Inference after human genetic clustering

Javier González-Delgado

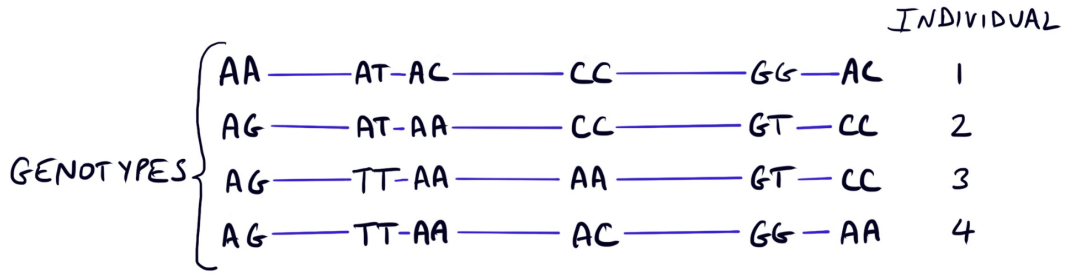
University of Rennes, ENSAI, CREST

MASCOT-NUM 2026

April 2, 2026

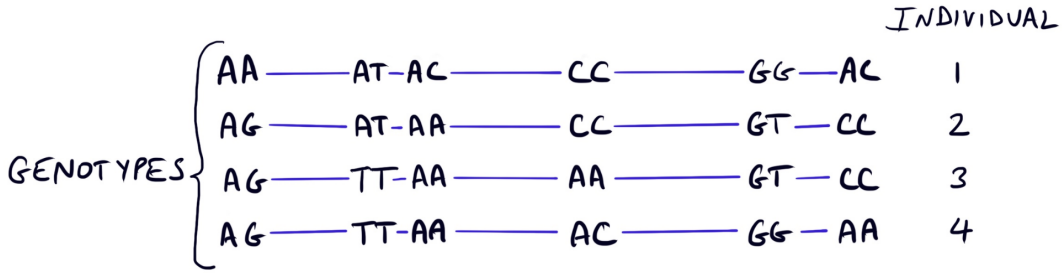
Genetic diversity and population structure

SNP data describing genotypes in a population



Figures : *An introduction to human population genetics, variation, and disease*, Jonathan K Pritchard.

SNP data describing genotypes in a population

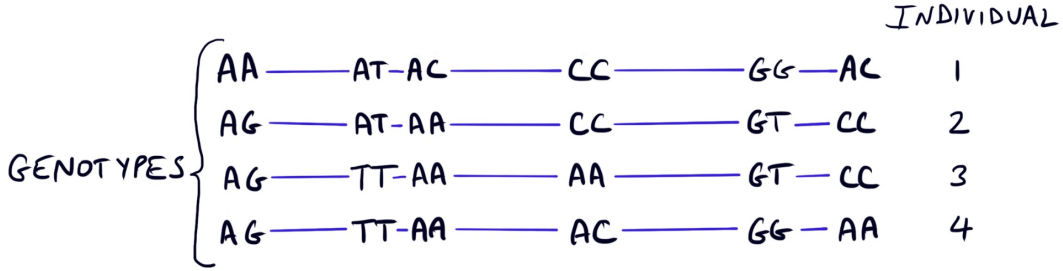


Allele configurations $\rightarrow$ SNP code

AA $\rightarrow$ 0, Aa $\rightarrow$ 1, aa $\rightarrow$ 2,
with A (resp. a) being the reference (resp. alternate) allele.

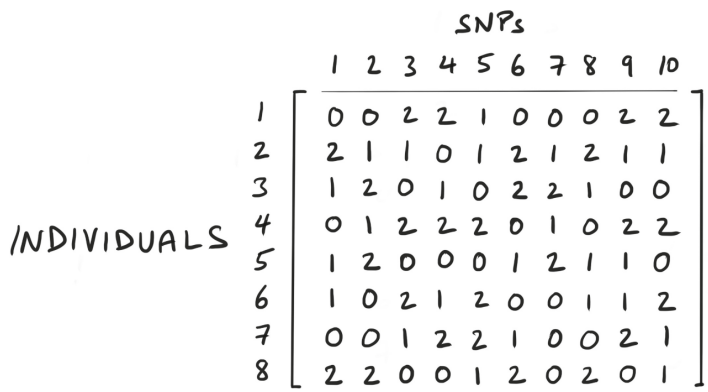
Figures : *An introduction to human population genetics, variation, and disease*, Jonathan K Pritchard.

SNP data describing genotypes in a population



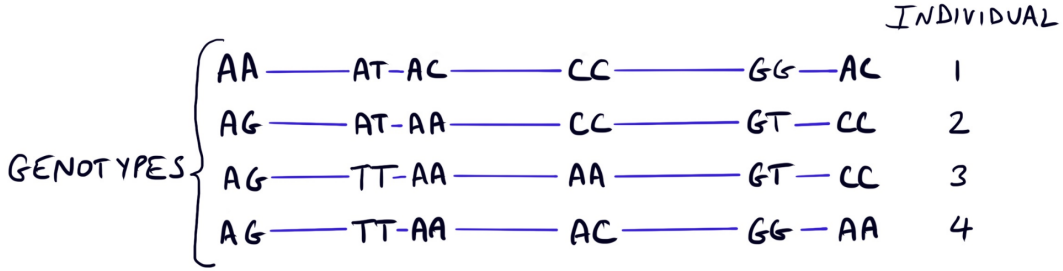
Allele configurations → SNP code

AA → 0, Aa → 1, aa → 2,
with A (resp. a) being the reference (resp. alternate) allele.



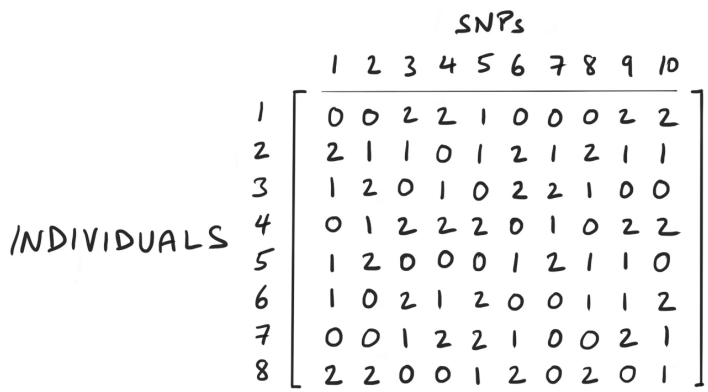
Figures : *An introduction to human population genetics, variation, and disease*, Jonathan K Pritchard.

SNP data describing genotypes in a population



Allele configurations → SNP code

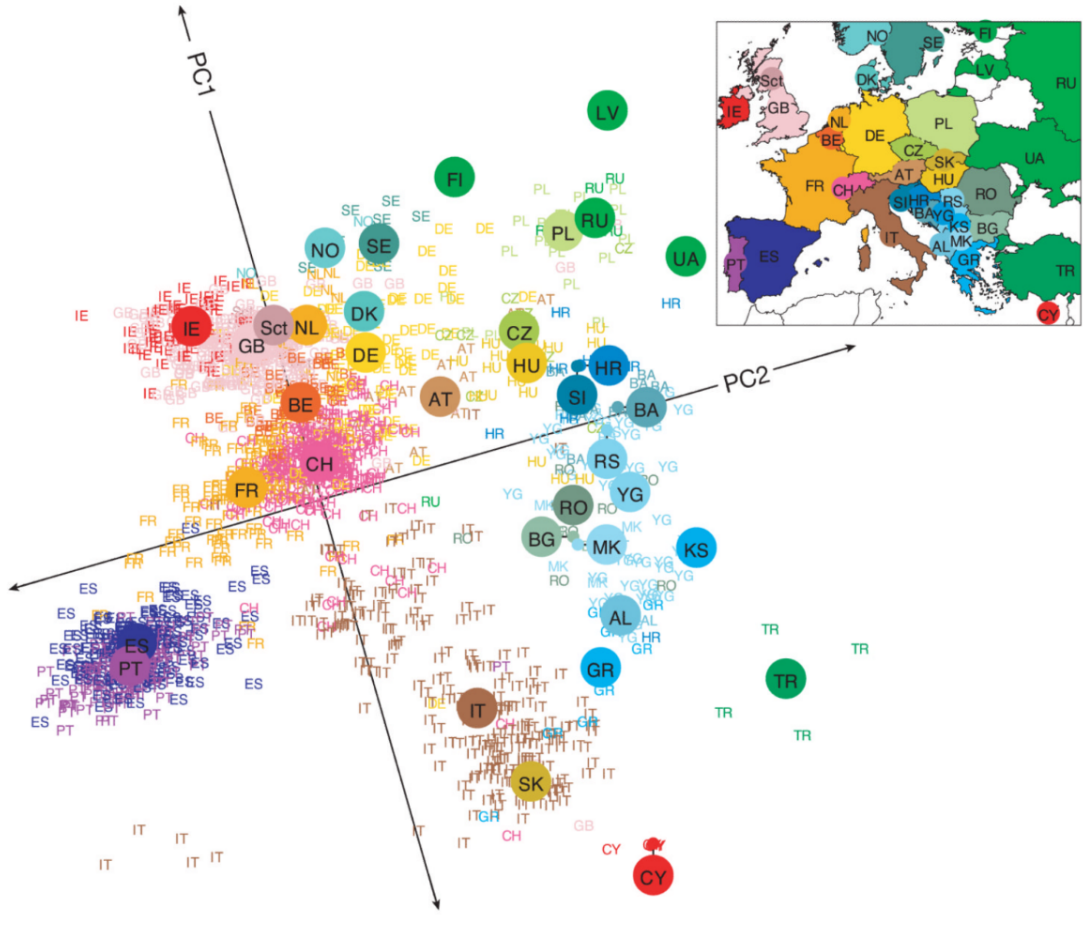
AA → 0, Aa → 1, aa → 2,
with A (resp. a) being the reference (resp. alternate) allele.



→ Is it possible to **find structure in that genotype matrix** ?

Figures : *An introduction to human population genetics, variation, and disease*, Jonathan K Pritchard.

Genotype data unveils population structure



PCA on genome-wide SNP data for 1387 European individuals (Novembre *et al.*, Nature, 2008).

Population structure and genotype data

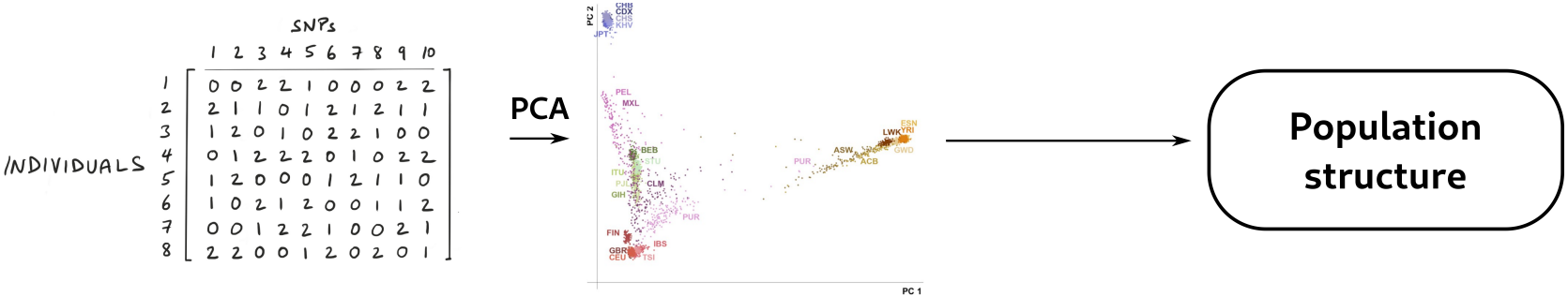
Fundamental question in Population Genetics

Can we infer/observe population structure and/or ancestry from genetic data ?

Population structure and genotype data

Fundamental question in Population Genetics

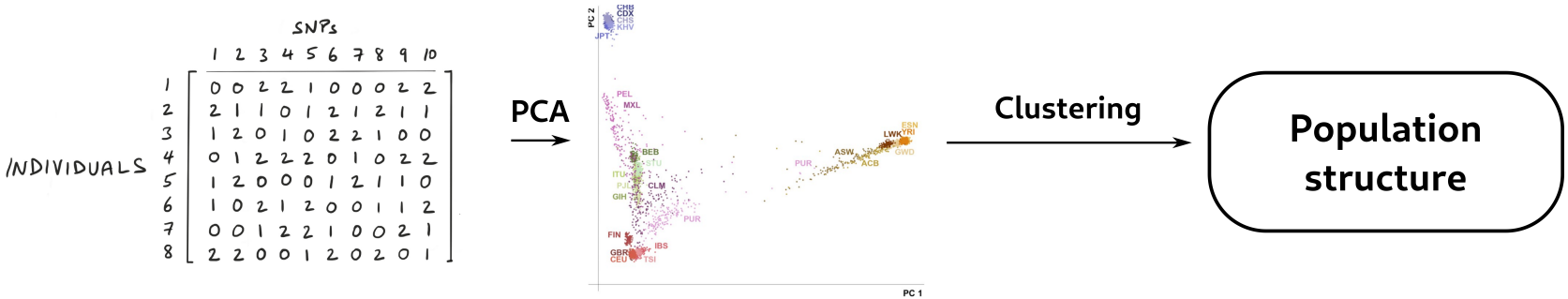
Can we infer/observe population structure and/or ancestry from genetic data ?



Population structure and genotype data

Fundamental question in Population Genetics

Can we infer/observe population structure and/or ancestry from genetic data ?



Population structure and genotype data

Fundamental question in Population Genetics

Can we infer/observe population structure and/or ancestry from genetic data ?

- Simple clustering algorithms like hierarchical clustering reveal broad (continental) structure[1].

1. Mountain and Cavalli-Sforza (1997).

Population structure and genotype data

Fundamental question in Population Genetics

Can we infer/observe population structure and/or ancestry from genetic data ?

- Simple clustering algorithms like hierarchical clustering reveal broad (continental) structure[1].
- More complex algorithms can reveal population structure at finer levels[2] (e.g. the UMAP+HDBSCAN($\hat{\epsilon}$) pipeline[3]).

1. Mountain and Cavalli-Sforza (1997), 2. Rosenberg *et al.* (2002), Hollfelder *et al.* (2017), 3. Diaz-Papkovich *et al.* (2025).

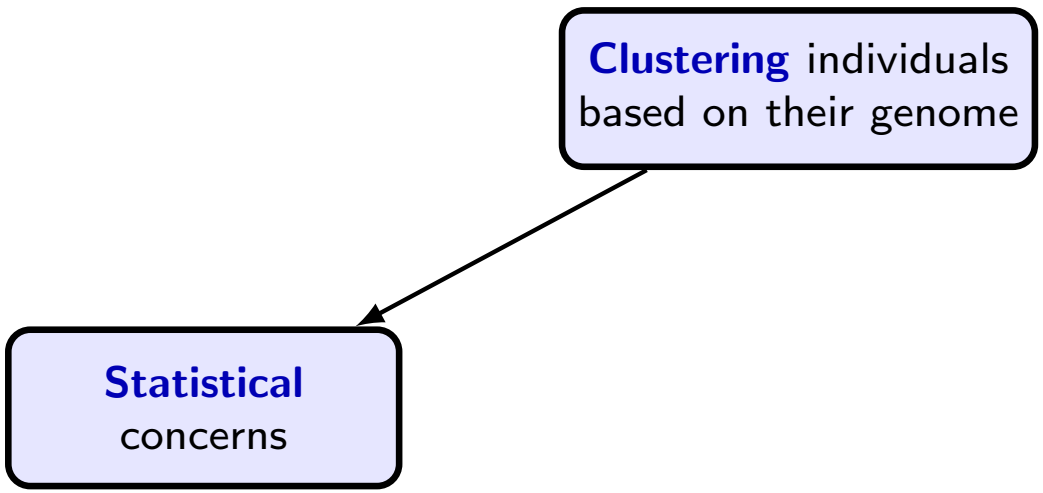
Clustering genetic data

Main challenges

Clustering individuals based on their genome

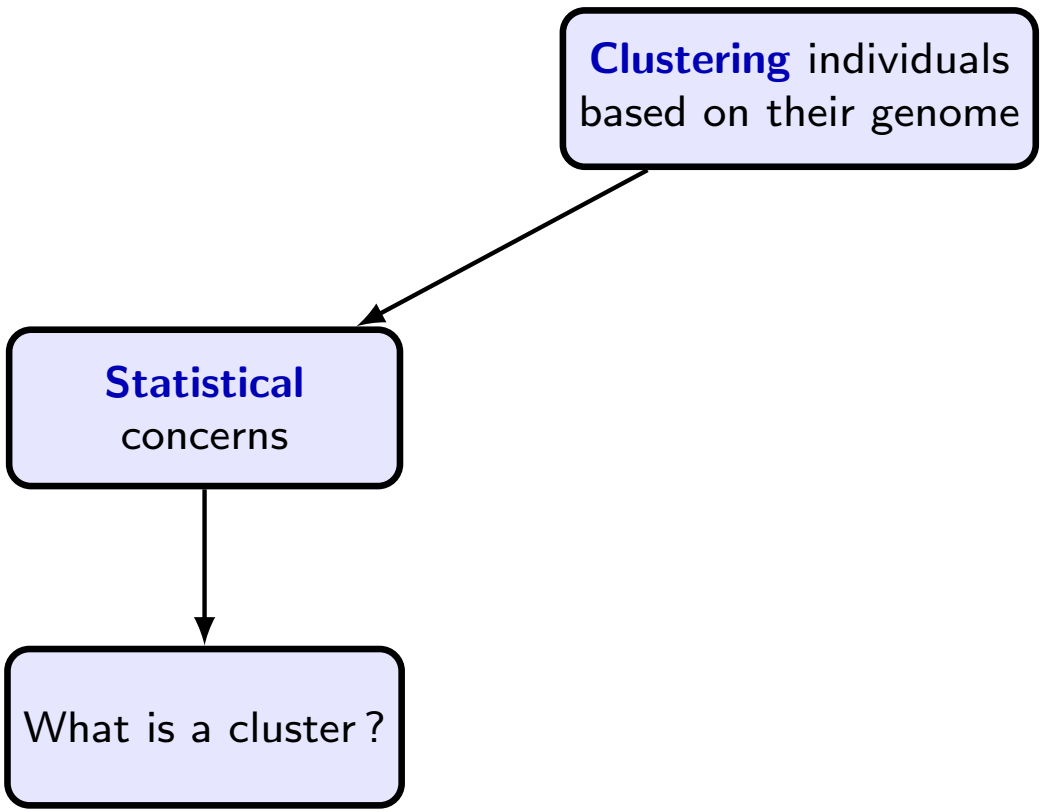
Clustering genetic data

Main challenges



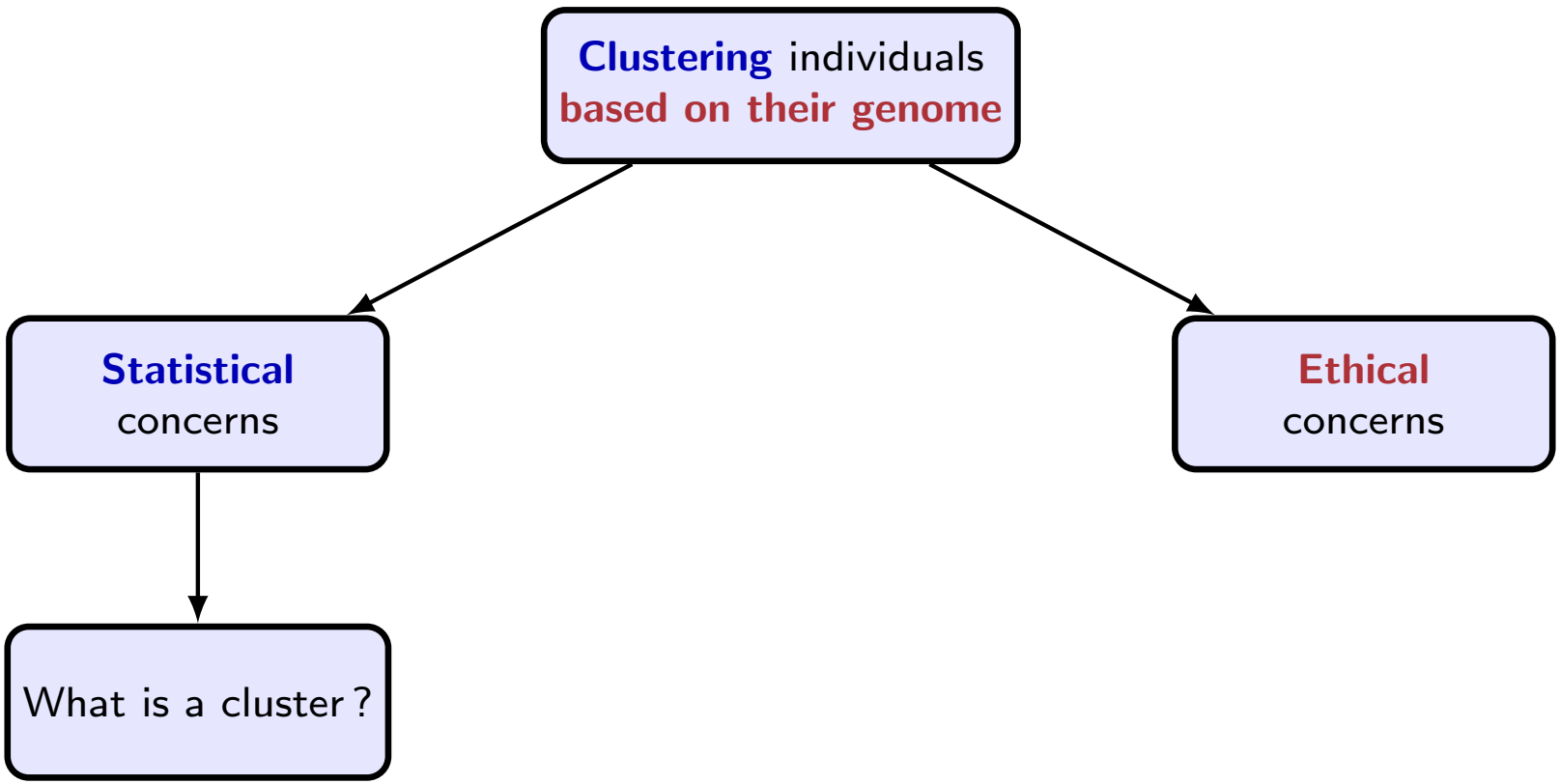
Clustering genetic data

Main challenges



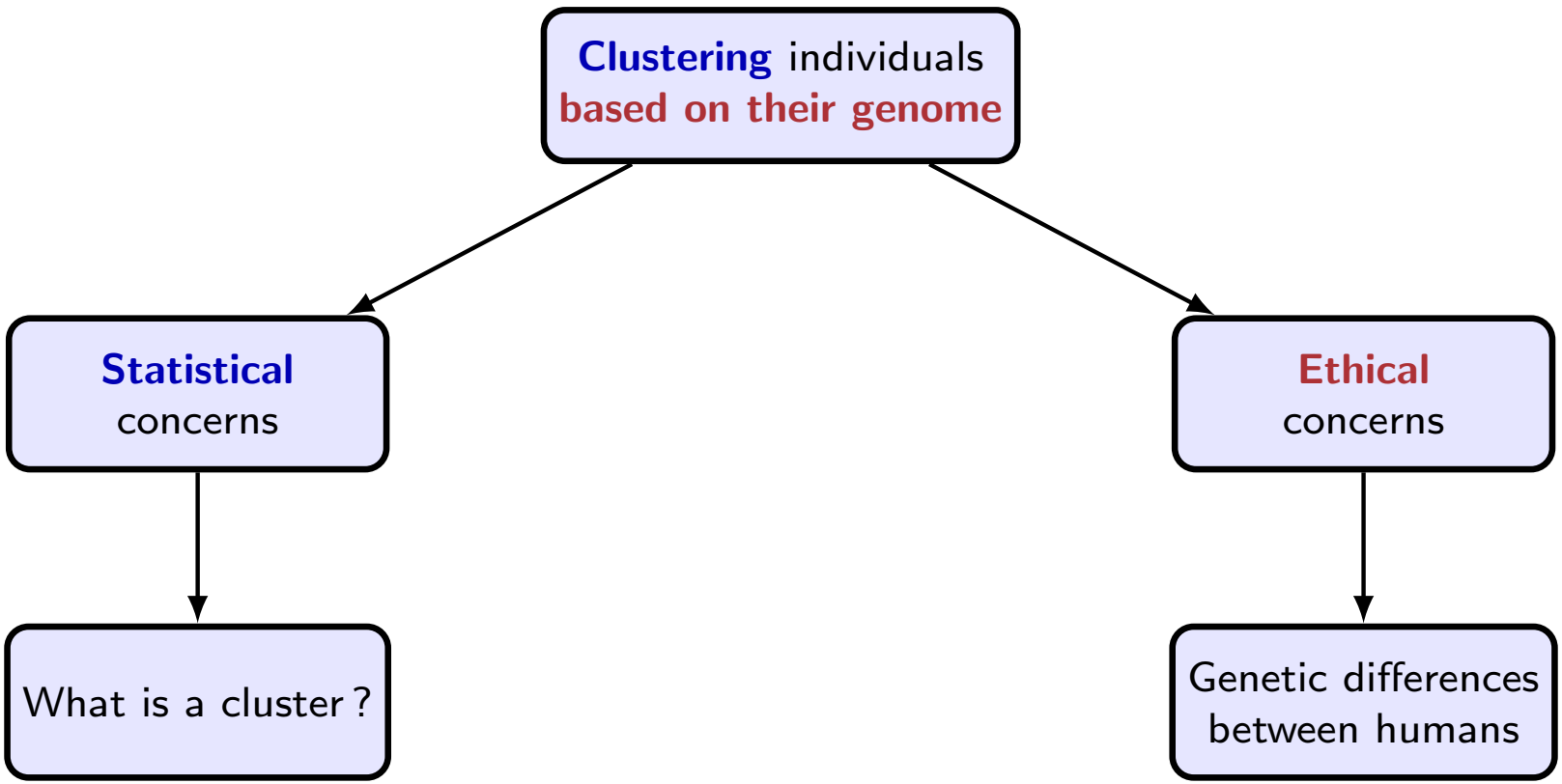
Clustering genetic data

Main challenges



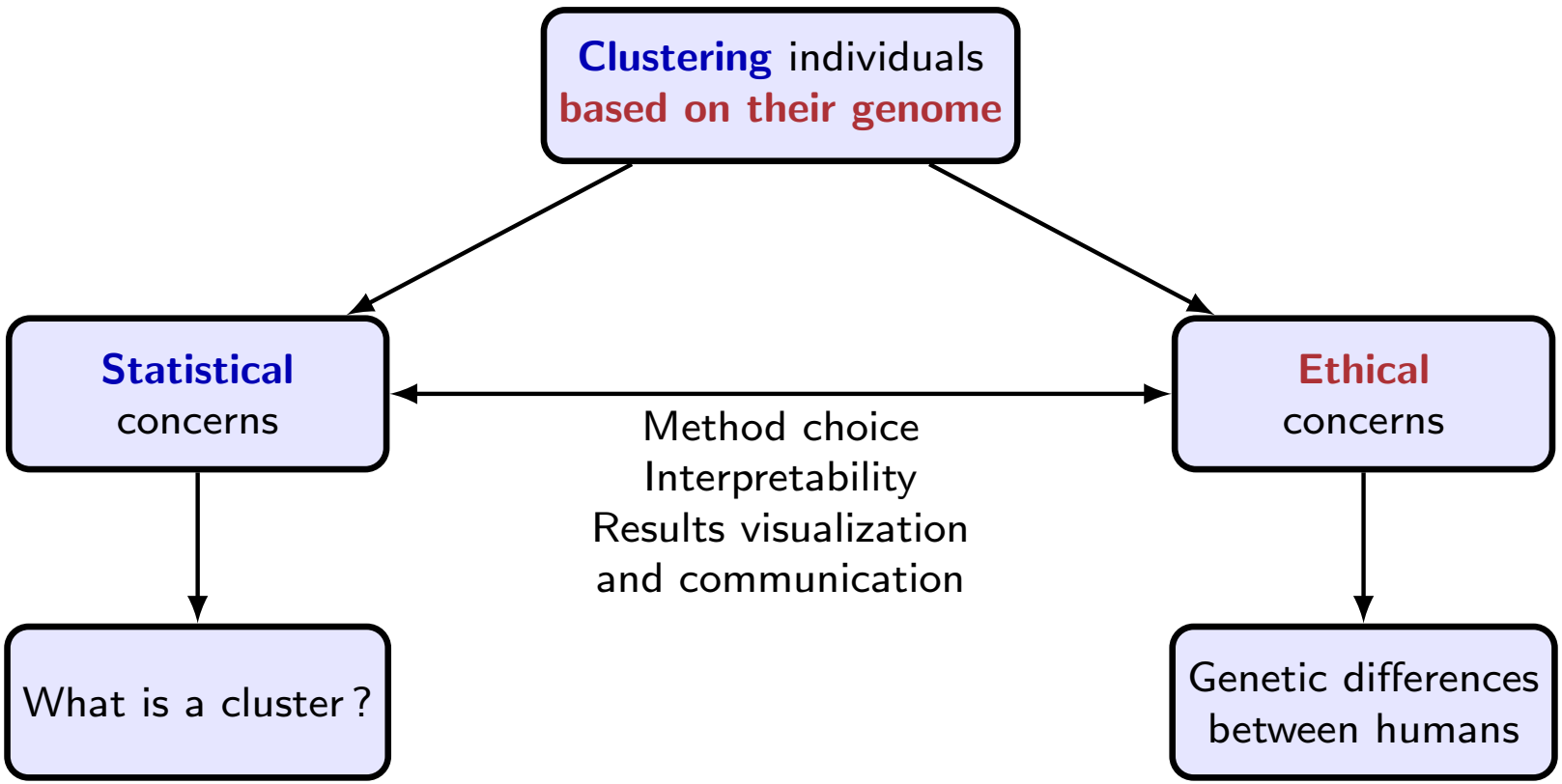
Clustering genetic data

Main challenges



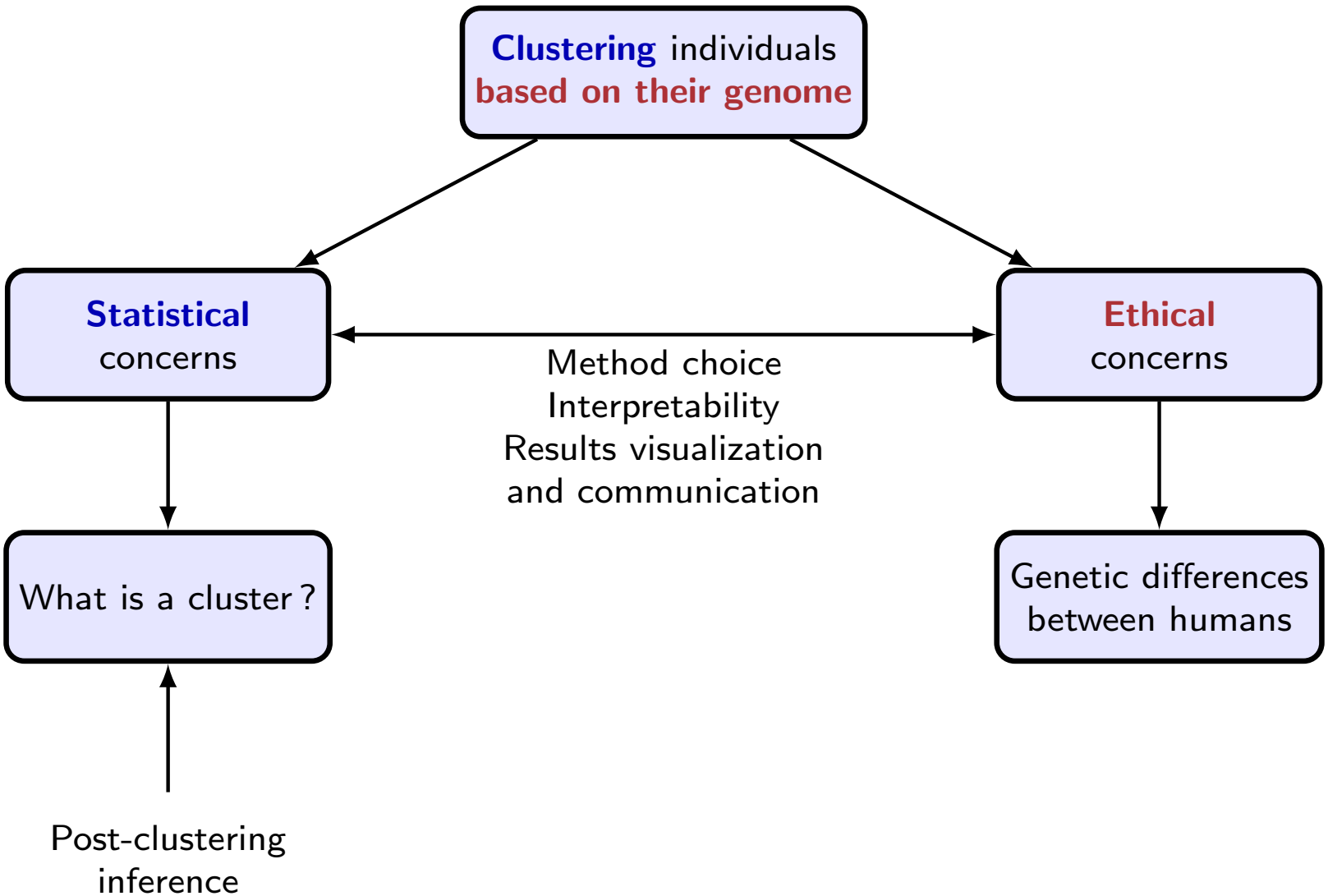
Clustering genetic data

Main challenges



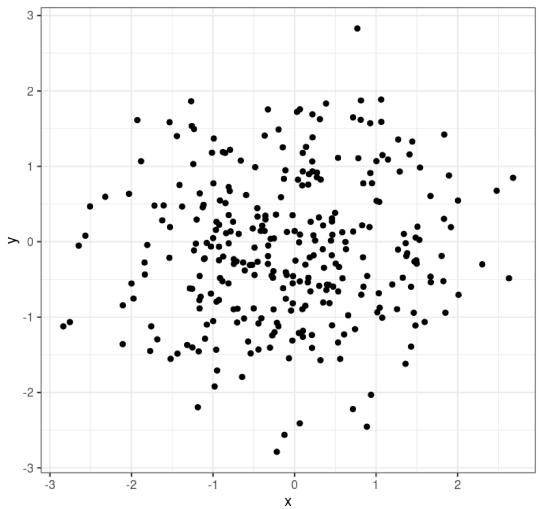
Clustering genetic data

Main challenges



Post-clustering inference

The problem of *double-dipping*

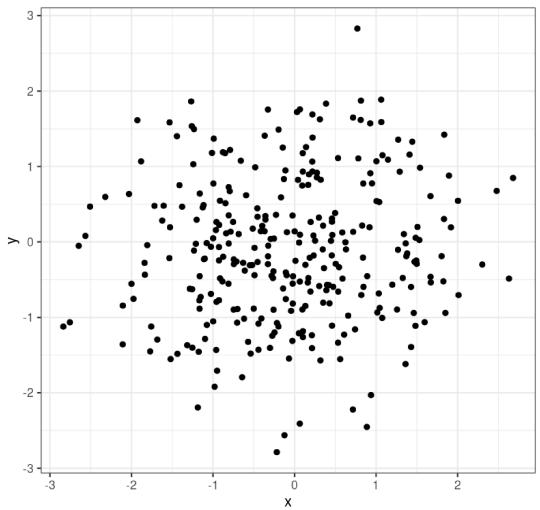


Data

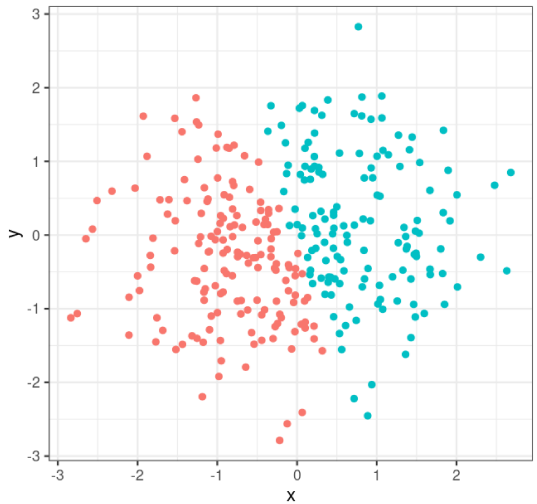
Inspired from Boris Hejblum's SMPGD 2026 slides.

Post-clustering inference

The problem of *double-dipping*



Data

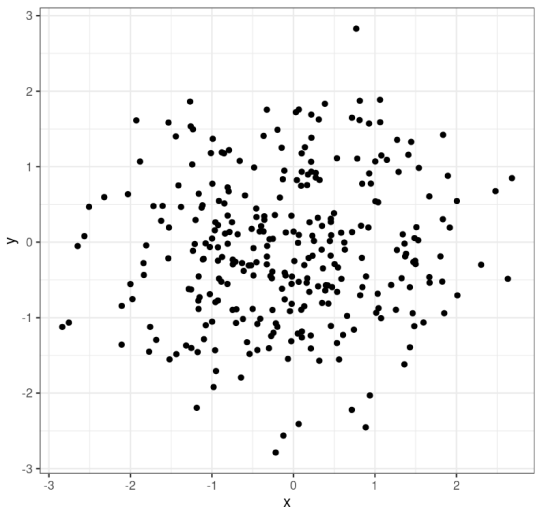


Clustering

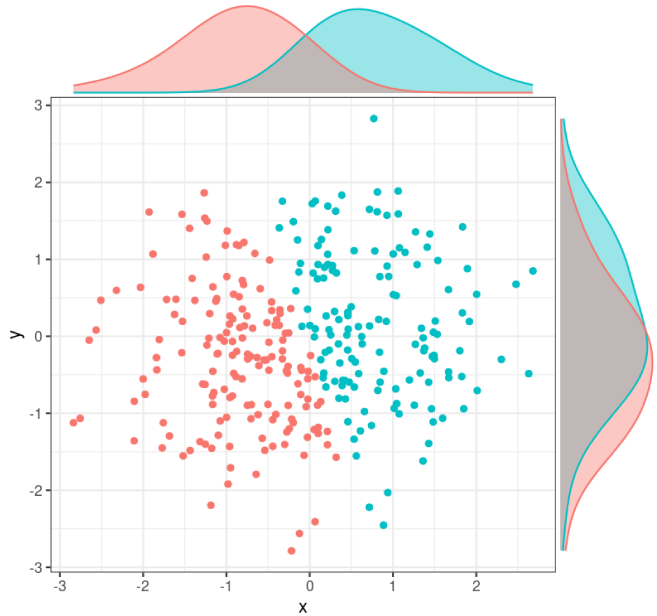
Inspired from Boris Hejblum's SMPGD 2026 slides.

Post-clustering inference

The problem of *double-dipping*



Data

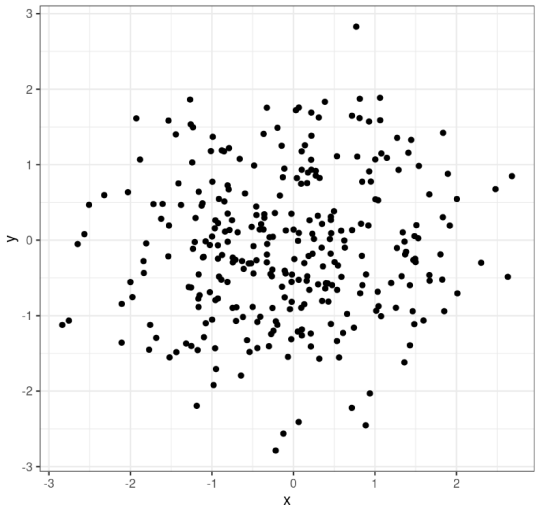


Cluster● = Cluster●?

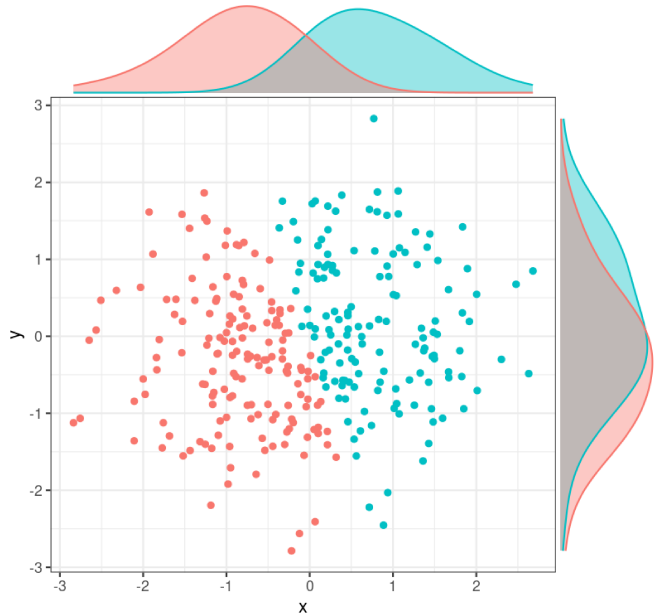
Inspired from Boris Hejblum's SMPGD 2026 slides.

Post-clustering inference

The problem of *double-dipping*



Data



Cluster● = Cluster●?

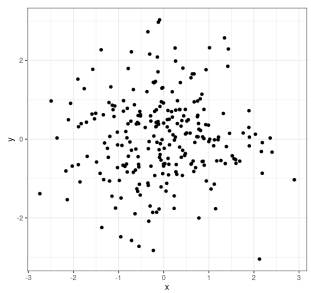
Double dipping

Same data used for both clustering and testing ⇒ Lost of type I error control.

Inspired from Boris Hejblum's SMPGD 2026 slides.

Post-clustering inference approaches

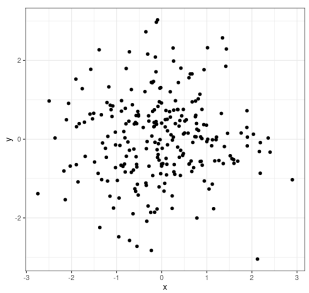
Sample splitting is not a solution



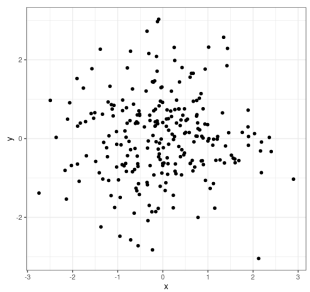
X

Inspired from Boris Hejblum's SMPGD 2026 slides.

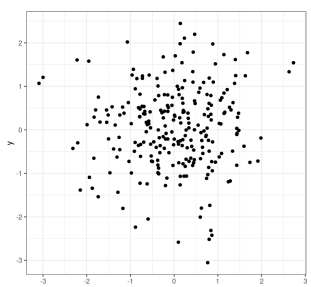
Sample splitting is not a solution



X



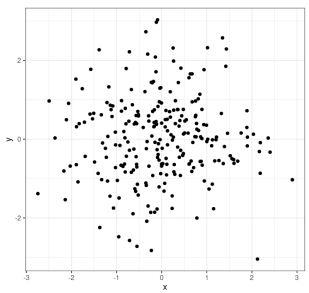
X_{train}



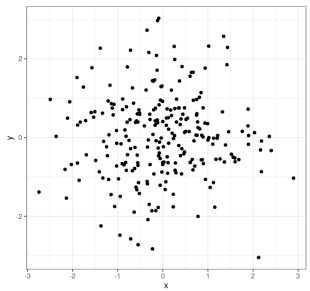
X_{test}

Inspired from Boris Hejblum's SMPGD 2026 slides.

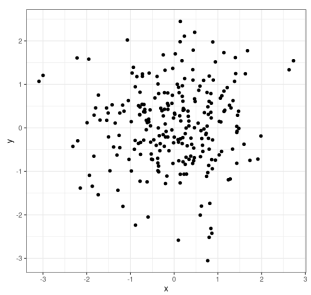
Sample splitting is not a solution



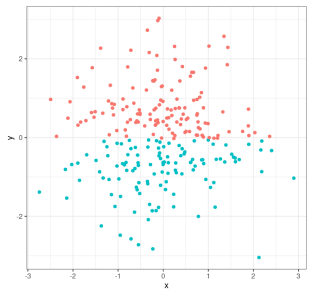
X



X_{train}



X_{test}

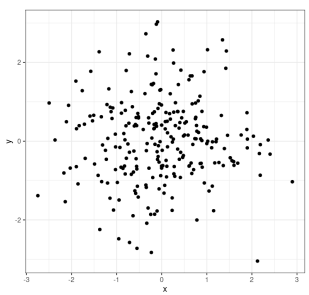


X_{train}

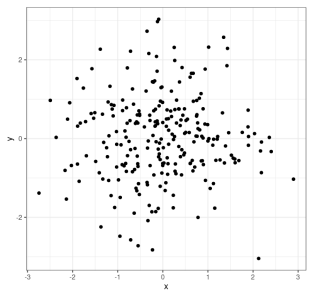
1. Learn a clustering on X_{train}

Inspired from Boris Hejblum's SMPGD 2026 slides.

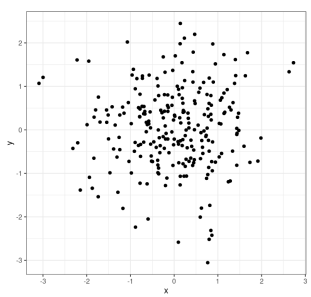
Sample splitting is not a solution



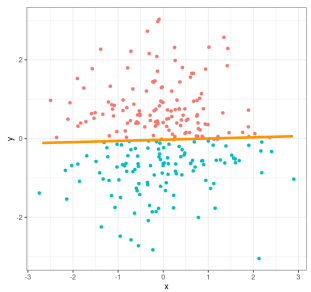
X



X_{train}



X_{test}

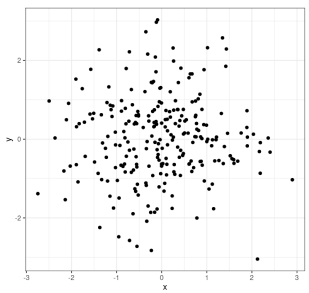


X_{train}

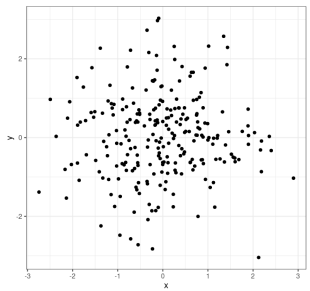
1. Learn a clustering on X_{train}
2. Train a classifier on X_{train}

Inspired from Boris Hejblum's SMPGD 2026 slides.

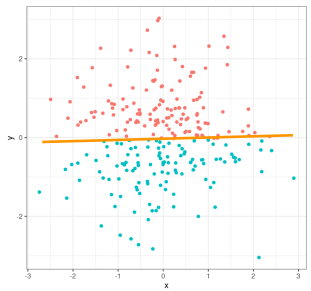
Sample splitting is not a solution



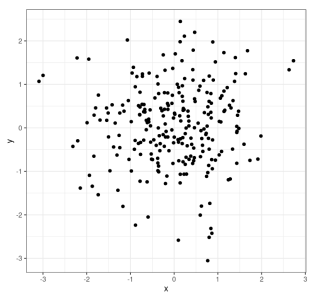
X



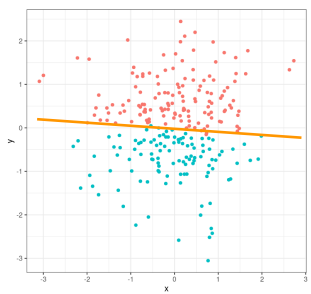
X_{train}



X_{train}



X_{test}

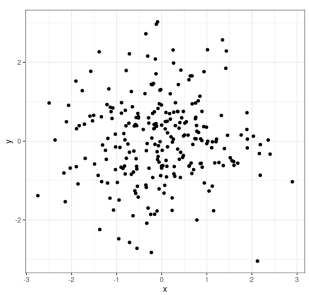


X_{test}

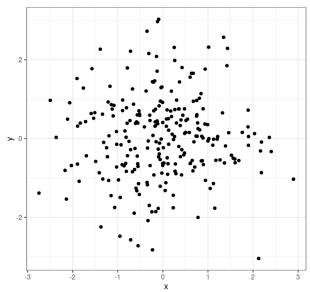
1. Learn a clustering on X_{train}
2. Train a classifier on X_{train}
3. Predict a clustering on X_{test}

Inspired from Boris Hejblum's SMPGD 2026 slides.

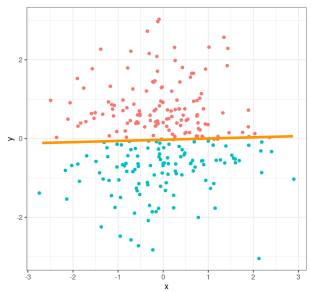
Sample splitting is not a solution



X

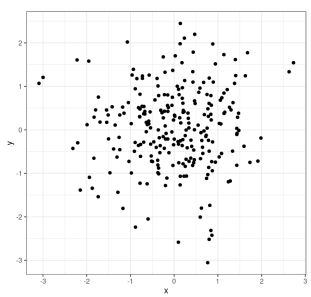


X_{train}

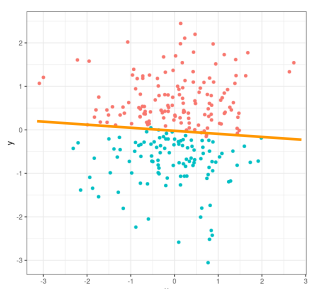


X_{train}

1. Learn a clustering on X_{train}
2. Train a classifier on X_{train}



X_{test}

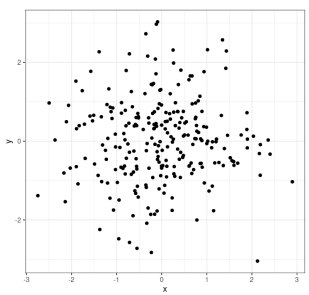


X_{test}

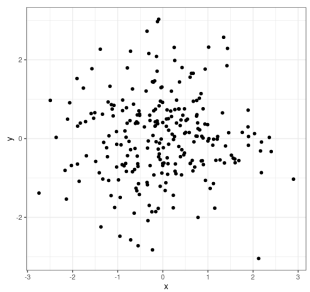
3. Predict a clustering on X_{test}
4. Test cluster differences on X_{test}

Inspired from Boris Hejblum's SMPGD 2026 slides.

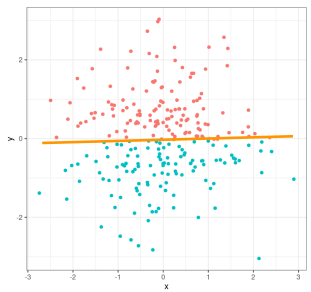
Sample splitting is not a solution



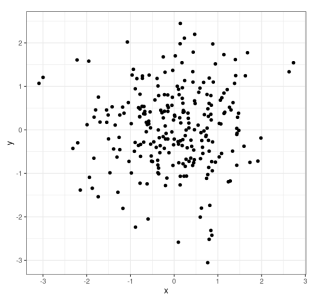
X



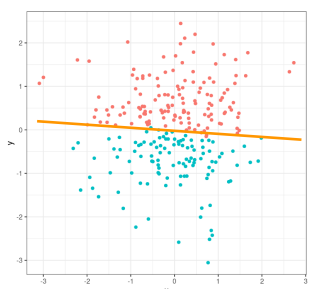
X_{train}



X_{train}



X_{test}



X_{test}

1. Learn a clustering on X_{train}
2. Train a classifier on X_{train}

3. Predict a clustering on X_{test}
4. Test cluster differences on X_{test}

3 and 4 both use the same data:

still double dipping!

Inspired from Boris Hejblum's SMPGD 2026 slides.

Information partitioning approaches

Data fission / data thinning

Information partitioning approaches

Data fission / data thinning

Idea

Decompose the dataset $\mathbf{X}$ into two independent datasets $\mathbf{X}^{(1)}$, $\mathbf{X}^{(2)}$ and :

- Use $\mathbf{X}^{(1)}$ for clustering,
- Use $\mathbf{X}^{(2)}$ for testing.

Inspired from Boris Hejblum's SMPGD 2026 slides.

Information partitioning approaches

Data fission / data thinning

Idea

Decompose the dataset $\mathbf{X}$ into two independent datasets $\mathbf{X}^{(1)}$, $\mathbf{X}^{(2)}$ and :

- Use $\mathbf{X}^{(1)}$ for clustering,
- Use $\mathbf{X}^{(2)}$ for testing.

Known decompositions for Gaussian data $X_1, \overset{\text{i.i.d.}}{\ldots}, X_n \sim \mathcal{N}(\mu, \mathbf{\Sigma})$

Data fission (Leiner *et al.*, 2023)

$$X_i^{(1)} = X_i + \tau W_i$$

$$X_i^{(2)} = X_i + W_i/\tau$$

with $W \sim \mathcal{N}(0, \mathbf{\Sigma})$ and $\tau > 0$.

Data thinning (Neufeld *et al.*, 2024)

$$X_i^{(1)} \mid X_i = x_i \sim \mathcal{N}(\tau x_i, \tau(1-\tau)\mathbf{\Sigma})$$

$$X_i^{(2)} = X_i - X_i^{(1)}$$

with $\tau \in (0, 1)$.

Inspired from Boris Hejblum's SMPGD 2026 slides.

Information partitioning approaches

Data fission / data thinning

Idea

Decompose the dataset $\mathbf{X}$ into two independent datasets $\mathbf{X}^{(1)}$, $\mathbf{X}^{(2)}$ and :

- Use $\mathbf{X}^{(1)}$ for clustering,
- Use $\mathbf{X}^{(2)}$ for testing.

Known decompositions for Gaussian data $X_1, \overset{\text{i.i.d.}}{\ldots}, X_n \sim \mathcal{N}(\mu, \mathbf{\Sigma})$

Data fission (Leiner *et al.*, 2023)

$$X_i^{(1)} = X_i + \tau W_i$$

$$X_i^{(2)} = X_i + W_i/\tau$$

with $W \sim \mathcal{N}(0, \mathbf{\Sigma})$ and $\tau > 0$.

Data thinning (Neufeld *et al.*, 2024)

$$X_i^{(1)} \mid X_i = x_i \sim \mathcal{N}(\tau x_i, \tau(1-\tau)\mathbf{\Sigma})$$

$$X_i^{(2)} = X_i - X_i^{(1)}$$

with $\tau \in (0, 1)$.

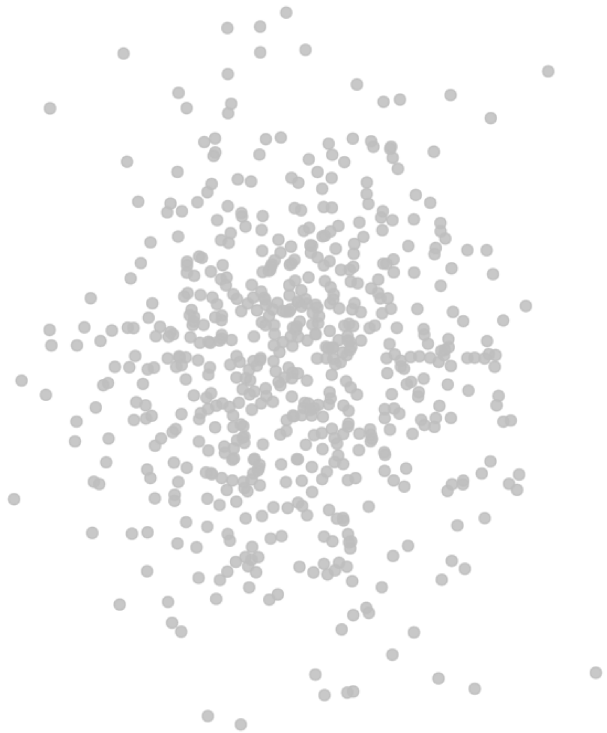
$$\mathbb{P}_{H_0}\left(\text{Reject } H_0 \text{ based on } \mathbf{X}^{(2)} \,\middle|\, \text{Select } H_0 \text{ based on } \mathbf{X}^{(1)}\right) = \mathbf{P}_{H_0}\left(\text{Reject } H_0 \text{ based on } \mathbf{X}^{(2)}\right)$$

$\rightarrow$ Thanks to independence.

Inspired from Boris Hejblum's SMPGD 2026 slides.

Information partitioning approaches

Data fission / data thinning



X

Inspired from Boris Hejblum's SMPGD 2026 slides.

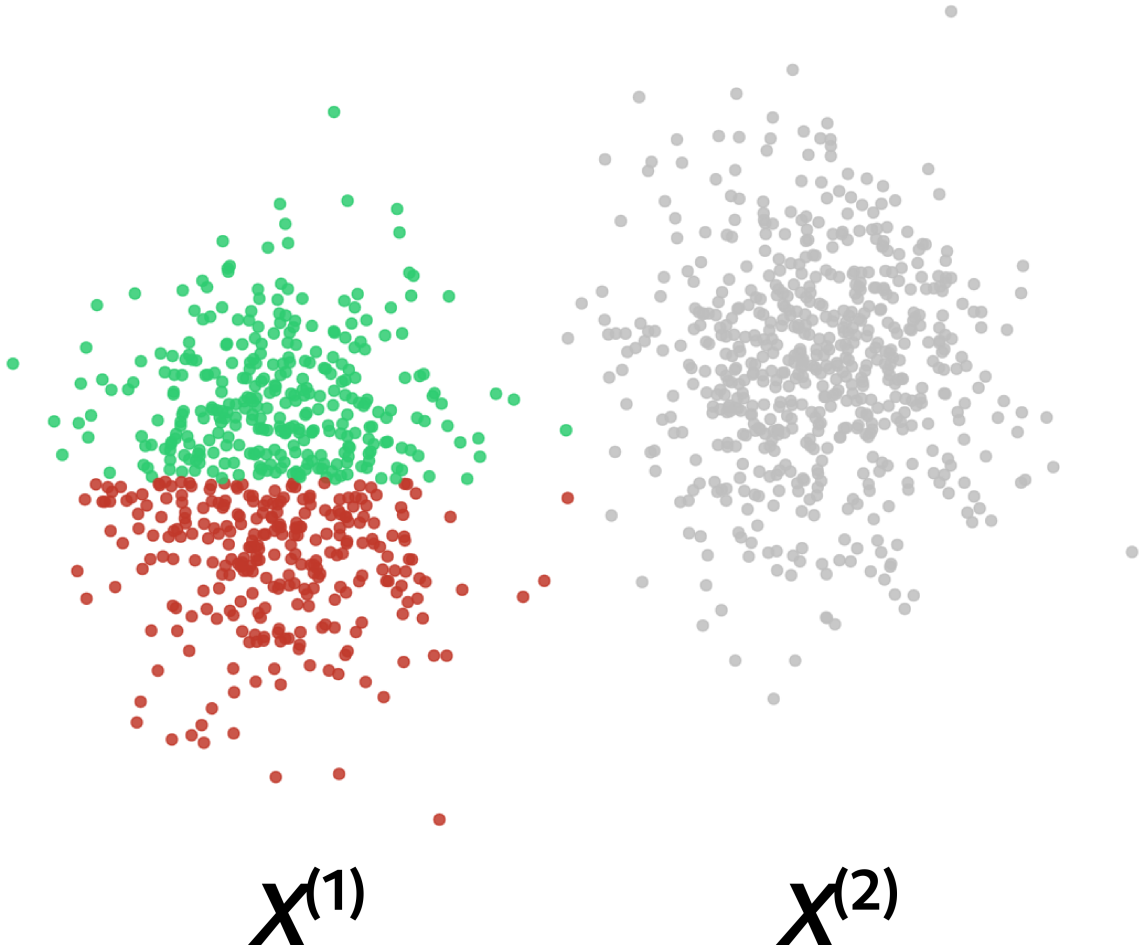
Information partitioning approaches

Data fission / data thinning

Inspired from Boris Hejblum's SMPGD 2026 slides.

Information partitioning approaches

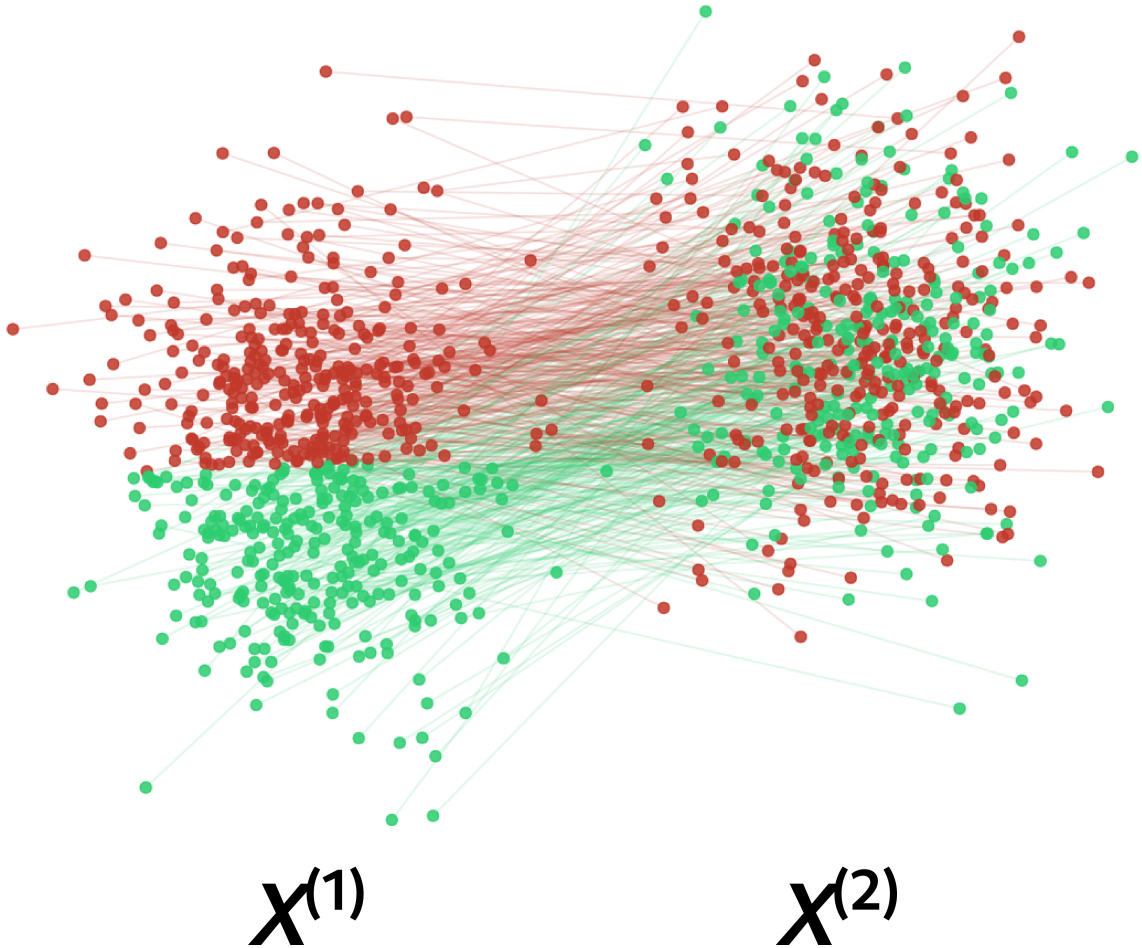
Data fission / data thinning



Inspired from Boris Hejblum's SMPGD 2026 slides.

Information partitioning approaches

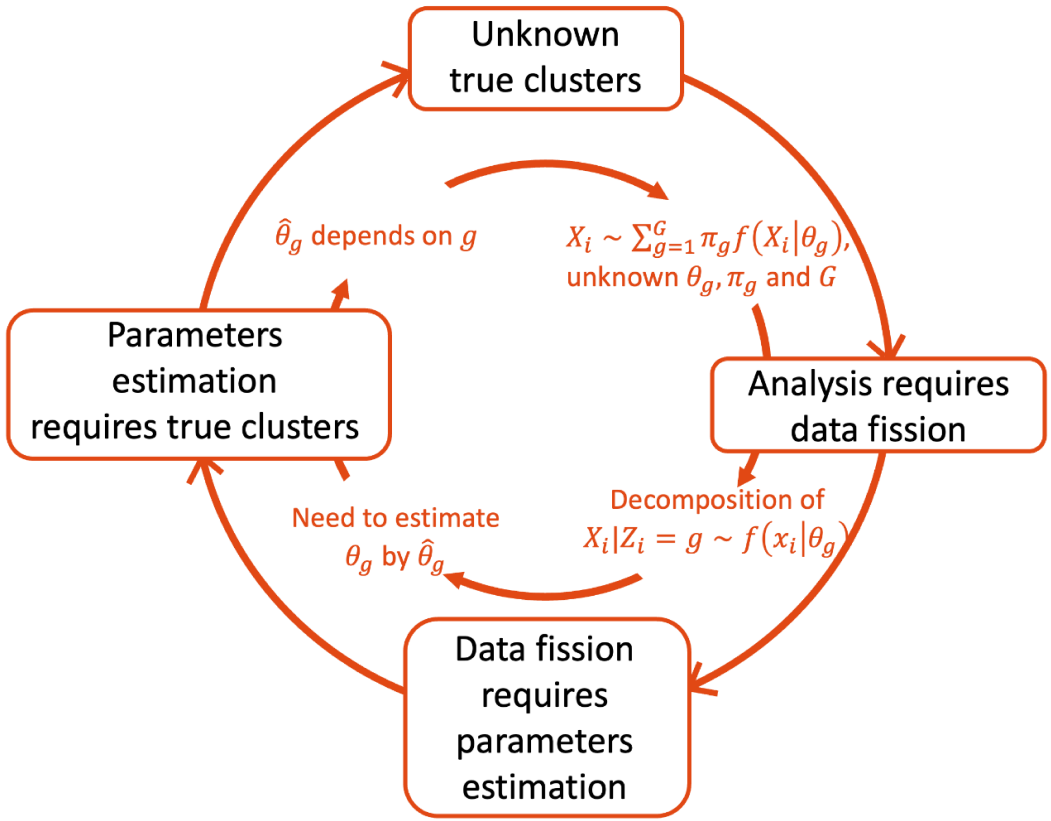
Data fission / data thinning



Inspired from Boris Hejblum's SMPGD 2026 slides.

Information partitioning approaches

Unsuitable for clustering



Extracted from Supplementary Figure 1 in Hivert *et al.* 2025, arXiv :2405.13591v3.

Conditional inference

Type I error control

$$\mathbb{P}_{H_0}\left(\text{Reject } H_0 \text{ based on } \mathbf{X}\right) \leq \alpha \quad \forall\, \alpha \in (0,1).$$

Conditional inference

Selective type I error control (Fithian *et al.* 2014)

$$\mathbb{P}_{H_0}\left(\text{Reject } H_0 \text{ based on } \mathbf{X} \,|\, H_0 \text{ selected using } \mathbf{X}\right) \leq \alpha \quad \forall\, \alpha \in (0,1).$$

Conditional inference

Selective type I error control (Fithian *et al.* 2014)

$$\mathbb{P}_{H_0}\left(\text{Reject } H_0 \text{ based on } \mathbf{X} \,|\, H_0 \text{ selected using } \mathbf{X}\right) \leq \alpha \quad \forall\, \alpha \in (0,1).$$

Selective type I error control for clustering

$$\mathbb{P}_{H_0}\left(\text{Reject } H_0 \text{ based on } \mathbf{X} \,|\, \mathcal{G}_1, \mathcal{G}_2 \in \mathcal{C}(\mathbf{X})\right) \leq \alpha \quad \forall\, \alpha \in (0,1),$$

$\rightarrow \mathcal{G}_1, \mathcal{G}_2$ are two non-overlapping groups of observations,

$\rightarrow \mathcal{C}$ is a clustering algorithm.

Conditional inference

Selective type I error control (Fithian *et al.* 2014)

$$\mathbb{P}_{H_0}\left(\text{Reject } H_0 \text{ based on } \mathbf{X} \,|\, H_0 \text{ selected using } \mathbf{X}\right) \leq \alpha \quad \forall\, \alpha \in (0,1).$$

Selective type I error control for clustering

$$\mathbb{P}_{H_0}\left(\text{Reject } H_0 \text{ based on } \mathbf{X} \,|\, \mathcal{G}_1, \mathcal{G}_2 \in \mathcal{C}(\mathbf{X})\right) \leq \alpha \quad \forall\, \alpha \in (0,1),$$

$\rightarrow \mathcal{G}_1, \mathcal{G}_2$ are two non-overlapping groups of observations,

$\rightarrow \mathcal{C}$ is a clustering algorithm.

To date, **conditional approaches** are the only methods that have been demonstrated to be suitable for post-clustering inference.

State-of-the-art and goal

State-of-the-art and goal

(Main) state-of-the-art and its model assumptions

- Independent observations $X_i \sim \mathcal{N}_p(\mu_i, \sigma^2 \mathbb{I}_p)$ + estimation of σ.

State-of-the-art and goal

(Main) state-of-the-art and its model assumptions

- Independent observations $X_i \sim \mathcal{N}_p(\mu_i, \sigma^2 \mathbb{I}_p)$ + estimation of σ.

  $\rightarrow$ Gao *et al.* (2022), Yun and Barber (2023), Chen and Witten (2023).

State-of-the-art and goal

(Main) state-of-the-art and its model assumptions

- Independent observations $X_i \sim \mathcal{N}_p(\mu_i, \sigma^2 \mathbb{I}_p)$ + estimation of σ.

  $\rightarrow$ Gao *et al.* (2022), Yun and Barber (2023), Chen and Witten (2023).

- General matrix normal model $\mathbf{X} \sim \mathcal{MN}_{n \times p}(\boldsymbol{\mu}, \mathbf{U}, \boldsymbol{\Sigma})$ + estimation of $\boldsymbol{\Sigma}$.

  $\rightarrow$ González-Delgado *et al.* (2025).

State-of-the-art and goal

(Main) state-of-the-art and its model assumptions

- Independent observations $X_i \sim \mathcal{N}_p(\mu_i, \sigma^2 \mathbb{I}_p)$ + estimation of σ.

  $\rightarrow$ Gao *et al.* (2022), Yun and Barber (2023), Chen and Witten (2023).

- General matrix normal model $\mathbf{X} \sim \mathcal{MN}_{n\times p}(\boldsymbol{\mu}, \mathbf{U}, \boldsymbol{\Sigma})$ + estimation of $\boldsymbol{\Sigma}$.

  $\rightarrow$ González-Delgado *et al.* (2025).

  For $\mathbf{U} \in \mathbb{R}^{n\times n}$ and $\boldsymbol{\Sigma} \in \mathbb{R}^{p\times p}$ positive definite, that means :

  - $\text{vec}(\mathbf{X}) \sim \mathcal{N}_{n\times p}\left(\text{vec}(\boldsymbol{\mu}), \boldsymbol{\Sigma} \otimes \mathbf{U}\right)$,
  - $X_i \sim \mathcal{N}_p(\mu_i, U_{ii}\boldsymbol{\Sigma})$ ($\boldsymbol{\Sigma} \leftrightarrow$ dependence between features),
  - $X^j \sim \mathcal{N}_n(\mu^j, \Sigma_{jj}\mathbf{U})$ ($\mathbf{U} \leftrightarrow$ dependence between observations).

State-of-the-art and goal

(Main) state-of-the-art and its model assumptions

- Independent observations $X_i \sim \mathcal{N}_p(\mu_i, \sigma^2 \mathbb{I}_p)$ + estimation of σ.

  $\rightarrow$ Gao *et al.* (2022), Yun and Barber (2023), Chen and Witten (2023).

- General matrix normal model $\mathbf{X} \sim \mathcal{MN}_{n \times p}(\boldsymbol{\mu}, \mathbf{U}, \boldsymbol{\Sigma})$ + estimation of $\boldsymbol{\Sigma}$.

  $\rightarrow$ González-Delgado *et al.* (2025).

Simplistic model

State-of-the-art and goal

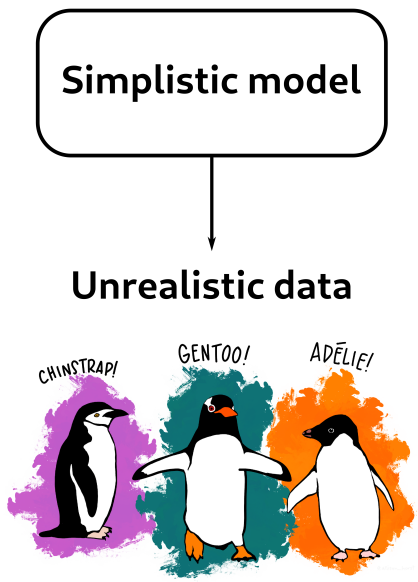
(Main) state-of-the-art and its model assumptions

- Independent observations $X_i \sim \mathcal{N}_p(\mu_i, \sigma^2 \mathbb{I}_p)$ + estimation of σ.

  $\rightarrow$ Gao *et al.* (2022), Yun and Barber (2023), Chen and Witten (2023).

- General matrix normal model $\mathbf{X} \sim \mathcal{MN}_{n \times p}(\boldsymbol{\mu}, \mathbf{U}, \boldsymbol{\Sigma})$ + estimation of $\boldsymbol{\Sigma}$.

  $\rightarrow$ González-Delgado *et al.* (2025).



State-of-the-art and goal

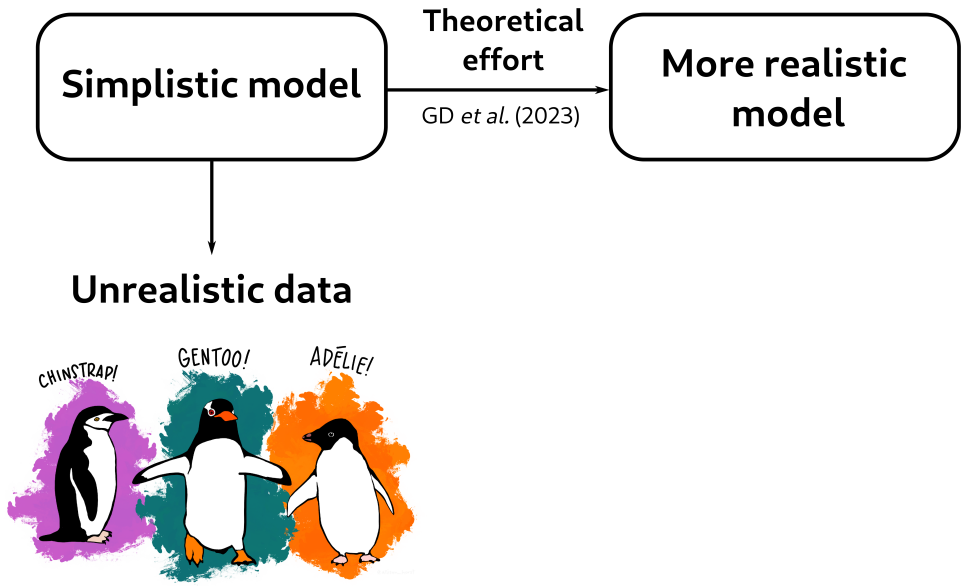
(Main) state-of-the-art and its model assumptions

- Independent observations $X_i \sim \mathcal{N}_p(\mu_i, \sigma^2 \mathbb{I}_p)$ + estimation of σ.

  $\rightarrow$ Gao *et al.* (2022), Yun and Barber (2023), Chen and Witten (2023).

- General matrix normal model $\mathbf{X} \sim \mathcal{MN}_{n \times p}(\boldsymbol{\mu}, \mathbf{U}, \boldsymbol{\Sigma})$ + estimation of $\boldsymbol{\Sigma}$.

  $\rightarrow$ González-Delgado *et al.* (2025).



State-of-the-art and goal

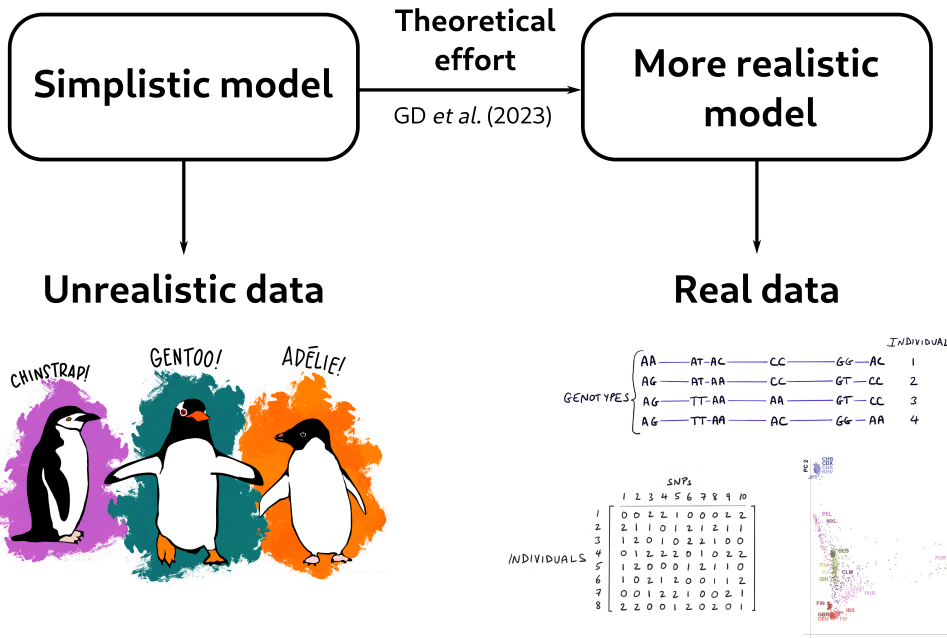
(Main) state-of-the-art and its model assumptions

- Independent observations $X_i \sim \mathcal{N}_p(\mu_i, \sigma^2 \mathbb{I}_p)$ + estimation of σ.

  $\rightarrow$ Gao *et al.* (2022), Yun and Barber (2023), Chen and Witten (2023).

- General matrix normal model $\mathbf{X} \sim \mathcal{MN}_{n \times p}(\boldsymbol{\mu}, \mathbf{U}, \boldsymbol{\Sigma})$ + estimation of $\boldsymbol{\Sigma}$.

  $\rightarrow$ González-Delgado *et al.* (2025).



State-of-the-art and goal

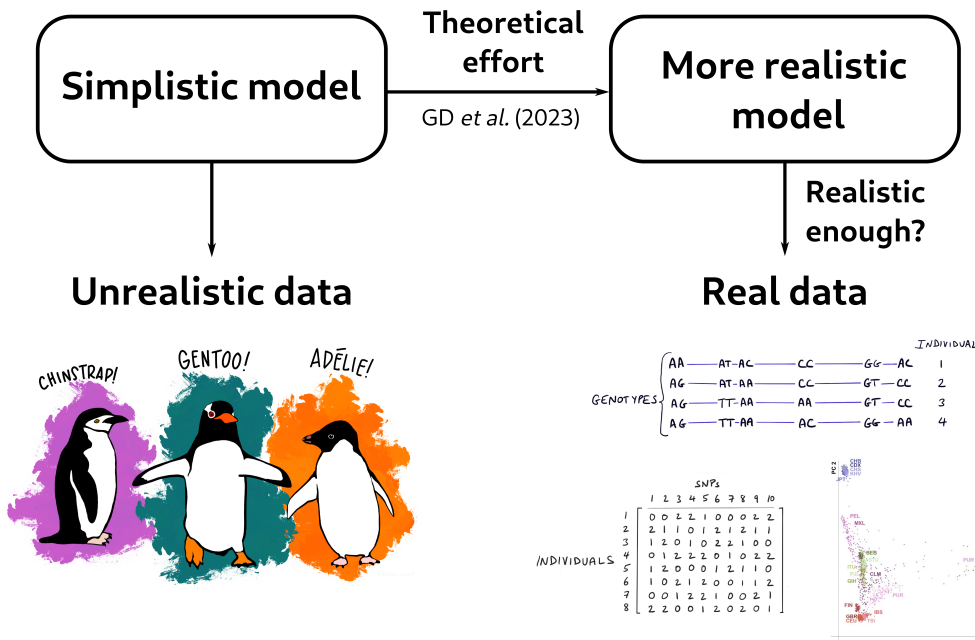
(Main) state-of-the-art and its model assumptions

- Independent observations $X_i \sim \mathcal{N}_p(\mu_i, \sigma^2 \mathbb{I}_p)$ + estimation of σ.

  $\rightarrow$ Gao *et al.* (2022), Yun and Barber (2023), Chen and Witten (2023).

- General matrix normal model $\mathbf{X} \sim \mathcal{MN}_{n\times p}(\boldsymbol{\mu}, \mathbf{U}, \boldsymbol{\Sigma})$ + estimation of $\boldsymbol{\Sigma}$.

  $\rightarrow$ González-Delgado *et al.* (2025).



Conditional approaches on real data

Framework setting

Framework setting

- Let $C(\cdot)$ be a clustering algorithm, $\mathbf{X}$ a $n \times p$ random matrix with $\mathbb{E}(\mathbf{X}) = \boldsymbol{\mu}$.

Framework setting

- Let $C(\cdot)$ be a clustering algorithm, $\mathbf{X}$ a $n \times p$ random matrix with $\mathbb{E}(\mathbf{X}) = \boldsymbol{\mu}$.
- Let X_i (resp. μ_i) denote the i-th row of $\mathbf{X}$ (resp. $\boldsymbol{\mu}$) for $i \in [n] = \{1, \ldots, n\}$.

Framework setting

- Let $C(\cdot)$ be a clustering algorithm, $\mathbf{X}$ a $n \times p$ random matrix with $\mathbb{E}(\mathbf{X}) = \boldsymbol{\mu}$.
- Let X_i (resp. μ_i) denote the i-th row of $\mathbf{X}$ (resp. $\boldsymbol{\mu}$) for $i \in [n] = \{1, \ldots, n\}$.
- For any $\mathcal{G} \subset \{1, \ldots, n\}$, let $\bar{X}_{\mathcal{G}} = \frac{1}{|\mathcal{G}|} \sum_{i \in \mathcal{G}} X_i$ and $\bar{\mu}_{\mathcal{G}} = \frac{1}{|\mathcal{G}|} \sum_{i \in \mathcal{G}} \mu_i$.

Framework setting

- Let $C(\cdot)$ be a clustering algorithm, $\mathbf{X}$ a $n \times p$ random matrix with $\mathbb{E}(\mathbf{X}) = \boldsymbol{\mu}$.
- Let X_i (resp. μ_i) denote the i-th row of $\mathbf{X}$ (resp. $\boldsymbol{\mu}$) for $i \in [n] = \{1, \ldots, n\}$.
- For any $\mathcal{G} \subset \{1, \ldots, n\}$, let $\bar{X}_{\mathcal{G}} = \frac{1}{|\mathcal{G}|} \sum_{i \in \mathcal{G}} X_i$ and $\bar{\mu}_{\mathcal{G}} = \frac{1}{|\mathcal{G}|} \sum_{i \in \mathcal{G}} \mu_i$.
- Let $\mathcal{G}_1, \mathcal{G}_2 \subset \{1, \ldots, n\}$ be two non-overlapping groups of observations.

Framework setting

- Let $C(\cdot)$ be a clustering algorithm, $\mathbf{X}$ a $n \times p$ random matrix with $\mathbb{E}(\mathbf{X}) = \boldsymbol{\mu}$.
- Let X_i (resp. μ_i) denote the i-th row of $\mathbf{X}$ (resp. $\boldsymbol{\mu}$) for $i \in [n] = \{1, \ldots, n\}$.
- For any $\mathcal{G} \subset \{1, \ldots, n\}$, let $\bar{X}_{\mathcal{G}} = \frac{1}{|\mathcal{G}|} \sum_{i \in \mathcal{G}} X_i$ and $\bar{\mu}_{\mathcal{G}} = \frac{1}{|\mathcal{G}|} \sum_{i \in \mathcal{G}} \mu_i$.
- Let $\mathcal{G}_1, \mathcal{G}_2 \subset \{1, \ldots, n\}$ be two non-overlapping groups of observations. Considering the column vector $\nu_{\mathcal{G}_1, \mathcal{G}_2} = \nu$ having as components

$$\nu_i = \mathbb{1}\{i \in \mathcal{G}_1\}/|\mathcal{G}_1| - \mathbb{1}\{i \in \mathcal{G}_2\}/|\mathcal{G}_2|,$$

for $i \in [n]$,

Framework setting

- Let $C(\cdot)$ be a clustering algorithm, $\mathbf{X}$ a $n \times p$ random matrix with $\mathbb{E}(\mathbf{X}) = \boldsymbol{\mu}$.
- Let X_i (resp. μ_i) denote the i-th row of $\mathbf{X}$ (resp. $\boldsymbol{\mu}$) for $i \in [n] = \{1, \dots, n\}$.
- For any $\mathcal{G} \subset \{1, \dots, n\}$, let $\bar{X}_{\mathcal{G}} = \frac{1}{|\mathcal{G}|} \sum_{i \in \mathcal{G}} X_i$ and $\bar{\mu}_{\mathcal{G}} = \frac{1}{|\mathcal{G}|} \sum_{i \in \mathcal{G}} \mu_i$.
- Let $\mathcal{G}_1, \mathcal{G}_2 \subset \{1, \dots, n\}$ be two non-overlapping groups of observations. Considering the column vector $\nu_{\mathcal{G}_1, \mathcal{G}_2} = \nu$ having as components

$$\nu_i = \mathbb{1}\{i \in \mathcal{G}_1\}/|\mathcal{G}_1| - \mathbb{1}\{i \in \mathcal{G}_2\}/|\mathcal{G}_2|,$$

for $i \in [n]$, we can write the difference between the (empirical) group means as

$$\bar{\mu}_{\mathcal{G}_1} - \bar{\mu}_{\mathcal{G}_2} = \boldsymbol{\mu}^T \nu \quad \text{and} \quad \bar{X}_{\mathcal{G}_1} - \bar{X}_{\mathcal{G}_2} = \mathbf{X}^T \nu.$$

Framework setting

- Let $C(\cdot)$ be a clustering algorithm, $\mathbf{X}$ a $n \times p$ random matrix with $\mathbb{E}(\mathbf{X}) = \boldsymbol{\mu}$.
- Let X_i (resp. μ_i) denote the i-th row of $\mathbf{X}$ (resp. $\boldsymbol{\mu}$) for $i \in [n] = \{1, \dots, n\}$.
- For any $\mathcal{G} \subset \{1, \dots, n\}$, let $\bar{X}_{\mathcal{G}} = \frac{1}{|\mathcal{G}|} \sum_{i \in \mathcal{G}} X_i$ and $\bar{\mu}_{\mathcal{G}} = \frac{1}{|\mathcal{G}|} \sum_{i \in \mathcal{G}} \mu_i$.
- Let $\mathcal{G}_1, \mathcal{G}_2 \subset \{1, \dots, n\}$ be two non-overlapping groups of observations. Considering the column vector $\nu_{\mathcal{G}_1, \mathcal{G}_2} = \nu$ having as components

$$\nu_i = \mathbb{1}\{i \in \mathcal{G}_1\}/|\mathcal{G}_1| - \mathbb{1}\{i \in \mathcal{G}_2\}/|\mathcal{G}_2|,$$

for $i \in [n]$, we can write the difference between the (empirical) group means as

$$\bar{\mu}_{\mathcal{G}_1} - \bar{\mu}_{\mathcal{G}_2} = \boldsymbol{\mu}^T \nu \quad \text{and} \quad \bar{X}_{\mathcal{G}_1} - \bar{X}_{\mathcal{G}_2} = \mathbf{X}^T \nu.$$

We are interested in the following null hypothesis :

$$H_0^{\{\mathcal{G}_1, \mathcal{G}_2\}} : \boldsymbol{\mu}^T \nu = 0. \qquad \text{(H0)}$$

Standard strategy

1. Choose a model

$$\mathbf{X} \sim \mathcal{MN}_{n\times p}(\boldsymbol{\mu}, \mathbf{U}, \boldsymbol{\Sigma}),$$

where $\boldsymbol{\Sigma}$ can be (over-)estimated and $\mathbf{U}$ is known.

González-Delgado *et al.*, 2025.

Standard strategy

1. Choose a model

$$\mathbf{X} \sim \mathcal{MN}_{n\times p}(\boldsymbol{\mu}, \mathbf{U}, \boldsymbol{\Sigma}),$$

where $\boldsymbol{\Sigma}$ can be (over-)estimated and $\mathbf{U}$ is known.

$\rightarrow$ Under (H0), $\mathbf{X}^T\nu \sim \mathcal{N}_p(\mathbf{0}_p, \mathbf{V})$, with $\mathbf{V} = \nu^T\mathbf{U}\nu\boldsymbol{\Sigma}$.

González-Delgado *et al.*, 2025.

Standard strategy

1. Choose a model

$$\mathbf{X} \sim \mathcal{MN}_{n\times p}(\boldsymbol{\mu}, \mathbf{U}, \boldsymbol{\Sigma}),$$

where $\boldsymbol{\Sigma}$ can be (over-)estimated and $\mathbf{U}$ is known.

$\rightarrow$ Under (H0), $\mathbf{X}^T\nu \sim \mathcal{N}_p(\mathbf{0}_p, \mathbf{V})$, with $\mathbf{V} = \nu^T\mathbf{U}\nu\boldsymbol{\Sigma}$.

$\rightarrow$ Therefore, $\|\mathbf{X}^T\nu\|^2_{\mathbf{V}} = (\mathbf{X}^T\nu)^T\mathbf{V}^{-1}(\mathbf{X}^T\nu) \sim \chi^2_p$ under the null.

González-Delgado *et al.*, 2025.

Standard strategy

1. Choose a model

$$\mathbf{X} \sim \mathcal{MN}_{n\times p}(\boldsymbol{\mu}, \mathbf{U}, \boldsymbol{\Sigma}),$$

where $\boldsymbol{\Sigma}$ can be (over-)estimated and $\mathbf{U}$ is known.

$\rightarrow$ Under (H0), $\mathbf{X}^T\nu \sim \mathcal{N}_p(\mathbf{0}_p, \mathbf{V})$, with $\mathbf{V} = \nu^T\mathbf{U}\nu\boldsymbol{\Sigma}$.

$\rightarrow$ Therefore, $\|\mathbf{X}^T\nu\|^2_{\mathbf{V}} = (\mathbf{X}^T\nu)^T\mathbf{V}^{-1}(\mathbf{X}^T\nu) \sim \chi^2_p$ under the null.

2. Consider a p-value of the form

$$p(\mathbf{x}; \{\mathcal{G}_1, \mathcal{G}_2\}) = \mathbb{P}_{H_0^{\{\mathcal{G}_1,\mathcal{G}_2\}}}\left(\|\mathbf{X}^T\nu\|_{\mathbf{V}} \geq \|\mathbf{x}^T\nu\|_{\mathbf{V}} \;\middle|\; \mathcal{G}_1, \mathcal{G}_2 \in \mathcal{C}(\mathbf{X}),\right.$$

González-Delgado *et al.*, 2025.

Standard strategy

1. Choose a model

$$\mathbf{X} \sim \mathcal{MN}_{n\times p}(\boldsymbol{\mu}, \mathbf{U}, \boldsymbol{\Sigma}),$$

where $\boldsymbol{\Sigma}$ can be (over-)estimated and $\mathbf{U}$ is known.

$\rightarrow$ Under (H0), $\mathbf{X}^T\nu \sim \mathcal{N}_p(\mathbf{0}_p, \mathbf{V})$, with $\mathbf{V} = \nu^T\mathbf{U}\nu\boldsymbol{\Sigma}$.

$\rightarrow$ Therefore, $\|\mathbf{X}^T\nu\|^2_{\mathbf{V}} = (\mathbf{X}^T\nu)^T\mathbf{V}^{-1}(\mathbf{X}^T\nu) \sim \chi^2_p$ under the null.

2. Consider a p-value of the form

$$p(\mathbf{x}; \{\mathcal{G}_1, \mathcal{G}_2\}) = \mathbb{P}_{H_0^{\{\mathcal{G}_1,\mathcal{G}_2\}}}\Big(\|\mathbf{X}^T\nu\|_{\mathbf{V}} \geq \|\mathbf{x}^T\nu\|_{\mathbf{V}} \;\Big|\; \mathcal{G}_1, \mathcal{G}_2 \in \mathcal{C}(\mathbf{X}), \boldsymbol{\pi}_\nu^\perp\mathbf{X} = \boldsymbol{\pi}_\nu^\perp\mathbf{x}\,,\, \mathrm{dir}_{\mathbf{V}}(\mathbf{X}^T\nu) = \mathrm{dir}_{\mathbf{V}}(\mathbf{x}^T\nu)\Big),$$

where $\boldsymbol{\pi}_\nu^\perp = \mathbb{I}_n - \frac{\nu\nu^T}{\|\nu\|_2^2}$ and $\mathrm{dir}_{\mathbf{V}}(u) = u/\|u\|_{\mathbf{V}}$ for all $u \in \mathbb{R}^p$.

González-Delgado *et al.*, 2025.

Standard strategy

1. Choose a model

$$\mathbf{X} \sim \mathcal{MN}_{n\times p}(\boldsymbol{\mu}, \mathbf{U}, \boldsymbol{\Sigma}),$$

where $\boldsymbol{\Sigma}$ can be (over-)estimated and $\mathbf{U}$ is known.

$\rightarrow$ Under (H0), $\mathbf{X}^T\nu \sim \mathcal{N}_p(\mathbf{0}_p, \mathbf{V})$, with $\mathbf{V} = \nu^T\mathbf{U}\nu\boldsymbol{\Sigma}$.

$\rightarrow$ Therefore, $\|\mathbf{X}^T\nu\|^2_{\mathbf{V}} = (\mathbf{X}^T\nu)^T\mathbf{V}^{-1}(\mathbf{X}^T\nu) \sim \chi^2_p$ under the null.

2. Consider a p-value of the form

$$p(\mathbf{x}; \{\mathcal{G}_1, \mathcal{G}_2\}) = \mathbb{P}_{H_0^{\{\mathcal{G}_1,\mathcal{G}_2\}}}\Big(\|\mathbf{X}^T\nu\|_{\mathbf{V}} \geq \|\mathbf{x}^T\nu\|_{\mathbf{V}} \;\Big|\; \mathcal{G}_1, \mathcal{G}_2 \in \mathcal{C}(\mathbf{X}), \boldsymbol{\pi}^{\perp}_{\nu}\mathbf{X} = \boldsymbol{\pi}^{\perp}_{\nu}\mathbf{x}\,,\, \mathrm{dir}_{\mathbf{V}}(\mathbf{X}^T\nu) = \mathrm{dir}_{\mathbf{V}}(\mathbf{x}^T\nu)\Big),$$

where $\boldsymbol{\pi}^{\perp}_{\nu} = \mathbb{I}_n - \frac{\nu\nu^T}{\|\nu\|_2^2}$ and $\mathrm{dir}_{\mathbf{V}}(u) = u/\|u\|_{\mathbf{V}}$ for all $u \in \mathbb{R}^p$.

- The p-value controls the selective type I error for clustering.

González-Delgado *et al.*, 2025.

Standard strategy

1. Choose a model

$$\mathbf{X} \sim \mathcal{MN}_{n\times p}(\boldsymbol{\mu}, \mathbf{U}, \boldsymbol{\Sigma}),$$

where $\boldsymbol{\Sigma}$ can be (over-)estimated and $\mathbf{U}$ is known.

$\rightarrow$ Under (H0), $\mathbf{X}^T\nu \sim \mathcal{N}_p(\mathbf{0}_p, \mathbf{V})$, with $\mathbf{V} = \nu^T\mathbf{U}\nu\boldsymbol{\Sigma}$.

$\rightarrow$ Therefore, $\|\mathbf{X}^T\nu\|^2_{\mathbf{V}} = (\mathbf{X}^T\nu)^T\mathbf{V}^{-1}(\mathbf{X}^T\nu) \sim \chi^2_p$ under the null.

2. Consider a p-value of the form

$$p(\mathbf{x}; \{\mathcal{G}_1, \mathcal{G}_2\}) = \mathbb{P}_{H_0^{\{\mathcal{G}_1,\mathcal{G}_2\}}}\Big(\|\mathbf{X}^T\nu\|_{\mathbf{V}} \geq \|\mathbf{x}^T\nu\|_{\mathbf{V}} \;\Big|\; \mathcal{G}_1, \mathcal{G}_2 \in \mathcal{C}(\mathbf{X}), \boldsymbol{\pi}^{\perp}_{\nu}\mathbf{X} = \boldsymbol{\pi}^{\perp}_{\nu}\mathbf{x}\,,\, \mathrm{dir}_{\mathbf{V}}(\mathbf{X}^T\nu) = \mathrm{dir}_{\mathbf{V}}(\mathbf{x}^T\nu)\Big),$$

where $\boldsymbol{\pi}^{\perp}_{\nu} = \mathbb{I}_n - \frac{\nu\nu^T}{\|\nu\|_2^2}$ and $\mathrm{dir}_{\mathbf{V}}(u) = u/\|u\|_{\mathbf{V}}$ for all $u \in \mathbb{R}^p$.

- The p-value controls the selective type I error for clustering.
- Main challenge : derive an analitically tractable form.

González-Delgado *et al.*, 2025.

Towards a more interpretable p-value

Rewriting the conditioning set

$$\mathbf{X} = \boldsymbol{\pi}_{\nu}^{\perp}\mathbf{X} + (\mathbb{I}_n - \boldsymbol{\pi}_{\nu}^{\perp})\mathbf{X}$$
$$= $$

González-Delgado *et al.*, 2025.

Towards a more interpretable p-value

Rewriting the conditioning set

$$\begin{aligned}\mathbf{X} &= \boldsymbol{\pi}_{\nu}^{\perp}\mathbf{X} + (\mathbb{I}_n - \boldsymbol{\pi}_{\nu}^{\perp})\mathbf{X} \\ &= \boldsymbol{\pi}_{\nu}^{\perp}\mathbf{X} + \frac{\nu\nu^T}{\|\nu\|_2^2}\mathbf{X}\end{aligned}$$

González-Delgado *et al.*, 2025.

Towards a more interpretable p-value

Rewriting the conditioning set

$$\begin{aligned}\mathbf{X} &= \boldsymbol{\pi}_{\nu}^{\perp}\mathbf{X} + (\mathbb{I}_n - \boldsymbol{\pi}_{\nu}^{\perp})\mathbf{X} \\ &= \boldsymbol{\pi}_{\nu}^{\perp}\mathbf{X} + \frac{\nu\nu^T}{\|\nu\|_2^2}\mathbf{X}\end{aligned}$$

González-Delgado *et al.*, 2025.

Towards a more interpretable p-value

Rewriting the conditioning set

$$\begin{aligned}\mathbf{X} &= \boldsymbol{\pi}_{\nu}^{\perp}\mathbf{X} + (\mathbb{I}_n - \boldsymbol{\pi}_{\nu}^{\perp})\mathbf{X} \\ &= \boldsymbol{\pi}_{\nu}^{\perp}\mathbf{X} + \frac{\nu}{\|\nu\|_2^2}\nu^T\mathbf{X}\end{aligned}$$

González-Delgado *et al.*, 2025.

Towards a more interpretable p-value

Rewriting the conditioning set

$$\begin{aligned}\mathbf{X} &= \boldsymbol{\pi}_{\nu}^{\perp}\mathbf{X} + (\mathbb{I}_n - \boldsymbol{\pi}_{\nu}^{\perp})\mathbf{X} \\ &= \boldsymbol{\pi}_{\nu}^{\perp}\mathbf{X} + \frac{\nu}{\|\nu\|_2^2}\nu^T\mathbf{X}\frac{\|\nu^T\mathbf{X}\|_{\mathbf{V}}}{\|\nu^T\mathbf{X}\|_{\mathbf{V}}}\end{aligned}$$

González-Delgado *et al.*, 2025.

Towards a more interpretable p-value

Rewriting the conditioning set

$$\begin{aligned}\mathbf{X} &= \boldsymbol{\pi}_{\nu}^{\perp}\mathbf{X} + (\mathbb{I}_n - \boldsymbol{\pi}_{\nu}^{\perp})\mathbf{X} \\ &= \boldsymbol{\pi}_{\nu}^{\perp}\mathbf{X} + \frac{\nu}{\|\nu\|_2^2}\nu^T\mathbf{X}\frac{\|\nu^T\mathbf{X}\|_{\mathbf{V}}}{\|\nu^T\mathbf{X}\|_{\mathbf{V}}}\end{aligned}$$

González-Delgado *et al.*, 2025.

Towards a more interpretable p-value

Rewriting the conditioning set

$$\begin{aligned}\mathbf{X} &= \boldsymbol{\pi}_{\nu}^{\perp}\mathbf{X} + (\mathbb{I}_n - \boldsymbol{\pi}_{\nu}^{\perp})\mathbf{X} \\ &= \boldsymbol{\pi}_{\nu}^{\perp}\mathbf{X} + \frac{\nu}{\|\nu\|_2^2}\|\nu^T\mathbf{X}\|_{\mathbf{V}}\mathrm{dir}_{\mathbf{V}}(\nu^T\mathbf{X})\end{aligned}$$

González-Delgado *et al.*, 2025.

Towards a more interpretable p-value

Rewriting the conditioning set

$$\begin{aligned}\mathbf{X} &= \boldsymbol{\pi}_{\nu}^{\perp}\mathbf{X} + (\mathbb{I}_n - \boldsymbol{\pi}_{\nu}^{\perp})\mathbf{X} \\ &= \boldsymbol{\pi}_{\nu}^{\perp}\mathbf{X} + \frac{\nu}{\|\nu\|_2^2}\|\nu^T\mathbf{X}\|_{\mathbf{V}}\mathrm{dir}_{\mathbf{V}}(\nu^T\mathbf{X})\end{aligned}$$

González-Delgado *et al.*, 2025.

Towards a more interpretable p-value

Rewriting the conditioning set

$$\begin{aligned}\mathbf{X} &= \boldsymbol{\pi}_{\nu}^{\perp}\mathbf{X} + (\mathbb{I}_n - \boldsymbol{\pi}_{\nu}^{\perp})\mathbf{X} \\ &= \boldsymbol{\pi}_{\nu}^{\perp}\mathbf{X} + \frac{\nu}{\|\nu\|_2^2}\|\nu^T\mathbf{X}\|_{\mathbf{V}}\mathrm{dir}_{\mathbf{V}}(\nu^T\mathbf{X})\end{aligned}$$

González-Delgado *et al.*, 2025.

Towards a more interpretable p-value

Rewriting the conditioning set

$$\begin{aligned}\mathbf{X} &= \boldsymbol{\pi}_{\nu}^{\perp}\mathbf{X} + (\mathbb{I}_n - \boldsymbol{\pi}_{\nu}^{\perp})\mathbf{X} \\ &= \boldsymbol{\pi}_{\nu}^{\perp}\mathbf{X} + \frac{\nu}{\|\nu\|_2^2}\|\nu^T\mathbf{X}\|_{\mathbf{V}}\mathrm{dir}_{\mathbf{V}}(\nu^T\mathbf{X})\end{aligned}$$

$$p(\mathbf{x}; \{\mathcal{G}_1, \mathcal{G}_2\}) = \mathbb{P}_{H_0^{\{\mathcal{G}_1,\mathcal{G}_2\}}}\Big(\|\mathbf{X}^T\nu\|_{\mathbf{V}} \geq \|\mathbf{x}^T\nu\|_{\mathbf{V}} \;\Big|$$

$$\mathcal{G}_1, \mathcal{G}_2 \in \mathcal{C}(\mathbf{X}),\ \boldsymbol{\pi}_{\nu}^{\perp}\mathbf{X} = \boldsymbol{\pi}_{\nu}^{\perp}\mathbf{x}\,,\ \mathrm{dir}_{\mathbf{V}}(\mathbf{X}^T\nu) = \mathrm{dir}_{\mathbf{V}}(\mathbf{x}^T\nu)\Big)$$

González-Delgado *et al.*, 2025.

Towards a more interpretable p-value

Rewriting the conditioning set

$$\begin{aligned}\mathbf{X} &= \boldsymbol{\pi}_{\nu}^{\perp}\mathbf{X} + (\mathbb{I}_n - \boldsymbol{\pi}_{\nu}^{\perp})\mathbf{X} \\ &= \boldsymbol{\pi}_{\nu}^{\perp}\mathbf{X} + \frac{\nu}{\|\nu\|_2^2}\|\nu^T\mathbf{X}\|_{\mathbf{V}}\mathrm{dir}_{\mathbf{V}}(\nu^T\mathbf{X})\end{aligned}$$

$$p(\mathbf{x}; \{\mathcal{G}_1, \mathcal{G}_2\}) = \mathbb{P}_{H_0^{\{\mathcal{G}_1,\mathcal{G}_2\}}}\Big(\|\mathbf{X}^T\nu\|_{\mathbf{V}} \geq \|\mathbf{x}^T\nu\|_{\mathbf{V}} \;\Big|$$

$$\mathcal{G}_1, \mathcal{G}_2 \in \mathcal{C}\Big(\boldsymbol{\pi}_{\nu}^{\perp}\mathbf{X} + \frac{\nu}{\|\nu\|_2^2}\|\nu^T\mathbf{X}\|_{\mathbf{V}}\mathrm{dir}_{\mathbf{V}}(\nu^T\mathbf{X})\Big),\ \boldsymbol{\pi}_{\nu}^{\perp}\mathbf{X} = \boldsymbol{\pi}_{\nu}^{\perp}\mathbf{x},\ \mathrm{dir}_{\mathbf{V}}(\mathbf{X}^T\nu) = \mathrm{dir}_{\mathbf{V}}(\mathbf{x}^T\nu)\Big)$$

González-Delgado *et al.*, 2025.

Towards a more interpretable p-value

Rewriting the conditioning set

$$p(\mathbf{x}; \{\mathcal{G}_1, \mathcal{G}_2\}) = \mathbb{P}_{H_0^{\{\mathcal{G}_1, \mathcal{G}_2\}}} \Big(\|\mathbf{X}^T \nu\|_{\mathbf{V}} \geq \|\mathbf{x}^T \nu\|_{\mathbf{V}} \;\Big|$$

$$\mathcal{G}_1, \mathcal{G}_2 \in \mathcal{C}\Big(\boldsymbol{\pi}_\nu^\perp \mathbf{X} + \frac{\nu}{\|\nu\|_2^2} \|\nu^T \mathbf{X}\|_{\mathbf{V}} \mathrm{dir}_{\mathbf{V}}(\nu^T \mathbf{X}) \Big),\ \boldsymbol{\pi}_\nu^\perp \mathbf{X} = \boldsymbol{\pi}_\nu^\perp \mathbf{x}\,,\ \mathrm{dir}_{\mathbf{V}}(\mathbf{X}^T \nu) = \mathrm{dir}_{\mathbf{V}}(\mathbf{x}^T \nu) \Big)$$

González-Delgado *et al.*, 2025.

Towards a more interpretable p-value

Rewriting the conditioning set

$$p(\mathbf{x}; \{\mathcal{G}_1, \mathcal{G}_2\}) = \mathbb{P}_{H_0^{\{\mathcal{G}_1, \mathcal{G}_2\}}} \Big(\|\mathbf{X}^T \nu\|_{\mathbf{V}} \geq \|\mathbf{x}^T \nu\|_{\mathbf{V}} \;\Big|$$

$$\mathcal{G}_1, \mathcal{G}_2 \in \mathcal{C}\Big(\boldsymbol{\pi}_\nu^\perp \mathbf{x} + \frac{\nu}{\|\nu\|_2^2} \|\nu^T \mathbf{X}\|_{\mathbf{V}} \mathrm{dir}_{\mathbf{V}}(\nu^T \mathbf{x})\Big),\ \boldsymbol{\pi}_\nu^\perp \mathbf{X} = \boldsymbol{\pi}_\nu^\perp \mathbf{x},\ \mathrm{dir}_{\mathbf{V}}(\mathbf{X}^T \nu) = \mathrm{dir}_{\mathbf{V}}(\mathbf{x}^T \nu)\Big)$$

González-Delgado *et al.*, 2025.

Towards a more interpretable p-value

Rewriting the conditioning set

$$p(\mathbf{x}; \{\mathcal{G}_1, \mathcal{G}_2\}) = \mathbb{P}_{H_0^{\{\mathcal{G}_1,\mathcal{G}_2\}}}\Big(\|\mathbf{X}^T\nu\|_{\mathbf{V}} \geq \|\mathbf{x}^T\nu\|_{\mathbf{V}} \;\Big|$$

$$\mathcal{G}_1, \mathcal{G}_2 \in \mathcal{C}\Big(\boldsymbol{\pi}_\nu^\perp \mathbf{x} + \frac{\nu}{\|\nu\|_2^2}\|\nu^T\mathbf{X}\|_{\mathbf{V}}\mathrm{dir}_{\mathbf{V}}(\nu^T\mathbf{x})\Big),\ \boldsymbol{\pi}_\nu^\perp \mathbf{X} = \boldsymbol{\pi}_\nu^\perp \mathbf{x},\ \underbrace{\mathrm{dir}_{\mathbf{V}}(\mathbf{X}^T\nu)}_{\perp\!\!\!\perp \|\mathbf{X}^T\nu\|_{\mathbf{V}}} = \mathrm{dir}_{\mathbf{V}}(\mathbf{x}^T\nu)\Big)$$

González-Delgado *et al.*, 2025.

Towards a more interpretable p-value

Rewriting the conditioning set

$$p(\mathbf{x}; \{\mathcal{G}_1, \mathcal{G}_2\}) = \mathbb{P}_{H_0^{\{\mathcal{G}_1,\mathcal{G}_2\}}}\Big(\|\mathbf{X}^T\nu\|_{\mathbf{V}} \geq \|\mathbf{x}^T\nu\|_{\mathbf{V}} \;\Big|$$

$$\mathcal{G}_1, \mathcal{G}_2 \in \mathcal{C}\Big(\boldsymbol{\pi}_\nu^\perp \mathbf{x} + \frac{\nu}{\|\nu\|_2^2}\|\nu^T\mathbf{X}\|_{\mathbf{V}}\mathrm{dir}_{\mathbf{V}}(\nu^T\mathbf{x})\Big), \quad \underbrace{\boldsymbol{\pi}_\nu^\perp \mathbf{X}}_{\substack{\perp\!\!\!\perp \|\mathbf{X}^T\nu\|_{\mathbf{V}} \\ \Leftrightarrow \mathbf{U} \in \mathcal{CS}(n)}} = \boldsymbol{\pi}_\nu^\perp \mathbf{x}\,,\ \underbrace{\mathrm{dir}_{\mathbf{V}}(\mathbf{X}^T\nu)}_{\perp\!\!\!\perp \|\mathbf{X}^T\nu\|_{\mathbf{V}}} = \mathrm{dir}_{\mathbf{V}}(\mathbf{x}^T\nu)\Big)$$

González-Delgado *et al.*, 2025.

Towards a more interpretable p-value

Rewriting the conditioning set

$$p(\mathbf{x}; \{\mathcal{G}_1, \mathcal{G}_2\}) = \mathbb{P}_{H_0^{\{\mathcal{G}_1, \mathcal{G}_2\}}} \Big(\|\mathbf{X}^T \nu\|_{\mathbf{V}} \geq \|\mathbf{x}^T \nu\|_{\mathbf{V}} \;\Big|\; \mathcal{G}_1, \mathcal{G}_2 \in \mathcal{C}\Big(\boldsymbol{\pi}_{\nu}^{\perp}\mathbf{x} + \frac{\nu}{\|\nu\|_2^2}\|\nu^T\mathbf{X}\|_{\mathbf{V}}\mathrm{dir}_{\mathbf{V}}(\nu^T\mathbf{x})\Big)\Big)$$

González-Delgado *et al.*, 2025.

Towards a more interpretable p-value

Rewriting the conditioning set

$$p(\mathbf{x}; \{\mathcal{G}_1, \mathcal{G}_2\}) = \mathbb{P}_{H_0^{\{\mathcal{G}_1, \mathcal{G}_2\}}} \left(\phi \geq \|\mathbf{x}^T \nu\|_{\mathbf{V}} \;\middle|\; \mathcal{G}_1, \mathcal{G}_2 \in \mathcal{C}\left(\boldsymbol{\pi}_{\nu}^{\perp} \mathbf{x} + \frac{\nu}{\|\nu\|_2^2} \phi \, \mathrm{dir}_{\mathbf{V}}(\nu^T \mathbf{x}) \right) \right)$$

González-Delgado *et al.*, 2025.

Towards a more interpretable p-value

Rewriting the conditioning set

$$p(\mathbf{x}; \{\mathcal{G}_1, \mathcal{G}_2\}) = \mathbb{P}_{H_0^{\{\mathcal{G}_1, \mathcal{G}_2\}}} \left(\phi \geq \|\mathbf{x}^T \nu\|_{\mathbf{V}} \;\middle|\; \mathcal{G}_1, \mathcal{G}_2 \in \mathcal{C}\left(\mathbf{x} + \frac{\nu}{\|\nu\|_2^2} (\phi - \|\mathbf{x}^T \nu\|_{\mathbf{V}}) \, \mathrm{dir}_{\mathbf{V}}(\mathbf{x}^T \nu) \right) \right)$$

González-Delgado *et al.*, 2025.

Towards a more interpretable p-value

Rewriting the conditioning set

$$p(\mathbf{x}; \{\mathcal{G}_1, \mathcal{G}_2\}) = \mathbb{P}_{H_0^{\{\mathcal{G}_1,\mathcal{G}_2\}}}\left(\phi \geq \|\mathbf{x}^T\nu\|_{\mathbf{V}} \;\middle|\; \mathcal{G}_1, \mathcal{G}_2 \in \mathcal{C}\Big(\underbrace{\mathbf{x} + \frac{\nu}{\|\nu\|_2^2}(\phi - \|\mathbf{x}^T\nu\|_{\mathbf{V}})\,\mathrm{dir}_{\mathbf{V}}(\mathbf{x}^T\nu)}_{\mathbf{x}'(\phi)}\Big)\right)$$

González-Delgado *et al.*, 2025.

Towards a more interpretable p-value

Rewriting the conditioning set

$$p(\mathbf{x}; \{\mathcal{G}_1, \mathcal{G}_2\}) = \mathbb{P}_{H_0^{\{\mathcal{G}_1, \mathcal{G}_2\}}} \left(\phi \geq \|\mathbf{x}^T \nu\|_{\mathbf{V}} \;\middle|\; \mathcal{G}_1, \mathcal{G}_2 \in \mathcal{C}(\mathbf{x}'(\phi)) \right)$$

González-Delgado *et al.*, 2025.

Towards a more interpretable p-value

Rewriting the conditioning set

$$p(\mathbf{x}; \{\mathcal{G}_1, \mathcal{G}_2\}) = \mathbb{P}_{H_0^{\{\mathcal{G}_1, \mathcal{G}_2\}}} \Big(\phi \geq \|\mathbf{x}^T \nu\|_{\mathbf{V}} \;\Big|\; \mathcal{G}_1, \mathcal{G}_2 \in \mathcal{C}\big(\mathbf{x}'(\phi)\big) \Big), \quad \text{with } \phi \sim \chi_p.$$

González-Delgado *et al.*, 2025.

Towards a more interpretable p-value

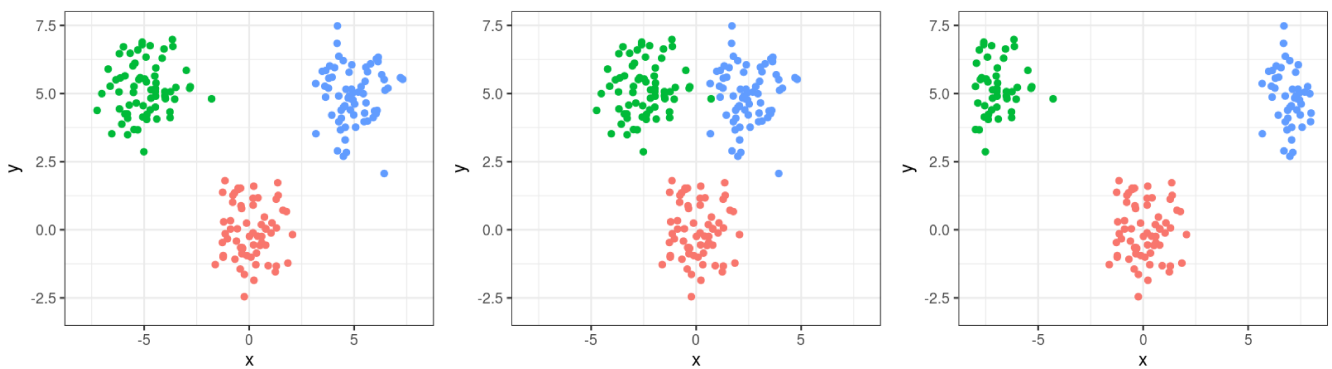
Rewriting the conditioning set

$$p(\mathbf{x}; \{\mathcal{G}_1, \mathcal{G}_2\}) = \mathbb{P}_{H_0^{\{\mathcal{G}_1,\mathcal{G}_2\}}}\left(\phi \geq \|\mathbf{x}^T\nu\|_{\mathbf{V}} \;\middle|\; \mathcal{G}_1, \mathcal{G}_2 \in \mathcal{C}(\mathbf{x}'(\phi))\right), \quad \text{with } \phi \sim \chi_p.$$

$\mathbf{x}'(\phi)$ is a perturbation of $\mathbf{x}$ such that

- If $\phi \geq \|\mathbf{x}^T\nu\|_{\mathbf{V}}$, $\mathcal{G}_1$ and $\mathcal{G}_2$ are pulled appart,
- If $\phi \leq \|\mathbf{x}^T\nu\|_{\mathbf{V}}$, $\mathcal{G}_1$ and $\mathcal{G}_2$ are pushed together.

$\phi = \|\bar{x}_{\mathcal{G}_1} - \bar{x}_{\mathcal{G}_2}\|$ $\phi \leq \|\bar{x}_{\mathcal{G}_1} - \bar{x}_{\mathcal{G}_2}\|$ $\phi \geq \|\bar{x}_{\mathcal{G}_1} - \bar{x}_{\mathcal{G}_2}\|$



González-Delgado *et al.*, 2025.

Interpretation of p-value

A consequence of conditioning

$$p(x) = \mathbb{P}_{H_0^{\{\mathcal{G}_1,\mathcal{G}_2\}}}\left(\phi \geq \|\mathbf{x}^T\nu\|_{\mathbf{V}} \,\middle|\, \mathcal{G}_1, \mathcal{G}_2 \in \mathcal{C}\left(\mathbf{x}'(\phi)\right)\right)$$
$$= \frac{\mathbb{E}\left[\mathbb{1}\{\phi \geq \|\mathbf{x}^T\nu\|_{\mathbf{V}}, \mathcal{G}_1, \mathcal{G}_2 \in \mathcal{C}\left(\mathbf{x}'(\phi)\right)\}\right]}{\mathbb{E}[\mathbb{1}\{\mathcal{G}_1, \mathcal{G}_2 \in \mathcal{C}\left(\mathbf{x}'(\phi)\right)\}]}$$

Interpretation of p-value

A consequence of conditioning

$$p(x) = \mathbb{P}_{H_0^{\{\mathcal{G}_1,\mathcal{G}_2\}}}\left(\phi \geq \|\mathbf{x}^T\nu\|_{\mathbf{V}} \,\middle|\, \mathcal{G}_1, \mathcal{G}_2 \in \mathcal{C}\left(\mathbf{x}'(\phi)\right)\right)$$
$$= \frac{\mathbb{E}\left[\mathbb{1}\{\phi \geq \|\mathbf{x}^T\nu\|_{\mathbf{V}}, \mathcal{G}_1, \mathcal{G}_2 \in \mathcal{C}\left(\mathbf{x}'(\phi)\right)\}\right]}{\mathbb{E}[\mathbb{1}\{\mathcal{G}_1, \mathcal{G}_2 \in \mathcal{C}\left(\mathbf{x}'(\phi)\right)\}]}$$

Informally

Do I have room to push the clusters together while preserving the clustering ?

Interpretation of p-value

A consequence of conditioning

$$p(x) = \mathbb{P}_{H_0^{\{\mathcal{G}_1,\mathcal{G}_2\}}}\left(\phi \geq \|\mathbf{x}^T\nu\|_{\mathbf{V}} \,\middle|\, \mathcal{G}_1, \mathcal{G}_2 \in \mathcal{C}\left(\mathbf{x}'(\phi)\right)\right)$$
$$= \frac{\mathbb{E}\left[\mathbb{1}\{\phi \geq \|\mathbf{x}^T\nu\|_{\mathbf{V}}, \mathcal{G}_1, \mathcal{G}_2 \in \mathcal{C}\left(\mathbf{x}'(\phi)\right)\}\right]}{\mathbb{E}[\mathbb{1}\{\mathcal{G}_1, \mathcal{G}_2 \in \mathcal{C}\left(\mathbf{x}'(\phi)\right)\}]}$$

Informally

Do I have room to push the clusters together while preserving the clustering ?

- Yes $\rightarrow$ reject (H0) $\rightarrow$ different clusters.
- No $\rightarrow$ accept (H0) $\rightarrow$ same cluster.

Interpretation of p-value

A consequence of conditioning

$$p(x) = \mathbb{P}_{H_0^{\{\mathcal{G}_1,\mathcal{G}_2\}}}\left(\phi \geq \|\mathbf{x}^T\nu\|_{\mathbf{V}} \,\middle|\, \mathcal{G}_1, \mathcal{G}_2 \in \mathcal{C}\left(\mathbf{x}'(\phi)\right)\right)$$
$$= \frac{\mathbb{E}\left[\mathbb{1}\{\phi \geq \|\mathbf{x}^T\nu\|_{\mathbf{V}}, \mathcal{G}_1, \mathcal{G}_2 \in \mathcal{C}\left(\mathbf{x}'(\phi)\right)\}\right]}{\mathbb{E}[\mathbb{1}\{\mathcal{G}_1, \mathcal{G}_2 \in \mathcal{C}\left(\mathbf{x}'(\phi)\right)\}]}$$

Informally

Do I have room to push the clusters together while preserving the clustering ?

- Yes $\rightarrow$ reject (H0) $\rightarrow$ different clusters.
- No $\rightarrow$ accept (H0) $\rightarrow$ same cluster.

In practice

- We need $\{\phi : \mathcal{G}_1, \mathcal{G}_2 \in \mathcal{C}(\mathbf{x}'(\phi))\} \neq \emptyset$.

Interpretation of p-value

A consequence of conditioning

$$p(x) = \mathbb{P}_{H_0^{\{\mathcal{G}_1,\mathcal{G}_2\}}}\left(\phi \geq \|\mathbf{x}^T\nu\|_{\mathbf{V}} \,\middle|\, \mathcal{G}_1, \mathcal{G}_2 \in \mathcal{C}\left(\mathbf{x}'(\phi)\right)\right)$$
$$= \frac{\mathbb{E}\left[\mathbb{1}\{\phi \geq \|\mathbf{x}^T\nu\|_{\mathbf{V}}, \mathcal{G}_1, \mathcal{G}_2 \in \mathcal{C}\left(\mathbf{x}'(\phi)\right)\}\right]}{\mathbb{E}[\mathbb{1}\{\mathcal{G}_1, \mathcal{G}_2 \in \mathcal{C}\left(\mathbf{x}'(\phi)\right)\}]}$$

Informally

Do I have room to push the clusters together while preserving the clustering ?

- Yes $\rightarrow$ reject (H0) $\rightarrow$ different clusters.
- No $\rightarrow$ accept (H0) $\rightarrow$ same cluster.

In practice

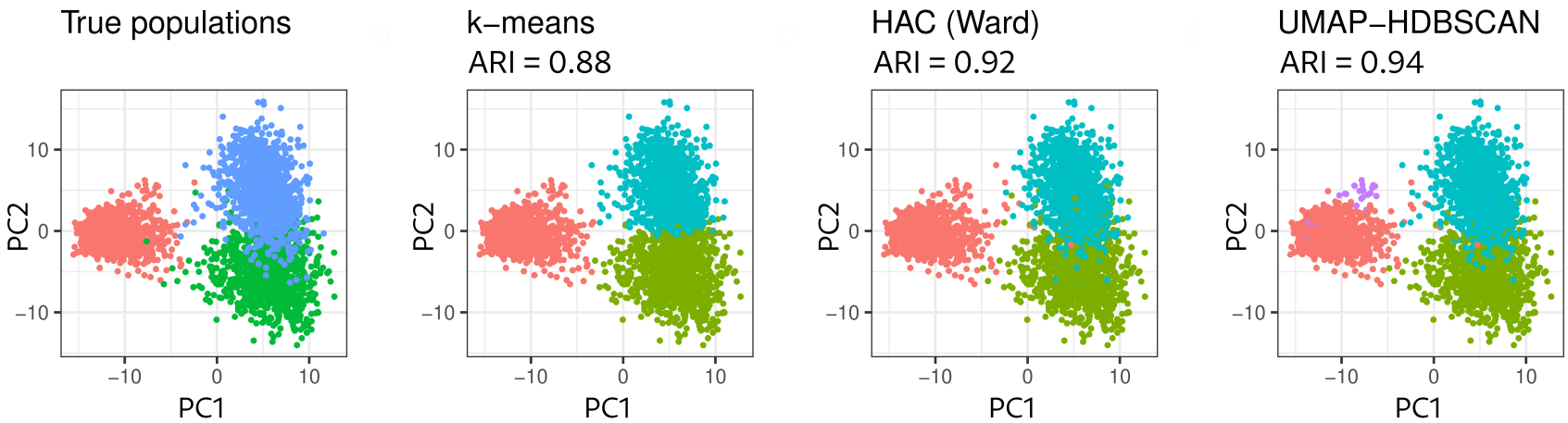
- We need $\{\phi : \mathcal{G}_1, \mathcal{G}_2 \in \mathcal{C}(\mathbf{x}'(\phi))\} \neq \emptyset$.
- If $\{\mathcal{G}_1, \mathcal{G}_2 \in \mathcal{C}(\mathbf{x}'(\phi))\}$ is unlikely, p-values will be very close to 0 or 1.

Consequences of conditioning (I)

Clustering need to be robust to small perturbations

Simulation setting

SNPs for $n = 3000$ individuals equally sampled from three populations[1] + Clustering performed on the first 100 PCs.



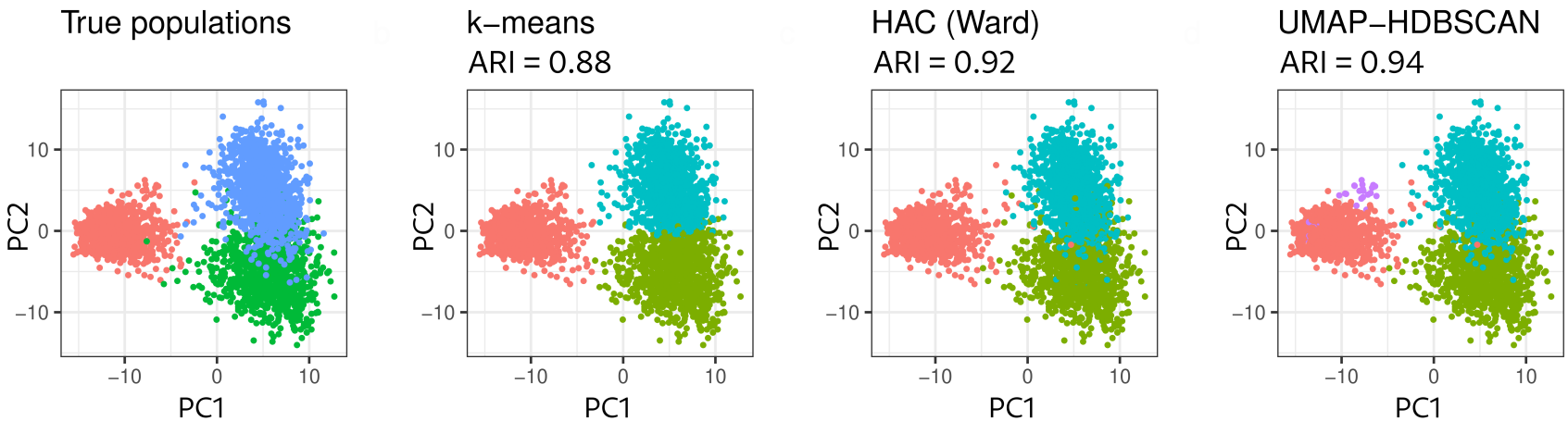
1. Data : Diaz-Papkovich, *unpublished*.

Consequences of conditioning (I)

Clustering need to be robust to small perturbations

Simulation setting

SNPs for $n = 3000$ individuals equally sampled from three populations[1] + Clustering performed on the first 100 PCs.



$\{\mathcal{G}_1, \mathcal{G}_2\}$	{●, ●}	{●, ●}	{●, ●}
k-means	0.279	0.554	0.178
HAC (Ward)	$< 10^{-16}$	$< 10^{-16}$	$< 10^{-16}$
UMAP+HDBSCAN	1	0.002	0.002

p-values obtained using `PCIdep` (González-Delgado *et al.* 2025).

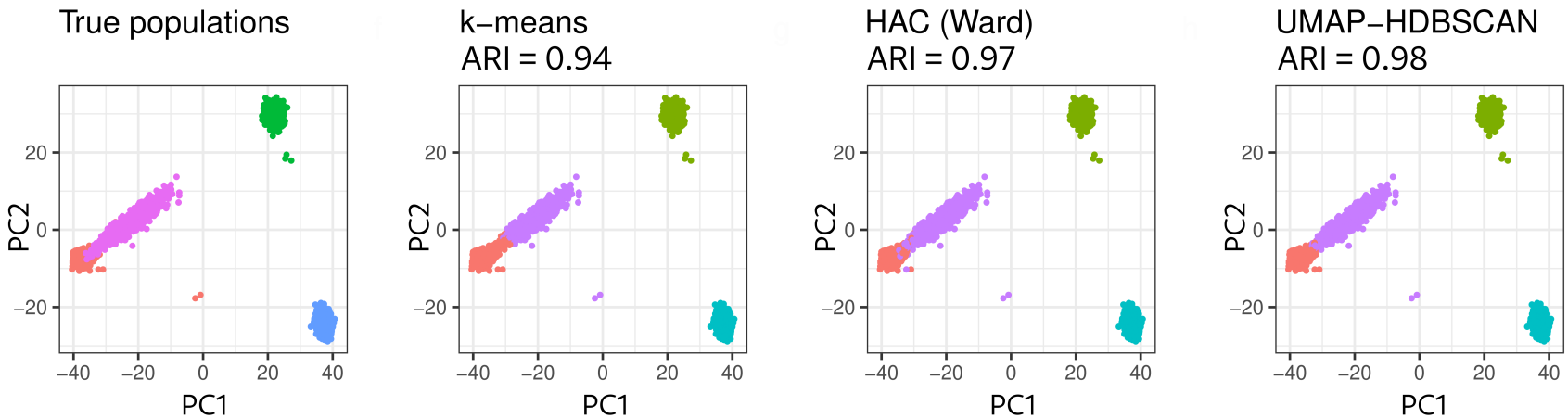
1. Data : Diaz-Papkovich, *unpublished*.

Consequences of conditioning (II)

Shapes and relative positions of clusters may hinder inference

Simulation setting

SNPs for $n = 4000$ individuals equally sampled from four populations : two sources, one unrelated population, and an admixed population derived from the sources[1] + Clustering performed on the first 100 PCs.



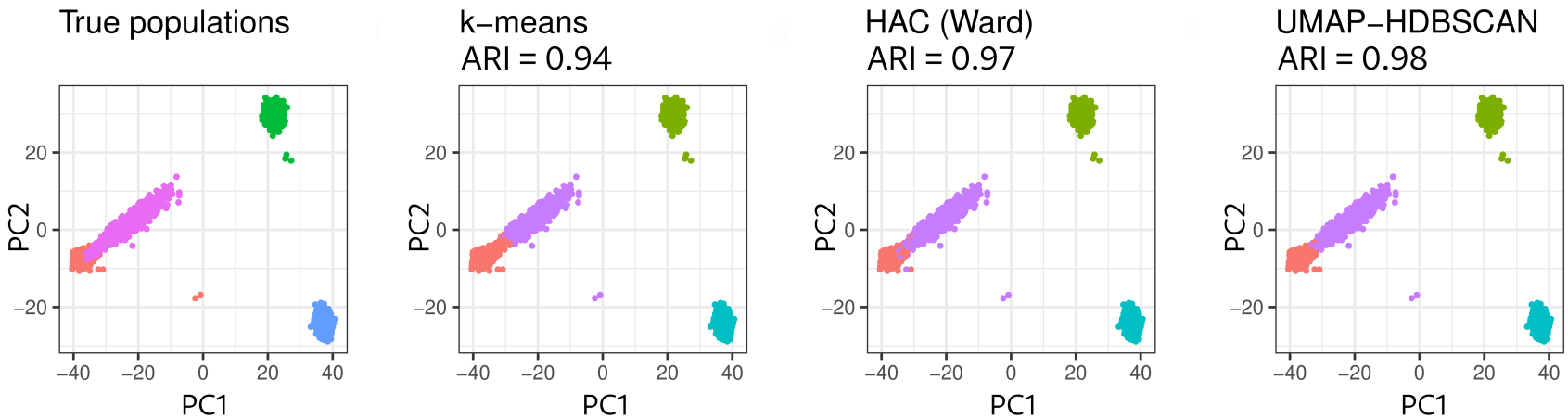
1. Data : Diaz-Papkovich, *unpublished*.

Consequences of conditioning (II)

Shapes and relative positions of clusters may hinder inference

Simulation setting

SNPs for $n = 4000$ individuals equally sampled from four populations : two sources, one unrelated population, and an admixed population derived from the sources[1] + Clustering performed on the first 100 PCs.



$\{\mathcal{G}_1, \mathcal{G}_2\}$	{●, ●}	{●, ●}	{●, ●}	{●, ●}	{●, ●}	{●, ●}
k-means	0.118	0.113	0.005	0.189	0.767	0.667
HAC	0.084	0.011	$< 10^{-16}$	0.343	$< 10^{-16}$	$< 10^{-16}$
UMAP+HDBSCAN	1	1	0.002	0.002	0.002	0.002

p-values obtained using `PCIdep` (González-Delgado *et al.* 2025).

1. Data : Diaz-Papkovich, *unpublished*.

Conclusion and perspectives

Main message

Conditional inference inherently limits post-clustering inference.

Conclusion and perspectives

Main message

Conditional inference inherently limits post-clustering inference.

The theory appears unsuitable in multiple practical applications

- Clustering algorithms sensitive to smooth boundaries,
- Realistic cluster placements.

Conclusion and perspectives

Main message

Conditional inference inherently limits post-clustering inference.

The theory appears unsuitable in multiple practical applications

- Clustering algorithms sensitive to smooth boundaries,
- Realistic cluster placements.

Limitations come from conditioning

Advances in this theory should proceed through a new selective inference framework, such as simultaneous inference :

$$\mathbb{P}_{H_0^{\{\mathcal{G}_1,\mathcal{G}_2\}}}\left(\forall\,\{\mathcal{G}_1,\mathcal{G}_2\}\ :\ \text{Reject } H_0^{\{\mathcal{G}_1,\mathcal{G}_2\}} \text{ based on } \mathbf{X} \text{ at level } \alpha\right) \leq \alpha.$$

Ongoing exploration...

(Some) references

Conditional inference methods

- Independence setting : L. L. Gao, J. Bien, and D. Witten. Selective inference for hierarchical clustering. *Journal of the American Statistical Association*, 119(545), 332–342, 2024.
- Extension to feature-level test : B. Hivert, D. Agniel, R. Thiébaut, and B. P. Hejblum. Post-clustering difference testing : Valid inference and practical considerations with applications to ecological and biological data. *Comput. Statist. Data Anal.*, 193 :107916, May 2024.
- General matrix normal model : J. González-Delgado, M. Deronzier, J. Cortés and P. Neuvial. Post-clustering inference under dependence, 2025. https://arxiv.org/abs/2310.11822.

Information partitioning approaches

- Data fission : J. Leiner, B. Duan, L. Wasserman and A. Ramdas. Data Fission : Splitting a Single Data Point. *Journal of the American Statistical Association*, 120(549), 135–146, 2025.
- Data thinning : A. Neufeld, A. Dharamshi, L. L. Gao, and D. Witten. Data thinning for convolution-closed distributions. *Journal of Machine Learning Research*, 25 :1-35, 2024.
- Limitations in post-clustering inference : B. Hivert, D. Agniel, R. Thiébaut and B. P. Hejblum. Practical limitations for real-life application of data fission and data thinning in post-clustering differential analysis, 2025. https://arxiv.org/abs/2405.13591.

Thank you for your attention !

https://gonzalez-delgado.github.io/