward!!!

A solution is keenly wanted

Lähnemann et al. *Genome Biology* (2020) 21:31
<https://doi.org/10.1186/s13059-020-1926-6>


Genome Biology

REVIEW

Open Access

Eleven grand challenges in single-cell data science



David Lähnemann^{1,2,3}, Johannes Köster^{1,4}, Ewa Szczurek⁵, Davis J. McCarthy^{6,7}, Stephanie C. Hicks⁸, Mark D. Robinson⁹ , Catalina A. Vallejos^{10,11}, Kieran R. Campbell^{12,13,14}, Niko Beerenwinkel^{15,16}, Ahmed Mahfouz^{17,18}, Luca Pinello^{19,20,21}, Pavel Skums²², Alexandros Stamatakis^{23,24}, Camille Stephan-Otto Attolini²⁵, Samuel A. M. Buys de Barbanson^{29,30,31}, Antonio Cappuccinelli³², Maria Florescu^{29,30,31}, Victor Guryev³⁵, Rebecca M. Keizer³⁷, Indu Khatri³⁸, Szymon J. Tzu-Hao Kuo³, Boudewijn P.F. Lelieveldt⁴, Tobias Marschall^{47,48}, Felix Mölder^{1,49}, Anthonie Jeroen de Ridder^{29,30}, Antoine-Emmanuel Fabian J. Theis⁵⁴, Huan Yang⁵⁵, Alex Zelikson⁵⁶, Sohrab P. Shah⁵⁹ and Alexander Schönhuth⁵⁷

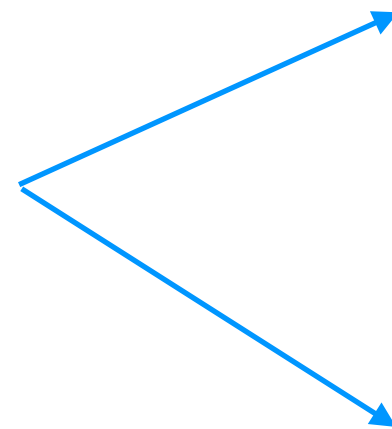
Status

Currently, the vast majority of differential expression detection methods assume that the groups of cells to be compared are known in advance (e.g., experimental conditions or cell types). However, current analysis pipelines typically rely on clustering or cell type assignment to identify such groups, before downstream differential analysis is performed, without propagating the uncertainty in these assignments or accounting for the double use of data (clustering, differential testing between clusters).

Recall: selective inference via sample splitting

Step 1: Estimation

	Gene 1	Gene 2	Gene 3
Cell 1	18	0	23
Cell 2	4	0	5
Cell 3	2	0	0
Cell 4	29	16	32



	Gene 1	Gene 2	Gene 3
Cell 1	18	0	23
Cell 2	4	0	5

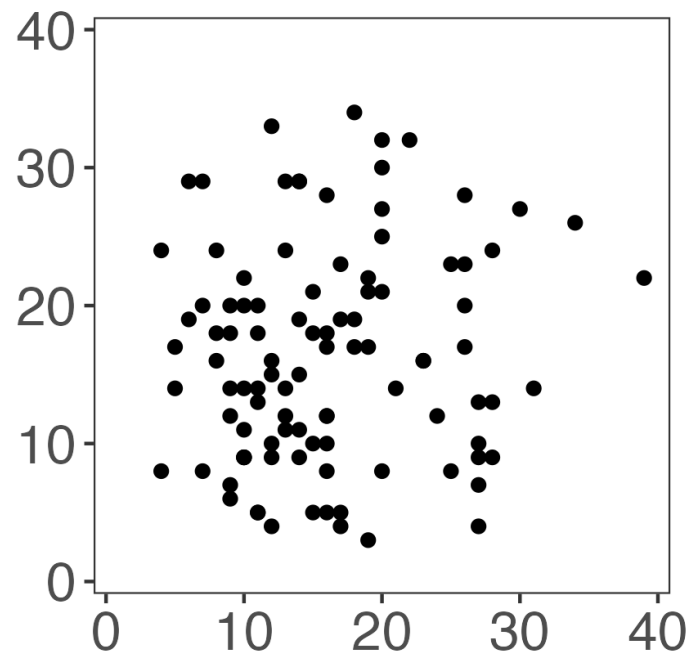
	Gene 1	Gene 2	Gene 3
Cell 3	2	0	0
Cell 4	29	16	32

Step 2: Testing

Last week: “It just works!”

Sample splitting doesn't work (picture)

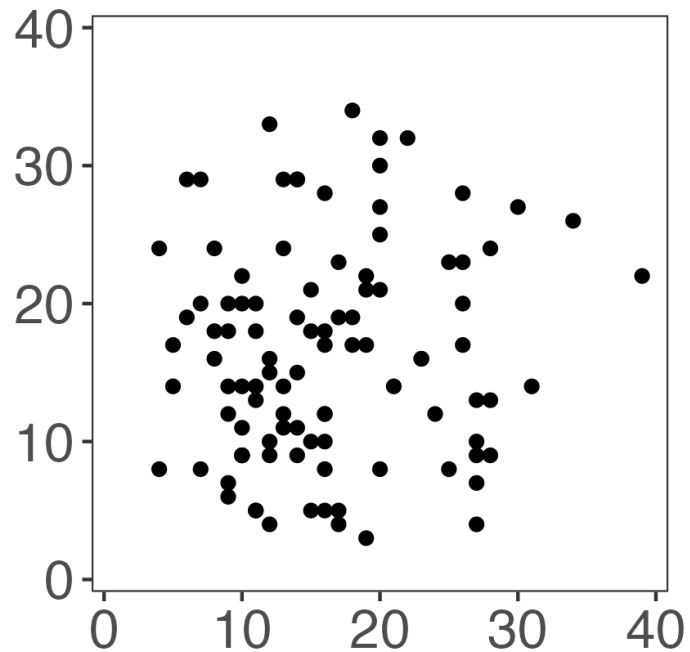
All



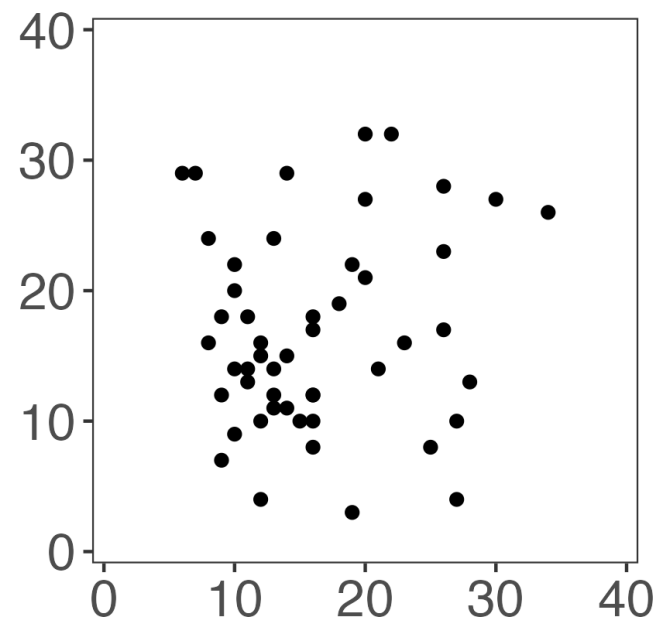
Step 1: Split
cells into train
and test sets.

Sample splitting doesn't work (picture)

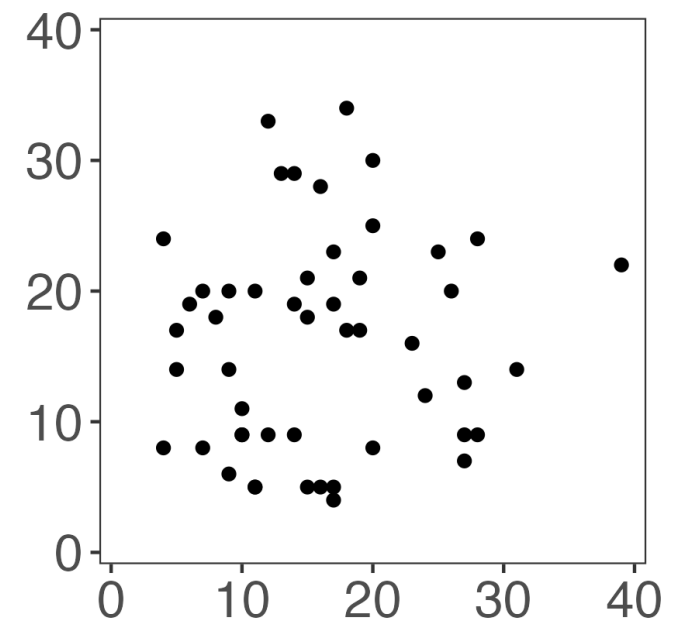
All



Train



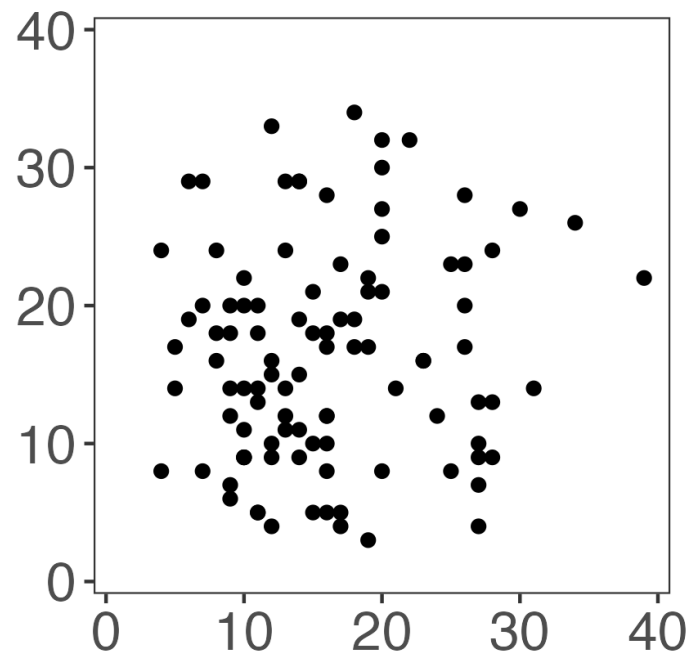
Test



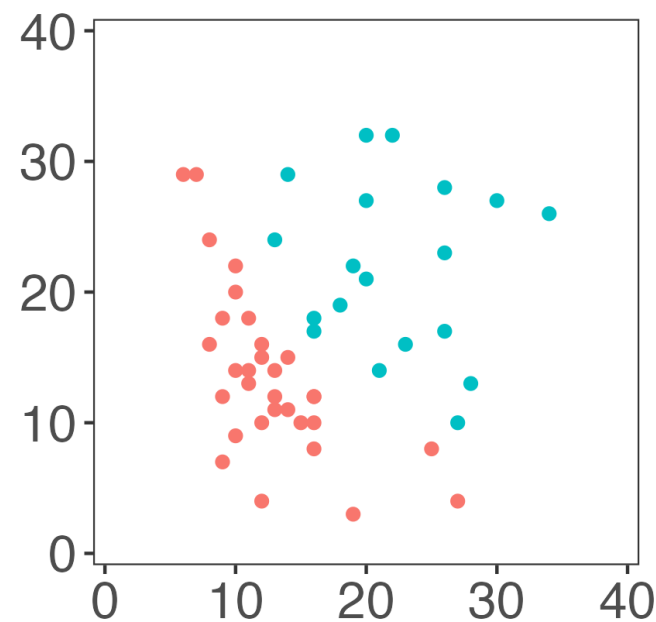
Step 1: Split
cells into train
and test sets.

Sample splitting doesn't work (picture)

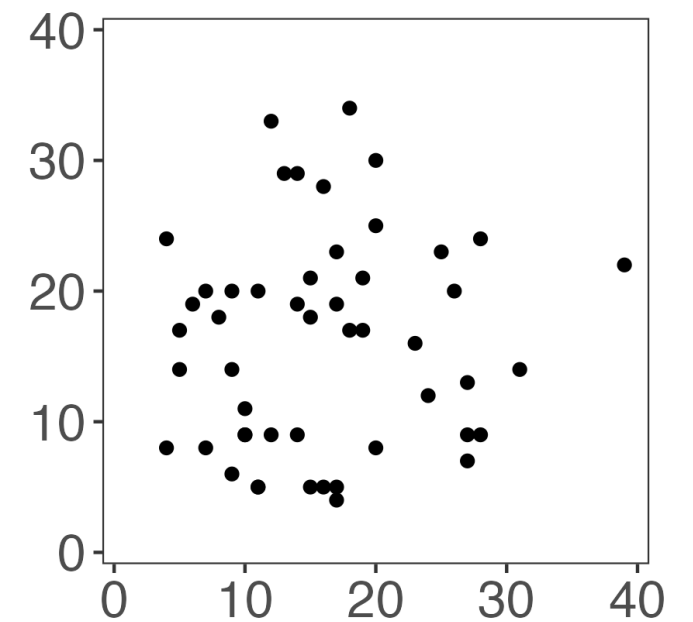
All



Train



Test

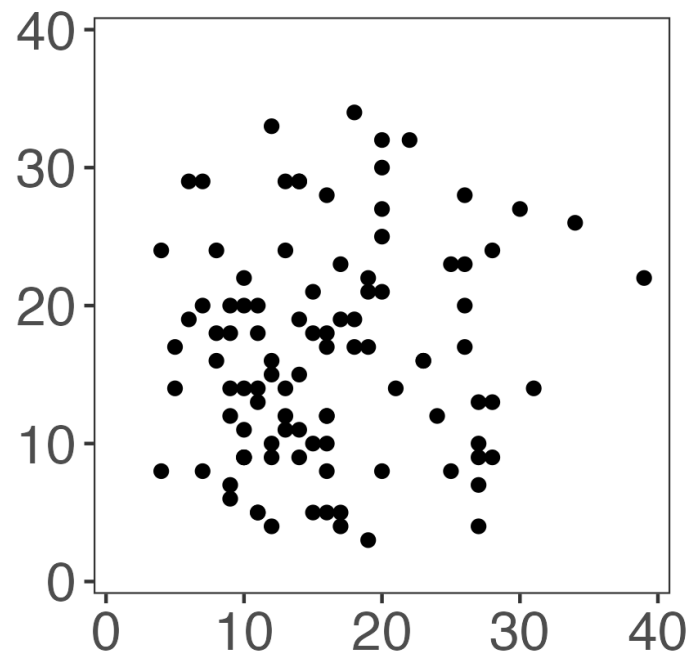


Step 1: Split cells into train and test sets.

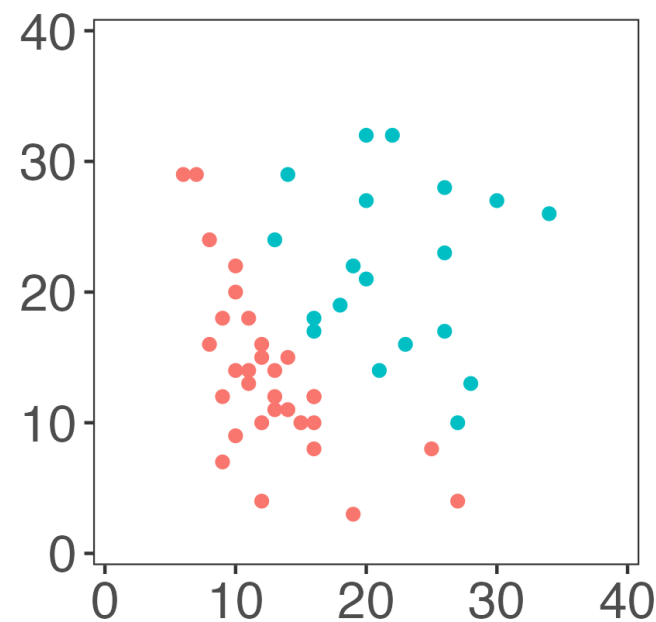
Step 2: Cluster the training set.

Sample splitting doesn't work (picture)

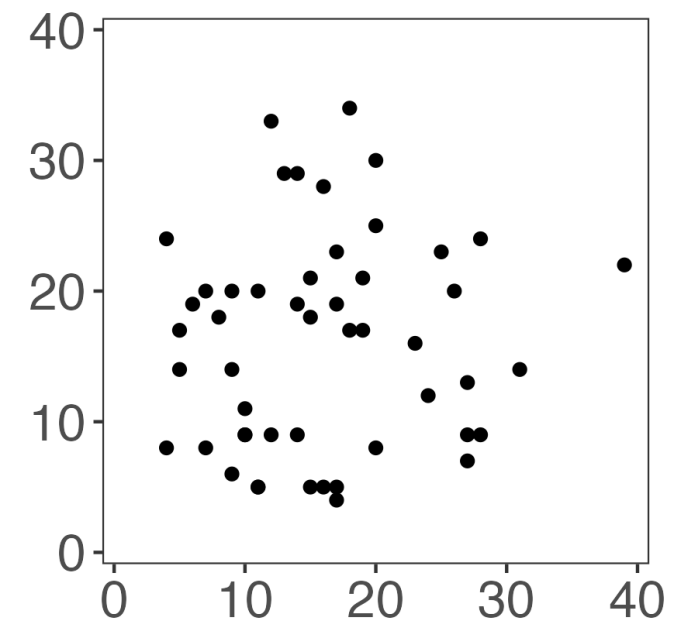
All



Train



Test

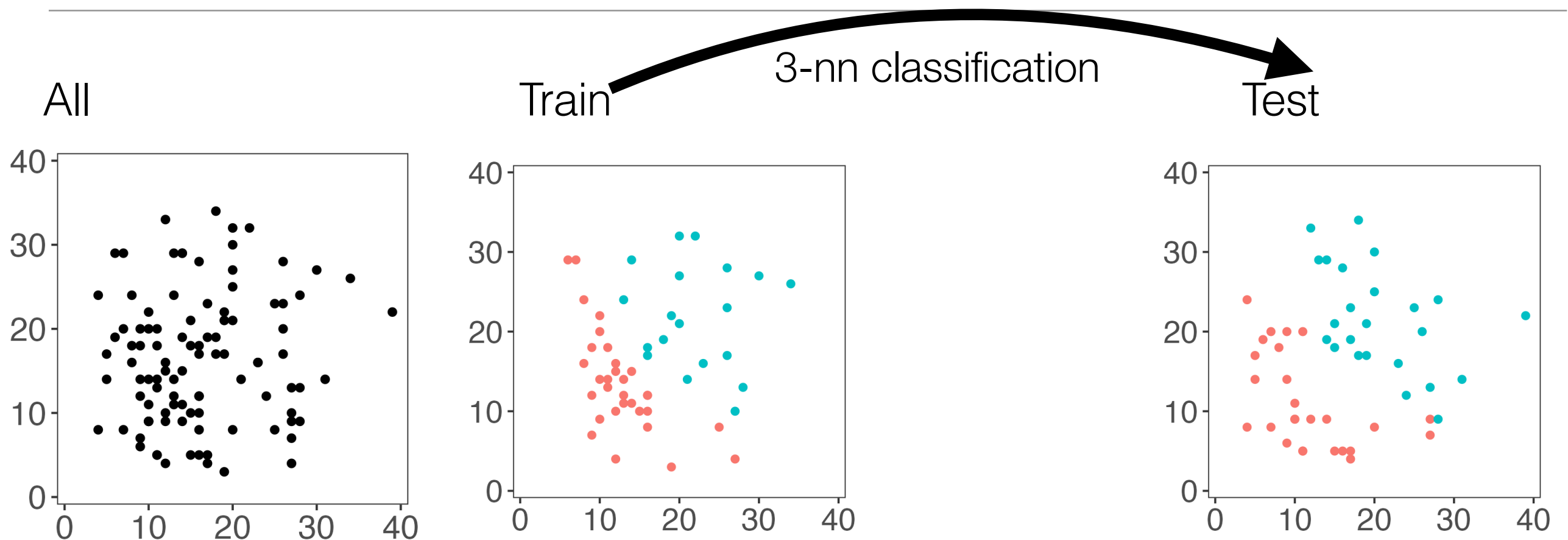


Step 1: Split cells into train and test sets.

Step 2: Cluster the training set.

Step 3: test for difference in means using test set.

Sample splitting doesn't work (picture)



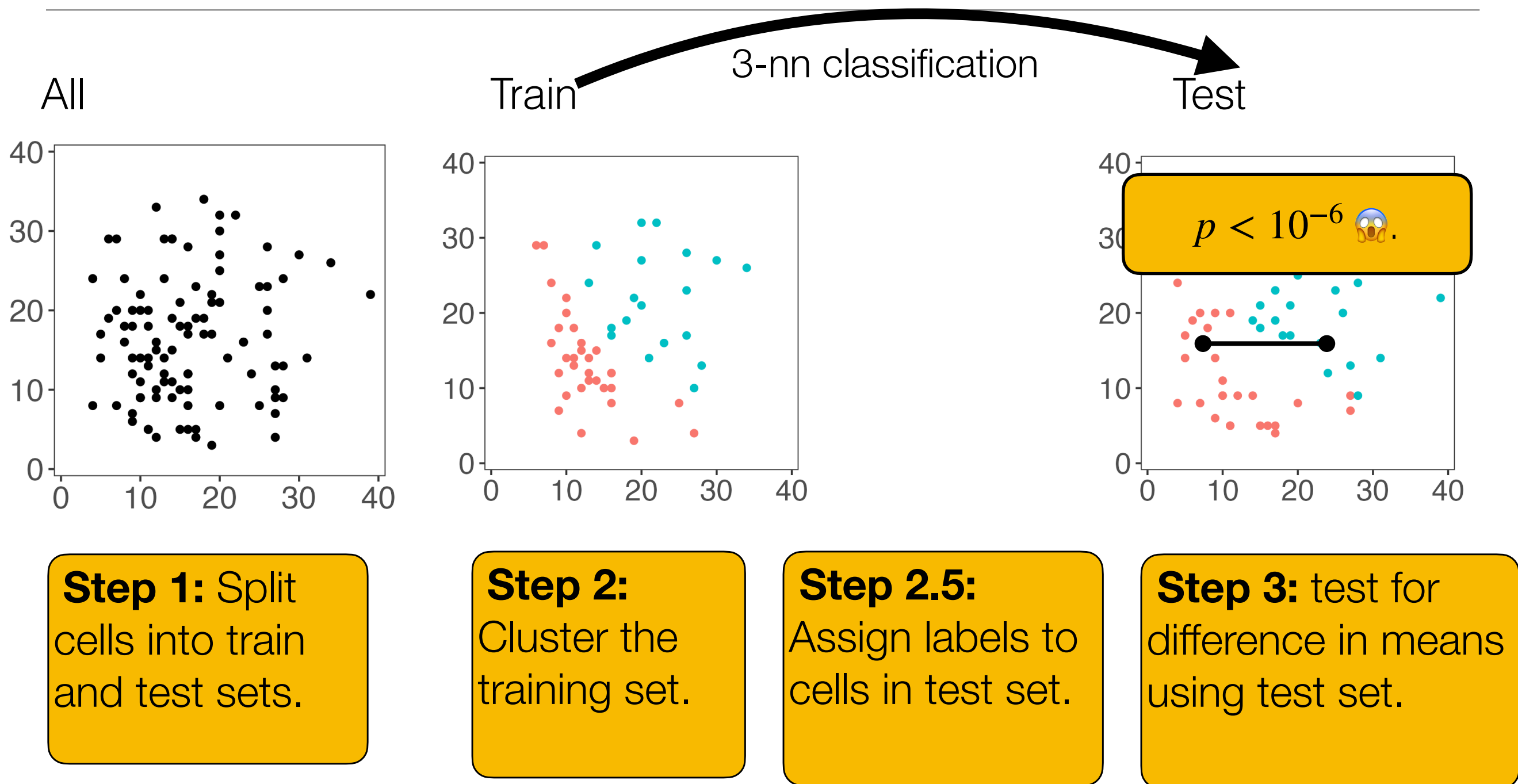
Step 1: Split cells into train and test sets.

Step 2: Cluster the training set.

Step 2.5: Assign labels to cells in test set.

Step 3: test for difference in means using test set.

Sample splitting doesn't work (picture)



Sample splitting doesn't work (math)

We observe x , a realization from $X \sim F_X$.

In step 1, we apply a clustering procedure to x^{tr} to get $\hat{L}(x^{tr})$; this is a realization from random variable $\hat{L}(X^{tr})$

In step 2 and 2.5, we select $H_{0j}(\hat{L}(x^{tr}, x^{te}))$: labels for observations in test set depend on $x^{tr}, \hat{L}(x^{tr})$ (classifier training) and x^{te} (classifier application)

We haven't avoided circularity at all! We select hypothesis and test it using a single data set, x^{tr} .

A subtle drawback of sample splitting

What makes this situation different from inference after variable selection in regression?

After all, they are very similar: we are using the data to “engineer” a categorical covariate to regress on.

In regression, the hypothesis selected on the training set (specific variables) could be transferred to the test set without touching the data because it simply selects columns, and those are “meta-data” shared by train/test

Here, the hypothesis selected on the training set involves rows, and those are *not* shared by train/test, causing problems.

What hypothesis are we trying to test?

Suppose that $X_i \stackrel{ind}{\sim} N_p(\mu_i, \sigma^2 \mathbf{I}_p)$, where σ^2 is known.

Cluster x (a realization from X) to get $\hat{L}(x) = \{\hat{C}_1, \dots, \hat{C}_K\}$, and

let $\bar{\mu}_{\hat{C}_k} = \frac{1}{|\hat{C}_k|} \sum_{i \in \hat{C}_k} \mu_i$ be the “mean” associated with cluster k .

Can we test $H_0 : \bar{\mu}_{\hat{C}_k} = \bar{\mu}_{\hat{C}_{k'}}?$

(Psst: $\{\bar{\mu}_{\hat{C}_k}\}_{k=1}^K$ are (essentially) the targets of “agnostic linear regression” of x on a matrix of dummy variables for $\hat{L}(x)$!)

What are we trying to control?

Want to control the selective type I error rate.

For all fixed non-overlapping subsets of observations

$G, G' \subseteq \{1, 2, \dots, n\}$, if $\bar{\mu}_G = \bar{\mu}_{G'}$ holds, then we want:

$$\mathbb{P} \left(\text{Reject } H_0 : \bar{\mu}_G = \bar{\mu}_{G'} \text{ using } X \mid G, G' \text{ are clusters in } \hat{L}(X) \right) \leq \alpha$$

“If we re-sequence the n cells lots of times, estimating cell types with clustering in each repetition, and only look at the experiments where we called G, G' different cell types, how often would we reject the null hypothesis that G, G' have the same expected expression levels when they really do have the same expected expression levels?”

Conditional selective inference

For all fixed non-overlapping subsets of observations

$G, G' \subseteq \{1, 2, \dots, n\}$, if $\bar{\mu}_G = \bar{\mu}_{G'}$ holds, we want:

$$\mathbb{P} \left(\text{Reject } H_0 : \bar{\mu}_G = \bar{\mu}_{G'} \text{ using } X \mid G, G' \text{ are clusters in } \hat{L}(X) \right) \leq \alpha$$

Conditional selective inference does this by characterizing the conditional distribution of:

Test statistic for $H_0 : \bar{\mu}_G = \bar{\mu}_{G'} \mid G, G' \text{ are clusters in } \hat{L}(X)$

By probability integral transform, produces valid inference.

Multivariate Z -test of $H_0 : \bar{\mu}_G = \bar{\mu}_{G'}$

Suppose $G, G' \subseteq \{1, 2, \dots, n\}$ are two fixed non-overlapping subsets of observations.

Test $H_0 : \bar{\mu}_G = \bar{\mu}_{G'}$ using X :

$$\mathbb{P}_{H_0} \left(\frac{\|\bar{X}_G - \bar{X}_{G'}\|_2^2}{\sigma^2(1/|G| + 1/|G'|)} \geq \frac{\|\bar{x}_G - \bar{x}_{G'}\|_2^2}{\sigma^2(1/|G| + 1/|G'|)} \right)$$

$$X_i \stackrel{\text{ind}}{\sim} N_p(\mu_i, \sigma^2 \mathbf{I}_p)$$

$$= \mathbb{P} \left(\chi_p^2 \geq \frac{\|\bar{x}_G - \bar{x}_{G'}\|_2^2}{\sigma^2(1/|G| + 1/|G'|)} \right).$$

Conditional selective inference framework

The idea is to compute the conditional p-value,

$$\mathbb{P} \left(\|\bar{X}_{\hat{C}_k} - \bar{X}_{\hat{C}_{k'}}\|_2 \geq \|\bar{x}_{\hat{C}_k} - \bar{x}_{\hat{C}_{k'}}\|_2 \mid \text{Clustering } X \text{ results in } \hat{C}_k, \hat{C}_{k'} \right),$$

taking the probability under $H_0 : \bar{\mu}_{\hat{C}_k} = \bar{\mu}_{\hat{C}_{k'}}$.

We can't do that because of nuisance parameters. So instead:

$$\mathbb{P}_{H_0} \left(\|\bar{X}_{\hat{C}_k} - \bar{X}_{\hat{C}_{k'}}\|_2 \geq \|\bar{x}_{\hat{C}_k} - \bar{x}_{\hat{C}_{k'}}\|_2 \mid \begin{array}{l} \text{Clustering } X \text{ results in } \hat{C}_k, \hat{C}_{k'}, \\ \text{dir} \left(\bar{X}_{\hat{C}_k} - \bar{X}_{\hat{C}_{k'}} \right) = \text{dir} \left(\bar{x}_{\hat{C}_k} - \bar{x}_{\hat{C}_{k'}} \right), \\ (\mathbf{I}_n - \hat{\nu}(\hat{\nu}^T \hat{\nu})^{-1} \hat{\nu}^T) X = (\mathbf{I}_n - \hat{\nu}(\hat{\nu}^T \hat{\nu})^{-1} \hat{\nu}^T) x \end{array} \right)$$

Definition: $\hat{\nu}$ satisfies $X^T \hat{\nu} = \bar{X}_{\hat{C}_k} - \bar{X}_{\hat{C}_{k'}}$

Importantly, conditioning on “more” still maintains selective Type 1 error control. (Homework)

What does the “more” get us?

Definition: $\hat{\nu}$ satisfies
 $\mathbf{X}^T \hat{\nu} = \bar{X}_{\hat{C}_k} - \bar{X}_{\hat{C}_{k'}}$

$$\mathbb{P}_{H_0} \left(\|\bar{X}_{\hat{C}_k} - \bar{X}_{\hat{C}_{k'}}\|_2 \geq \|\bar{x}_{\hat{C}_k} - \bar{x}_{\hat{C}_{k'}}\|_2 \mid \begin{array}{l} \text{Clustering } \mathbf{X} \text{ results in } \hat{C}_k, \hat{C}_{k'}, \\ \text{dir}(\bar{X}_{\hat{C}_k} - \bar{X}_{\hat{C}_{k'}}) = \text{dir}(\bar{x}_{\hat{C}_k} - \bar{x}_{\hat{C}_{k'}}), \\ (\mathbf{I}_n - \hat{\nu}(\hat{\nu}^T \hat{\nu})^{-1} \hat{\nu}^T) \mathbf{X} = (\mathbf{I}_n - \hat{\nu}(\hat{\nu}^T \hat{\nu})^{-1} \hat{\nu}^T) \mathbf{x} \end{array} \right)$$

$$= \mathbb{P}_{H_0} \left(\phi \geq \|\bar{x}_{\hat{C}_k} - \bar{x}_{\hat{C}_{k'}}\|_2 \mid \text{Clustering } \mathbf{x}'(\phi) \text{ results in } \hat{C}_k, \hat{C}_{k'} \right),$$

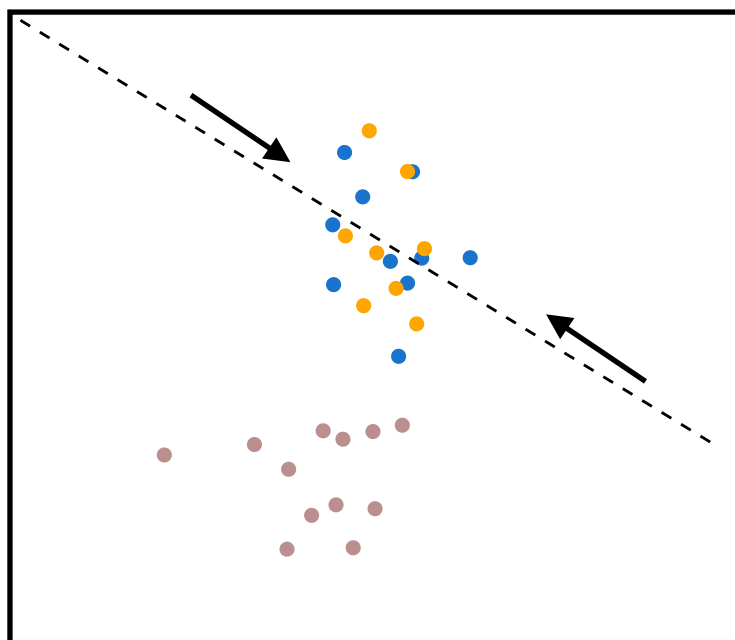
where $\phi = \|\bar{X}_{\hat{C}_k} - \bar{X}_{\hat{C}_{k'}}\|_2$, and $\mathbf{x}'(\phi)$ is an $n \times p$ data set with rows

$$[\mathbf{x}'(\phi)]_i = \begin{cases} x_i + \left(\frac{|\hat{C}_2|}{|\hat{C}_1| + |\hat{C}_2|} \right) \left(\phi - \|\bar{x}_{\hat{C}_k} - \bar{x}_{\hat{C}_{k'}}\|_2 \right) \text{dir}(\bar{x}_{\hat{C}_k} - \bar{x}_{\hat{C}_{k'}}), & \text{if } i \in \hat{C}_k \\ x_i - \left(\frac{|\hat{C}_1|}{|\hat{C}_1| + |\hat{C}_2|} \right) \left(\phi - \|\bar{x}_{\hat{C}_k} - \bar{x}_{\hat{C}_{k'}}\|_2 \right) \text{dir}(\bar{x}_{\hat{C}_k} - \bar{x}_{\hat{C}_{k'}}), & \text{if } i \in \hat{C}_{k'} \\ x_i, & \text{otherwise} \end{cases}$$

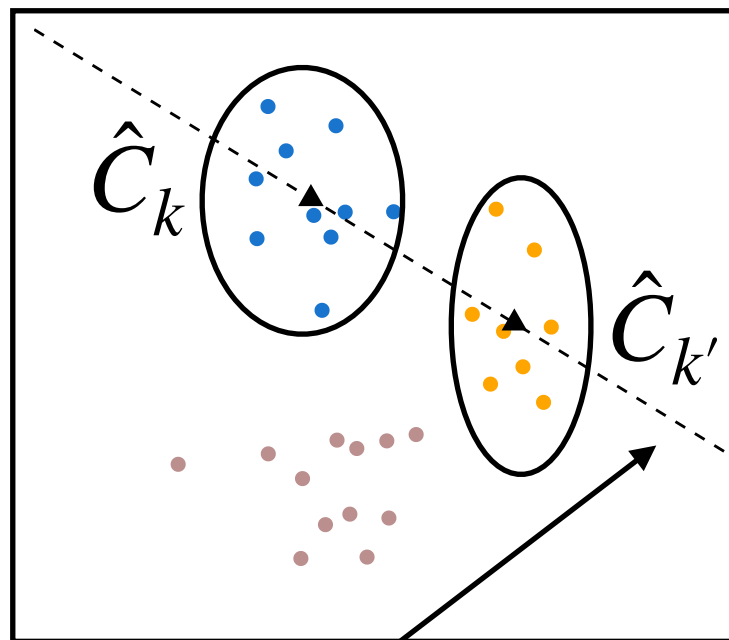
Illustrating $\mathbf{x}'(\phi)$

$$[\mathbf{x}'(\phi)]_i = \begin{cases} x_i + \left(\frac{|\hat{C}_2|}{|\hat{C}_1| + |\hat{C}_2|} \right) \left(\phi - \|\bar{x}_{\hat{C}_k} - \bar{x}_{\hat{C}_{k'}}\|_2 \right) \text{dir} \left(\bar{x}_{\hat{C}_k} - \bar{x}_{\hat{C}_{k'}} \right), & \text{if } i \in \hat{C}_k \\ x_i - \left(\frac{|\hat{C}_1|}{|\hat{C}_1| + |\hat{C}_2|} \right) \left(\phi - \|\bar{x}_{\hat{C}_k} - \bar{x}_{\hat{C}_{k'}}\|_2 \right) \text{dir} \left(\bar{x}_{\hat{C}_k} - \bar{x}_{\hat{C}_{k'}} \right), & \text{if } i \in \hat{C}_{k'} \\ x_i, & \text{otherwise} \end{cases}$$

$\mathbf{x}'(0)$

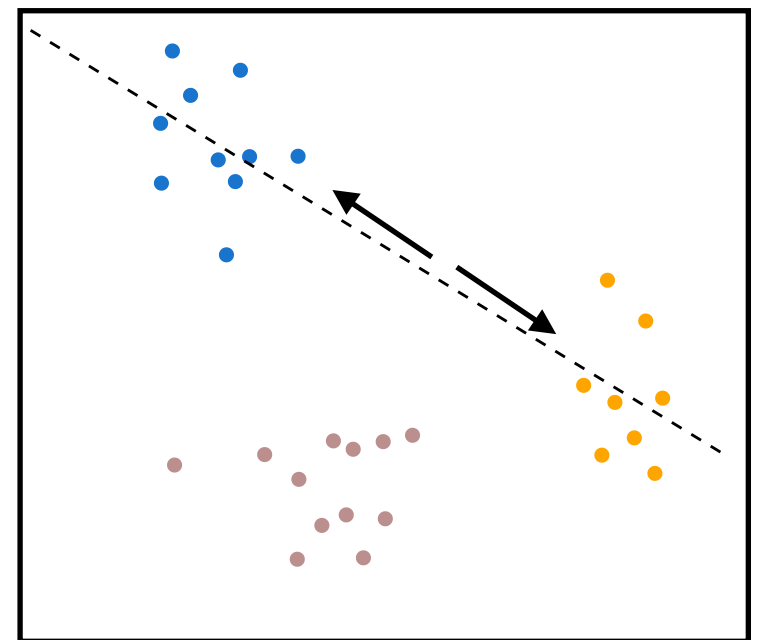


$\mathbf{x} = \mathbf{x}'(4)$



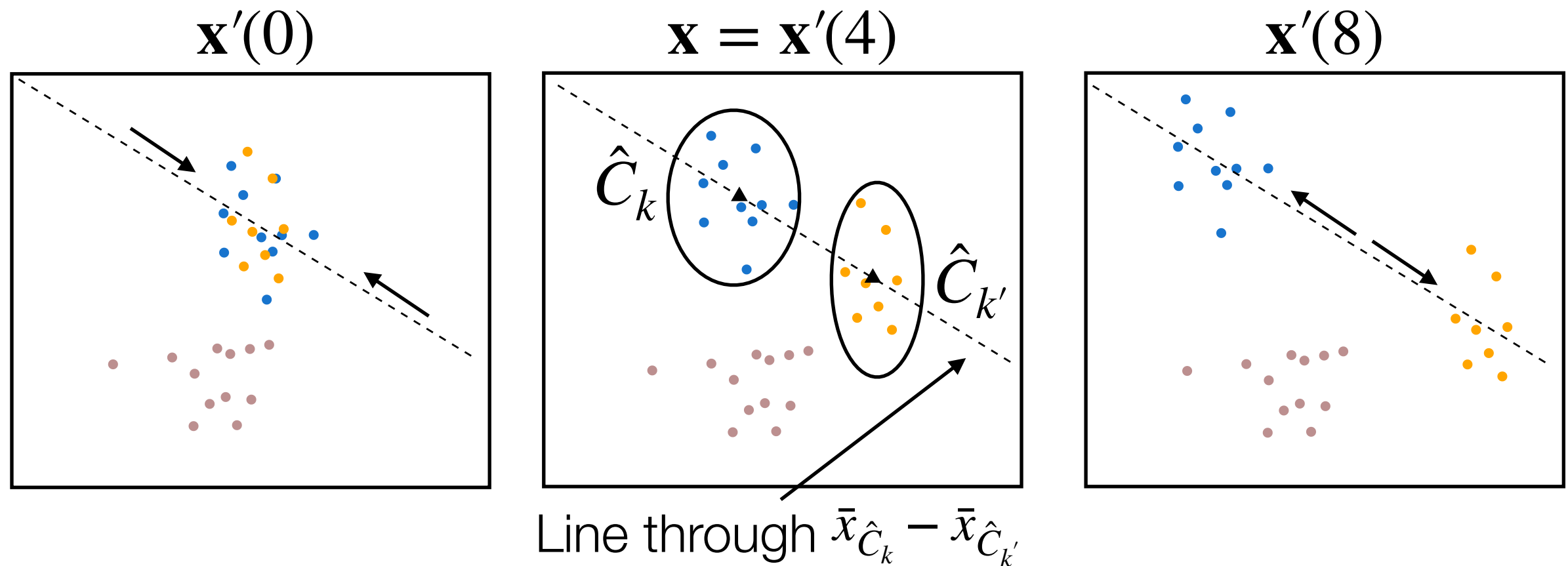
Line through $\bar{x}_{\hat{C}_k} - \bar{x}_{\hat{C}_{k'}}$

$\mathbf{x}'(8)$



Power loss through conditioning

Adding “extra stuff” to the conditioning set ignores all data realizations that don’t look like this:



That’s a lot of data sets to rule out for purely computational reasons! Power loss could be substantial.

Computing the p-value for conditional SI

$$\mathbb{P}_{H_0} \left(\|\bar{X}_{\hat{C}_k} - \bar{X}_{\hat{C}_{k'}}\|_2 \geq \|\bar{x}_{\hat{C}_k} - \bar{x}_{\hat{C}_{k'}}\|_2 \mid \begin{array}{l} \text{Clustering } \mathbf{X} \text{ results in } \hat{C}_k, \hat{C}_{k'}, \\ \text{dir} \left(\bar{X}_{\hat{C}_k} - \bar{X}_{\hat{C}_{k'}} \right) = \text{dir} \left(\bar{x}_{\hat{C}_k} - \bar{x}_{\hat{C}_{k'}} \right), \\ (\mathbf{I}_n - \hat{\nu}(\hat{\nu}^T \hat{\nu})^{-1} \hat{\nu}^T) \mathbf{X} = (\mathbf{I}_n - \hat{\nu}(\hat{\nu}^T \hat{\nu})^{-1} \hat{\nu}^T) \mathbf{x} \end{array} \right)$$

$$= \mathbb{P}_{H_0} \left(\phi \geq \|\bar{x}_{\hat{C}_k} - \bar{x}_{\hat{C}_{k'}}\|_2 \mid \text{Clustering } \mathbf{x}'(\phi) \text{ results in } \hat{C}_k, \hat{C}_{k'} \right),$$

$$\text{where } \phi = \|\bar{X}_{\hat{C}_k} - \bar{X}_{\hat{C}_{k'}}\|_2$$

Computing the p-value for conditional SI

$$\mathbb{P}_{H_0} \left(\|\bar{X}_{\hat{C}_k} - \bar{X}_{\hat{C}_{k'}}\|_2 \geq \|\bar{x}_{\hat{C}_k} - \bar{x}_{\hat{C}_{k'}}\|_2 \mid \begin{array}{l} \text{Clustering } \mathbf{X} \text{ results in } \hat{C}_k, \hat{C}_{k'}, \\ \text{dir}(\bar{X}_{\hat{C}_k} - \bar{X}_{\hat{C}_{k'}}) = \text{dir}(\bar{x}_{\hat{C}_k} - \bar{x}_{\hat{C}_{k'}}), \\ (\mathbf{I}_n - \hat{\nu}(\hat{\nu}^T \hat{\nu})^{-1} \hat{\nu}^T) \mathbf{X} = (\mathbf{I}_n - \hat{\nu}(\hat{\nu}^T \hat{\nu})^{-1} \hat{\nu}^T) \mathbf{x} \end{array} \right)$$

$$= \mathbb{P}_{H_0} \left(\phi \geq \|\bar{x}_{\hat{C}_k} - \bar{x}_{\hat{C}_{k'}}\|_2 \mid \text{Clustering } \mathbf{x}'(\phi) \text{ results in } \hat{C}_k, \hat{C}_{k'} \right),$$

$$\text{where } \phi = \|\bar{X}_{\hat{C}_k} - \bar{X}_{\hat{C}_{k'}}\|_2 \stackrel{H_0}{\sim} \sigma \sqrt{1/|\hat{C}_k| + 1/|\hat{C}_{k'}|} \cdot \chi_p$$

Computing the p-value for conditional SI

$$\mathbb{P}_{H_0} \left(\|\bar{X}_{\hat{C}_k} - \bar{X}_{\hat{C}_{k'}}\|_2 \geq \|\bar{x}_{\hat{C}_k} - \bar{x}_{\hat{C}_{k'}}\|_2 \mid \begin{array}{l} \text{Clustering } \mathbf{X} \text{ results in } \hat{C}_k, \hat{C}_{k'}, \\ \text{dir}(\bar{X}_{\hat{C}_k} - \bar{X}_{\hat{C}_{k'}}) = \text{dir}(\bar{x}_{\hat{C}_k} - \bar{x}_{\hat{C}_{k'}}), \\ (\mathbf{I}_n - \hat{\nu}(\hat{\nu}^T \hat{\nu})^{-1} \hat{\nu}^T) \mathbf{X} = (\mathbf{I}_n - \hat{\nu}(\hat{\nu}^T \hat{\nu})^{-1} \hat{\nu}^T) \mathbf{x} \end{array} \right)$$

$$= \mathbb{P}_{H_0} \left(\phi \geq \|\bar{x}_{\hat{C}_k} - \bar{x}_{\hat{C}_{k'}}\|_2 \mid \text{Clustering } \mathbf{x}'(\phi) \text{ results in } \hat{C}_k, \hat{C}_{k'} \right),$$

$$= 1 - \mathbb{F}_{\mathcal{S}} \left(\|\bar{x}_{\hat{C}_k} - \bar{x}_{\hat{C}_{k'}}\|_2 \right), \text{ where } \mathbb{F}_{\mathcal{S}} \text{ is the C.D.F. of a } \sigma \|\hat{\nu}\|_2 \cdot \chi_p$$

random variable truncated to the set

$$\mathcal{S} = \{ \phi \geq 0 : \text{Clustering } \mathbf{x}'(\phi) \text{ results in } \hat{C}_k, \hat{C}_{k'} \}.$$

Only one question left: How do we compute \mathcal{S} ?

Computing the p-value for conditional SI

Strategy for calculating conditioning set \mathcal{S} depends heavily on algorithm used for clustering.

For hierarchical clustering with squared Euclidean distance and most linkages:

<https://www.lucylgao.com/clusterpval/>

Also includes a brute force approximation to the p-value for any clustering method.

For k-means clustering: <https://github.com/yiqunchen/KmeansInference>

For single-gene testing: <https://github.com/yiqunchen/CADET>

Conditional SI for clustering

Strengths:

- Controls selective type I error rate and attains nominal selective coverage
- Get to use all of the available data for clustering and inference

Issues:

- Need to assume data is independently Gaussian
- Need to assume that covariance matrix is known (VERY VERY hard to estimate in practice)
- Lose power through additional conditioning to eliminate nuisance parameters
- Either need to use specific clustering method for which there is a paper, or live with slow approximate brute force computation
- Unclear how much information is left after clustering for inference

Alternatives to condition SI?

Sadly, can't use sample splitting!! It doesn't work.

But stay tuned for Ivy's presentation next Tuesday!