# Post-clustering inference under dependency

Javier González-Delgado[1,2] and Pierre Neuvial[1]

1. Institut de Mathématiques de Toulouse, 2. LAAS-CNRS

**Séminaire rennais de Statistiques**

October 11, 2023


# Outline

- The problem of selective (and post-clustering) inference
- Post-clustering inference with known model parameters
- Post-clustering inference with compatible parameter estimation
- Clustering to describe ensembles of highly flexible proteins

# The problem of selective inference

## Statistical investigation

1. **Selection** : Choose a statistical model for the data and formulate a problem
2. **Inference** : Investigate the chosen problem using the data

# The problem of selective inference

## Statistical investigation

1. **Selection** : Choose a statistical model for the data and formulate a problem
2. **Inference** : Investigate the chosen problem using the data

## Selection step

- **Non-adaptive** : Selection is made before seeing the data
- **Adaptive** : Selection is made at least partially guided by the data

# The problem of selective inference

## Statistical investigation

1. **Selection** : Choose a statistical model for the data and formulate a problem
2. **Inference** : Investigate the chosen problem using the data

## Selection step

- **Non-adaptive** : Selection is made before seeing the data
- **Adaptive** : Selection is made at least partially guided by the data

Goal of selective inference : valid inference in presence of adaptive selection

# The problem of selective inference

## Example (file drawer effect)

$Y \sim \mathcal{N}(\mu, 1)$ $\rightarrow H_0 : \mu = 0$ (reject at level $\alpha = 0.05$)

Fithian *et al.*, *Optimal Inference After Model Selection*, 2017.

# The problem of selective inference

## Example (file drawer effect)

$$Y \sim \mathcal{N}(\mu, 1) \rightarrow \underbrace{\boxed{\text{Test if } |Y| > 1}}_{\text{Selection}} \rightarrow H_0 : \mu = 0 \text{ (reject at level } \alpha = 0.05)$$

Fithian *et al.*, *Optimal Inference After Model Selection*, 2017.

# The problem of selective inference

## Example (file drawer effect)

For $i \in \{1, \ldots, n\}$ :

$$Y_i \sim \mathcal{N}(\mu_i, 1) \rightarrow \underbrace{\boxed{\text{Test if } |Y_i| > 1}}_{\text{Selection}} \rightarrow H_{0,i} : \mu_i = 0 \text{ (reject at level } \alpha = 0.05)$$

Fithian *et al.*, *Optimal Inference After Model Selection*, 2017.

# The problem of selective inference

## Example (file drawer effect)

For $i \in \{1, \dots, n\}$ :

$$Y_i \sim \mathcal{N}(\mu_i, 1) \rightarrow \underbrace{\boxed{\text{Test if } |Y_i| > 1}}_{\text{Selection}} \rightarrow H_{0,i} : \mu_i = 0 \text{ (reject at level } \alpha = 0.05)$$

$$\frac{\#\text{False rejections}}{\#\text{True nulls selected}} \xrightarrow[n]{} \frac{\mathbb{P}_{H_0}(\text{Reject } H_0 \,,\, H_0 \text{ tested})}{\mathbb{P}_{H_0}(H_0 \text{ tested})} = \frac{\Phi(-1.96)}{\Phi(-1)} \approx 0.16 \gg \alpha$$

Fithian *et al.*, *Optimal Inference After Model Selection*, 2017.

# The problem of selective inference

## Example (file drawer effect)

For $i \in \{1, \dots, n\}$ :

$$Y_i \sim \mathcal{N}(\mu_i, 1) \to \underbrace{\boxed{\text{Test if } |Y_i| > 1}}_{\text{Selection}} \to H_{0,i} : \mu_i = 0 \text{ (reject at level } \alpha = 0.05)$$

$$\frac{\#\text{False rejections}}{\#\text{True nulls selected}} \xrightarrow[n]{} \frac{\mathbb{P}_{H_0}(\text{Reject } H_0 \,,\, H_0 \text{ tested})}{\mathbb{P}_{H_0}(H_0 \text{ tested})} = \frac{\Phi(-1.96)}{\Phi(-1)} \approx 0.16 \gg \alpha$$

Control the **selective type I error**

$$\mathbb{P}_{H_0}(\text{Reject } H_0 \mid H_0 \text{ tested})$$

"The answer must be valid, given that the question was asked"

Fithian *et al.*, *Optimal Inference After Model Selection*, 2017.

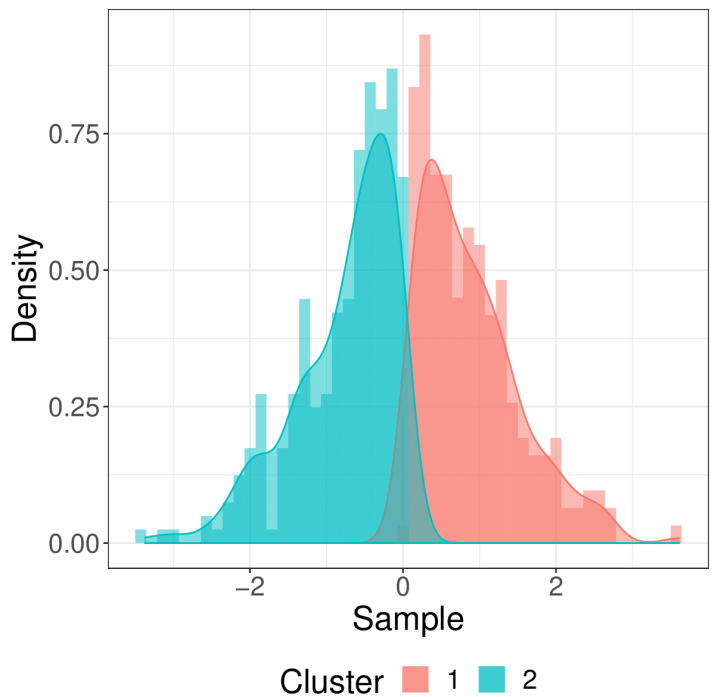
# Post-clustering inference

## Toy example

# Post-clustering inference

## Toy example

- Simulate $\mathcal{N}(0, 1) + \mathcal{U}(-0.2, 0.2)$
- Ask $k$-means to find 2 clusters (data-driven hypothesis selection)
- Test for the difference of cluster means (inference after selection)

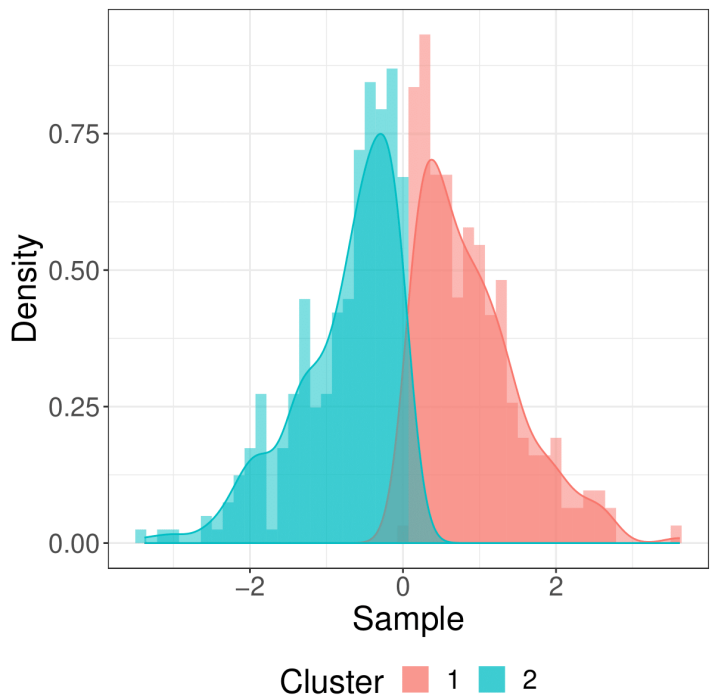



Adapted from Hivert *et al.* 2022

# Post-clustering inference

## Toy example

- Simulate $\mathcal{N}(0, 1) + \mathcal{U}(-0.2, 0.2)$
- Ask $k$-means to find 2 clusters (data-driven hypothesis selection)
- Test for the difference of cluster means (inference after selection)



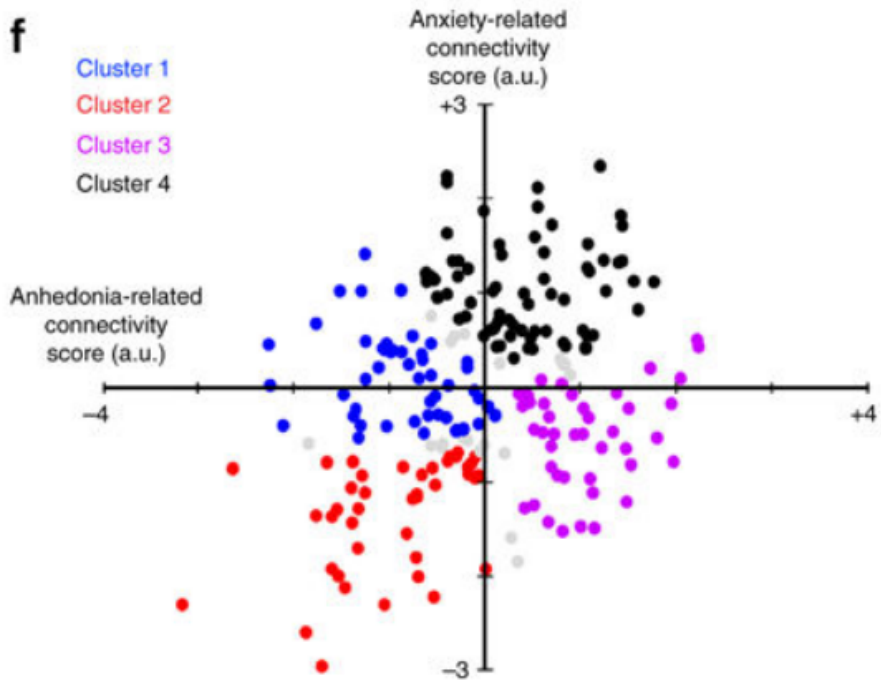

$p_Z = 10^{-67}$, $p_{SI} = 0.84$ (using Chen and Witten 2023).

Adapted from Hivert *et al.* 2022

# Post-clustering inference

## The risk of ignoring the selection step

Drysdale *et al.*, Resting-state connectivity biomarkers define neurophysiological subtypes of depression, *Nat Med.* 2017.




Refuted using post-clustering inference by Dinga *et al.* (2019)

# Post-clustering inference

## General strategy

### Notation

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

# Post-clustering inference

## General strategy

### Notation

- 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. $\mu_i$) denote the $i$-th row of $\mathbf{X}$ (resp. $\boldsymbol{\mu}$) for $i \in \{1, \ldots, n\}$.

# Post-clustering inference

## General strategy

### Notation

- 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. $\mu_i$) denote the $i$-th row of $\mathbf{X}$ (resp. $\boldsymbol{\mu}$) for $i \in \{1, \ldots, n\}$.
- 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 $\bar{\mu}_{\mathcal{G}} = \frac{1}{|\mathcal{G}|} \sum_{i \in \mathcal{G}} \mu_i$.

# Post-clustering inference

## General strategy

### Notation

- 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. $\mu_i$) denote the $i$-th row of $\mathbf{X}$ (resp. $\boldsymbol{\mu}$) for $i \in \{1, \dots, n\}$.
- 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 $\hat{C}_1, \hat{C}_2 \subset \{1, \dots, n\}$ be two clusters estimated by $C(\cdot)$ on $\mathbf{X}$, that is, $\hat{C}_1, \hat{C}_2 \in \mathcal{C}(\mathbf{X})$.

# Post-clustering inference

## General strategy

### Notation

- 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. $\mu_i$) denote the $i$-th row of $\mathbf{X}$ (resp. $\boldsymbol{\mu}$) for $i \in \{1, \dots, n\}$.
- 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 $\hat{C}_1, \hat{C}_2 \subset \{1, \dots, n\}$ be two clusters estimated by $C(\cdot)$ on $\mathbf{X}$, that is, $\hat{C}_1, \hat{C}_2 \in \mathcal{C}(\mathbf{X})$.
- Consider the null hypothesis $H_0^{\{\hat{C}_1, \hat{C}_2\}} : \bar{\mu}_{\hat{C}_1} = \bar{\mu}_{\hat{C}_2}$.

# Post-clustering inference

## General strategy

### Notation

- 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. $\mu_i$) denote the $i$-th row of $\mathbf{X}$ (resp. $\boldsymbol{\mu}$) for $i \in \{1, \dots, n\}$.
- 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 $\hat{C}_1, \hat{C}_2 \subset \{1, \dots, n\}$ be two clusters estimated by $C(\cdot)$ on $\mathbf{X}$, that is, $\hat{C}_1, \hat{C}_2 \in \mathcal{C}(\mathbf{X})$.
- Consider the null hypothesis $H_0^{\{\hat{C}_1, \hat{C}_2\}} : \bar{\mu}_{\hat{C}_1} = \bar{\mu}_{\hat{C}_2}$.

### Goal

Define a $p$-value for $H_0^{\{\hat{C}_1, \hat{C}_2\}}$ that controls the selective type I error, that is,

$$\mathbb{P}_{H_0^{\{\hat{C}_1, \hat{C}_2\}}} \left( \text{reject } H_0^{\{\hat{C}_1, \hat{C}_2\}} \text{ based on } \mathbf{X} \text{ at level } \alpha \;\middle|\; \hat{C}_1, \hat{C}_2 \in \mathcal{C}(\mathbf{X}) \right) \leq \alpha \quad \forall \alpha \in [0, 1].$$

# Post-clustering inference

## Conditioning to define the $p$-value

Ideal $p$-value :

$$p_{\text{ideal}} = \mathbb{P}_{H_0^{\{\hat{C}_1, \hat{C}_2\}}} \left( \text{Critical region} \,\middle|\, \hat{C}_1, \hat{C}_2 \in \mathcal{C}(\mathbf{X}) \right)$$

# Post-clustering inference

## Conditioning to define the $p$-value

Ideal $p$-value :

$$p_{\text{ideal}} = \mathbb{P}_{H_0^{\{\hat{C}_1, \hat{C}_2\}}} \left( \text{Critical region} \,\middle|\, \hat{C}_1, \hat{C}_2 \in \mathcal{C}(\mathbf{X}) \right)$$

Additional technical conditions must be added to the conditioning set to ensure the analytical tractability of $p$-values :

$$p_{\text{tractable}} = \mathbb{P}_{H_0^{\{\hat{C}_1, \hat{C}_2\}}} \left( \text{Critical region} \,\middle|\, \hat{C}_1, \hat{C}_2 \in \mathcal{C}(\mathbf{X}) \cap E(\mathbf{X}) \right)$$

# Post-clustering inference

## Conditioning to define the $p$-value

Ideal $p$-value :

$$p_{\text{ideal}} = \mathbb{P}_{H_0^{\{\hat{C}_1, \hat{C}_2\}}} \left( \text{Critical region} \,\middle|\, \hat{C}_1, \hat{C}_2 \in \mathcal{C}(\mathbf{X}) \right)$$

Additional technical conditions must be added to the conditioning set to ensure the analytical tractability of $p$-values :

$$p_{\text{tractable}} = \mathbb{P}_{H_0^{\{\hat{C}_1, \hat{C}_2\}}} \left( \text{Critical region} \,\middle|\, \hat{C}_1, \hat{C}_2 \in \mathcal{C}(\mathbf{X}) \cap E(\mathbf{X}) \right)$$

Conditioning on too much information entails a loss of power

The "more strict" is $E(\mathbf{X})$, the less powerful the test is [1].

1. Jewell *et al.* 2022, Chen *et al.* 2022, Liu *et al.* 2018, Fithian *et al.* 2017.

# Independence setting

Gao *et al.* 2022

Framework

Consider the model

$$\mathbf{X} \sim \mathcal{MN}_{n\times p}(\boldsymbol{\mu}, \mathbf{I}_n, \sigma^2\mathbf{I}_p),$$

and the null hypothesis

$$H_0^{\{\hat{C}_1,\hat{C}_2\}} : \bar{\mu}_{\hat{C}_1} = \bar{\mu}_{\hat{C}_2}, \qquad \text{(null)}$$

for $\hat{C}_1, \hat{C}_2 \in \mathcal{C}(\mathbf{X})$.

# Independence setting

Gao *et al.* 2022

## Framework

Consider the model

$$\mathbf{X} \sim \mathcal{MN}_{n\times p}(\boldsymbol{\mu}, \mathbf{I}_n, \sigma^2\mathbf{I}_p),$$

and the null hypothesis

$$H_0^{\{\hat{C}_1,\hat{C}_2\}} : \bar{\mu}_{\hat{C}_1} = \bar{\mu}_{\hat{C}_2}, \qquad \text{(null)}$$

for $\hat{C}_1, \hat{C}_2 \in \mathcal{C}(\mathbf{X})$.

Gao *et al.* define a $p$-value for (null) that

- Controls the selective type I error,
- Can be be efficiently computed for hierarchical clustering (HAC) with several types of linkages and $k$-means (Chen and Witten 2023),

# Independence setting

## $p$-value definition

$p$-value for $H_0^{\{\hat{C}_1, \hat{C}_2\}}$ when $\mathbf{U} = \mathbf{I}_n$, $\mathbf{\Sigma} = \sigma^2 \mathbf{I}_p$ (Gao *et al.* 2022)

$$p(\mathbf{x}; \{\hat{C}_1, \hat{C}_2\}) = \mathbb{P}_{H_0^{\{\hat{C}_1, \hat{C}_2\}}} \Big( \|\bar{X}_{\hat{C}_1} - \bar{X}_{\hat{C}_2}\|_2 \geq \|\bar{x}_{\hat{C}_1} - \bar{x}_{\hat{C}_2}\|_2 \;\Big|\; \hat{C}_1, \hat{C}_2 \in \mathcal{C}(\mathbf{X}),$$

$$\boldsymbol{\pi}^{\perp}_{\nu(\hat{C}_1, \hat{C}_2)} \mathbf{X} = \boldsymbol{\pi}^{\perp}_{\nu(\hat{C}_1, \hat{C}_2)} \mathbf{x}, \, \mathrm{dir}(\bar{X}_{\hat{C}_1} - \bar{X}_{\hat{C}_2}) = \mathrm{dir}(\bar{x}_{\hat{C}_1} - \bar{x}_{\hat{C}_2}) \Big).$$

# Independence setting

## $p$-value definition

$p$-value for $H_0^{\{\hat{C}_1, \hat{C}_2\}}$ when $\mathbf{U} = \mathbf{I}_n$, $\mathbf{\Sigma} = \sigma^2 \mathbf{I}_p$ (Gao *et al.* 2022)

$$p(\mathbf{x}; \{\hat{C}_1, \hat{C}_2\}) = \mathbb{P}_{H_0^{\{\hat{C}_1, \hat{C}_2\}}} \Big( \|\bar{X}_{\hat{C}_1} - \bar{X}_{\hat{C}_2}\|_2 \geq \|\bar{x}_{\hat{C}_1} - \bar{x}_{\hat{C}_2}\|_2 \;\Big|\; \hat{C}_1, \hat{C}_2 \in \mathcal{C}(\mathbf{X}),$$

$$\boldsymbol{\pi}^{\perp}_{\nu(\hat{C}_1, \hat{C}_2)} \mathbf{X} = \boldsymbol{\pi}^{\perp}_{\nu(\hat{C}_1, \hat{C}_2)} \mathbf{x}, \, \mathrm{dir}(\bar{X}_{\hat{C}_1} - \bar{X}_{\hat{C}_2}) = \mathrm{dir}(\bar{x}_{\hat{C}_1} - \bar{x}_{\hat{C}_2}) \Big).$$

The $p$-value is computationally tractable (Gao *et al.* 2022)

$$p(\mathbf{x}; \{\hat{C}_1, \hat{C}_2\}) = 1 - \mathbb{F}_p \left( \|\bar{x}_{\hat{C}_1} - \bar{x}_{\hat{C}_2}\|_2; \sigma \sqrt{\frac{1}{|\hat{C}_1|} + \frac{1}{|\hat{C}_2|}}, \mathcal{S}_2(\mathbf{x}; \{\hat{C}_1, \hat{C}_2\}) \right)$$

where $\mathbb{F}_p(t; c, \mathcal{S})$ denotes the CDF of a $c\chi_p$ random variable truncated to the set $\mathcal{S}$.

# Independence setting

## $p$-value definition

$p$-value for $H_0^{\{\hat{C}_1, \hat{C}_2\}}$ when $\mathbf{U} = \mathbf{I}_n$, $\mathbf{\Sigma} = \sigma^2 \mathbf{I}_p$ (Gao *et al.* 2022)

$$p(\mathbf{x}; \{\hat{C}_1, \hat{C}_2\}) = \mathbb{P}_{H_0^{\{\hat{C}_1, \hat{C}_2\}}} \Big( \|\bar{X}_{\hat{C}_1} - \bar{X}_{\hat{C}_2}\|_2 \geq \|\bar{x}_{\hat{C}_1} - \bar{x}_{\hat{C}_2}\|_2 \;\Big|\; \hat{C}_1, \hat{C}_2 \in \mathcal{C}(\mathbf{X}),$$

$$\boldsymbol{\pi}^{\perp}_{\nu(\hat{C}_1, \hat{C}_2)} \mathbf{X} = \boldsymbol{\pi}^{\perp}_{\nu(\hat{C}_1, \hat{C}_2)} \mathbf{x}, \, \mathrm{dir}(\bar{X}_{\hat{C}_1} - \bar{X}_{\hat{C}_2}) = \mathrm{dir}(\bar{x}_{\hat{C}_1} - \bar{x}_{\hat{C}_2}) \Big).$$

The $p$-value is computationally tractable (Gao *et al.* 2022)

$$p(\mathbf{x}; \{\hat{C}_1, \hat{C}_2\}) = 1 - \mathbb{F}_p \Big( \|\bar{x}_{\hat{C}_1} - \bar{x}_{\hat{C}_2}\|_2; \sigma \sqrt{\frac{1}{|\hat{C}_1|} + \frac{1}{|\hat{C}_2|}}, \underbrace{\boxed{\mathcal{S}_2(\mathbf{x}; \{\hat{C}_1, \hat{C}_2\})}}_{\text{HAC, } k\text{-means}} \Big)$$

where $\mathbb{F}_p(t; c, \mathcal{S})$ denotes the CDF of a $c\chi_p$ random variable truncated to the set $\mathcal{S}$.

# Arbitrary dependence setting

## Adapt Gao *et al.* 2022 to realistic practical scenarios

### Framework

Consider the model

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

where $\mathbf{U} \in \mathcal{M}_{n\times n}(\mathbb{R})$ and $\boldsymbol{\Sigma} \in \mathcal{M}_{p\times p}(\mathbb{R})$. We ask $\mathbf{U}$ and $\boldsymbol{\Sigma}$ to be positive definite. Let

$$H_0^{\{\hat{C}_1, \hat{C}_2\}} : \bar{\mu}_{\hat{C}_1} = \bar{\mu}_{\hat{C}_2}, \tag{null}$$

for $\hat{C}_1, \hat{C}_2 \in \mathcal{C}(\mathbf{X})$.

# Arbitrary dependence setting

## Adapt Gao *et al.* 2022 to realistic practical scenarios

### Framework

Consider the model

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

where $\mathbf{U} \in \mathcal{M}_{n\times n}(\mathbb{R})$ and $\boldsymbol{\Sigma} \in \mathcal{M}_{p\times p}(\mathbb{R})$. We ask $\mathbf{U}$ and $\boldsymbol{\Sigma}$ to be positive definite. Let

$$H_0^{\{\hat{C}_1, \hat{C}_2\}} : \bar{\mu}_{\hat{C}_1} = \bar{\mu}_{\hat{C}_2}, \tag{null}$$

for $\hat{C}_1, \hat{C}_2 \in \mathcal{C}(\mathbf{X})$.

- Definition of a $p$-value for (null) that controls selective type I error under (dep),
- Efficient computation for HAC and $k$-means clustering.

# Choice of the test statistic

- Let $\mathcal{G}_1, \mathcal{G}_2 \subset \{1, \ldots, n\}$ with $\mathcal{G}_1 \cap \mathcal{G}_2 = \emptyset$
- Let

$$\mathbf{D}_{\mathcal{G}_1,\mathcal{G}_2} = \left( \underbrace{\frac{1}{|\mathcal{G}_1|}\mathbf{I}_p \quad \cdots \quad \frac{1}{|\mathcal{G}_1|}\mathbf{I}_p}_{|\mathcal{G}_1|} \quad \underbrace{-\frac{1}{|\mathcal{G}_2|}\mathbf{I}_p \quad \cdots \quad -\frac{1}{|\mathcal{G}_2|}\mathbf{I}_p}_{|\mathcal{G}_2|} \right).$$

# Choice of the test statistic

- Let $\mathcal{G}_1, \mathcal{G}_2 \subset \{1, \ldots, n\}$ with $\mathcal{G}_1 \cap \mathcal{G}_2 = \emptyset$
- Let

$$\mathbf{D}_{\mathcal{G}_1,\mathcal{G}_2} = \Big( \underbrace{\tfrac{1}{|\mathcal{G}_1|}\mathbf{I}_p \quad \cdots \quad \tfrac{1}{|\mathcal{G}_1|}\mathbf{I}_p}_{|\mathcal{G}_1|} \quad \underbrace{-\tfrac{1}{|\mathcal{G}_2|}\mathbf{I}_p \quad \cdots \quad -\tfrac{1}{|\mathcal{G}_2|}\mathbf{I}_p}_{|\mathcal{G}_2|} \Big).$$

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

$$\bar{X}_{\mathcal{G}_1} - \bar{X}_{\mathcal{G}_2} \overset{H_0^{\{\mathcal{G}_1,\mathcal{G}_2\}}}{\sim} \mathcal{N}_p\left(0, \mathbf{V}_{\mathcal{G}_1,\mathcal{G}_2}\right),$$

where

$$\mathbf{V}_{\mathcal{G}_1,\mathcal{G}_2} = \mathbf{D}_{\mathcal{G}_1,\mathcal{G}_2}(\mathbf{U}_{\mathcal{G}_1,\mathcal{G}_2} \otimes \boldsymbol{\Sigma})\mathbf{D}^T_{\mathcal{G}_1,\mathcal{G}_2}.$$

# Choice of the test statistic

- Let $\mathcal{G}_1, \mathcal{G}_2 \subset \{1, \ldots, n\}$ with $\mathcal{G}_1 \cap \mathcal{G}_2 = \emptyset$
- Let

$$\mathbf{D}_{\mathcal{G}_1,\mathcal{G}_2} = \left( \underbrace{\tfrac{1}{|\mathcal{G}_1|}\mathbf{I}_p \quad \cdots \quad \tfrac{1}{|\mathcal{G}_1|}\mathbf{I}_p}_{|\mathcal{G}_1|} \quad \underbrace{-\tfrac{1}{|\mathcal{G}_2|}\mathbf{I}_p \quad \cdots \quad -\tfrac{1}{|\mathcal{G}_2|}\mathbf{I}_p}_{|\mathcal{G}_2|} \right).$$

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

$$\bar{X}_{\mathcal{G}_1} - \bar{X}_{\mathcal{G}_2} \overset{H_0^{\{\mathcal{G}_1,\mathcal{G}_2\}}}{\sim} \mathcal{N}_p\left(0, \mathbf{V}_{\mathcal{G}_1,\mathcal{G}_2}\right),$$

where

$$\mathbf{V}_{\mathcal{G}_1,\mathcal{G}_2} = \mathbf{D}_{\mathcal{G}_1,\mathcal{G}_2}(\mathbf{U}_{\mathcal{G}_1,\mathcal{G}_2} \otimes \boldsymbol{\Sigma})\mathbf{D}_{\mathcal{G}_1,\mathcal{G}_2}^T.$$

Consequently,

$$\|\bar{X}_{\mathcal{G}_1} - \bar{X}_{\mathcal{G}_2}\|^2_{\mathbf{V}_{\mathcal{G}_1,\mathcal{G}_2}} \overset{H_0^{\{\mathcal{G}_1,\mathcal{G}_2\}}}{\sim} \chi^2_p.$$

with $\|x\|_{\mathbf{V}_{\mathcal{G}_1,\mathcal{G}_2}} = \sqrt{x^T \mathbf{V}^{-1}_{\mathbf{V}_{\mathcal{G}_1,\mathcal{G}_2}} x}, \quad \forall\, x \in \mathbb{R}^p.$

# Arbitrary dependence setting

## $p$-value definition

Key idea : Replace the norm $\|\cdot\|_2$ by the *Mahalanobis distance* between the cluster means w.r.t. the null distribution of their difference.

$p$-value for $H_0^{\{\hat{C}_1,\hat{C}_2\}}$ for arbitrary $\mathbf{U}$ and $\mathbf{\Sigma}$

$$p\mathbf{v}_{\hat{C}_1,\hat{C}_2}(\mathbf{x};\{\hat{C}_1,\hat{C}_2\}) = \mathbb{P}_{H_0^{\{\hat{C}_1,\hat{C}_2\}}}\Big(\|\bar{X}_{\hat{C}_1} - \bar{X}_{\hat{C}_2}\|\mathbf{v}_{\hat{C}_1,\hat{C}_2} \geq \|\bar{x}_{\hat{C}_1} - \bar{x}_{\hat{C}_2}\|\mathbf{v}_{\hat{C}_1,\hat{C}_2} \;\Big|\; \hat{C}_1,\hat{C}_2 \in \mathcal{C}(\mathbf{X}),$$

$$\boldsymbol{\pi}^{\perp}_{\nu(\hat{C}_1,\hat{C}_2)}\mathbf{X} = \boldsymbol{\pi}^{\perp}_{\nu(\hat{C}_1,\hat{C}_2)}\mathbf{x}\,,\ \mathrm{dir}\mathbf{v}_{\hat{C}_1,\hat{C}_2}(\bar{X}_{\hat{C}_1} - \bar{X}_{\hat{C}_2}) = \mathrm{dir}\mathbf{v}_{\hat{C}_1,\hat{C}_2}(\bar{x}_{\hat{C}_1} - \bar{x}_{\hat{C}_2})\Big).$$

# Arbitrary dependence setting

## $p$-value definition

Key idea : Replace the norm $\|\cdot\|_2$ by the *Mahalanobis distance* between the cluster means w.r.t. the null distribution of their difference.

$p$-value for $H_0^{\{\hat{C}_1,\hat{C}_2\}}$ for arbitrary $\mathbf{U}$ and $\mathbf{\Sigma}$

$$p_{\mathbf{v}_{\hat{C}_1,\hat{C}_2}}(\mathbf{x};\{\hat{C}_1,\hat{C}_2\}) = \mathbb{P}_{H_0^{\{\hat{C}_1,\hat{C}_2\}}}\Big(\|\bar{X}_{\hat{C}_1} - \bar{X}_{\hat{C}_2}\|_{\mathbf{v}_{\hat{C}_1,\hat{C}_2}} \geq \|\bar{x}_{\hat{C}_1} - \bar{x}_{\hat{C}_2}\|_{\mathbf{v}_{\hat{C}_1,\hat{C}_2}} \;\Big|\; \hat{C}_1,\hat{C}_2 \in \mathcal{C}(\mathbf{X}),$$

$$\boldsymbol{\pi}^{\perp}_{\nu(\hat{C}_1,\hat{C}_2)}\mathbf{X} = \boldsymbol{\pi}^{\perp}_{\nu(\hat{C}_1,\hat{C}_2)}\mathbf{x}\,,\ \mathrm{dir}_{\mathbf{v}_{\hat{C}_1,\hat{C}_2}}(\bar{X}_{\hat{C}_1} - \bar{X}_{\hat{C}_2}) = \mathrm{dir}_{\mathbf{v}_{\hat{C}_1,\hat{C}_2}}(\bar{x}_{\hat{C}_1} - \bar{x}_{\hat{C}_2})\Big).$$

Theorem : The $p$-value is computationally tractable (and controls sel. type I error)

$$p_{\mathbf{v}_{\hat{C}_1,\hat{C}_2}}(\mathbf{x};\{\hat{C}_1,\hat{C}_2\}) = 1 - \mathbb{F}_p\Big(\|\bar{x}_{\hat{C}_1} - \bar{x}_{\hat{C}_2}\|_{\mathbf{v}_{\hat{C}_1,\hat{C}_2}}; \mathcal{S}_{\mathbf{v}_{\hat{C}_1,\hat{C}_2}}(\mathbf{x},\{\hat{C}_1,\hat{C}_2\})\Big)$$

where $\mathbb{F}_p(t;\mathcal{S})$ denotes the CDF of a $\chi_p$ random variable truncated to the set $\mathcal{S}$.

# Arbitrary dependence setting

## $p$-value definition

Key idea : Replace the norm $\|\cdot\|_2$ by the *Mahalanobis distance* between the cluster means w.r.t. the null distribution of their difference.

$p$-value for $H_0^{\{\hat{C}_1,\hat{C}_2\}}$ for arbitrary $\mathbf{U}$ and $\mathbf{\Sigma}$

$$p_{\mathbf{v}_{\hat{C}_1,\hat{C}_2}}(\mathbf{x};\{\hat{C}_1,\hat{C}_2\}) = \mathbb{P}_{H_0^{\{\hat{C}_1,\hat{C}_2\}}}\Big(\|\bar{X}_{\hat{C}_1} - \bar{X}_{\hat{C}_2}\|_{\mathbf{v}_{\hat{C}_1,\hat{C}_2}} \geq \|\bar{x}_{\hat{C}_1} - \bar{x}_{\hat{C}_2}\|_{\mathbf{v}_{\hat{C}_1,\hat{C}_2}} \;\Big|\; \hat{C}_1,\hat{C}_2 \in \mathcal{C}(\mathbf{X}),$$

$$\boldsymbol{\pi}^{\perp}_{\nu(\hat{C}_1,\hat{C}_2)}\mathbf{X} = \boldsymbol{\pi}^{\perp}_{\nu(\hat{C}_1,\hat{C}_2)}\mathbf{x}\,,\ \mathrm{dir}_{\mathbf{v}_{\hat{C}_1,\hat{C}_2}}(\bar{X}_{\hat{C}_1} - \bar{X}_{\hat{C}_2}) = \mathrm{dir}_{\mathbf{v}_{\hat{C}_1,\hat{C}_2}}(\bar{x}_{\hat{C}_1} - \bar{x}_{\hat{C}_2})\Big).$$

Theorem : The $p$-value is computationally tractable (and controls sel. type I error)

$$p_{\mathbf{v}_{\hat{C}_1,\hat{C}_2}}(\mathbf{x};\{\hat{C}_1,\hat{C}_2\}) = 1 - \mathbb{F}_p\Big(\|\bar{x}_{\hat{C}_1} - \bar{x}_{\hat{C}_2}\|_{\mathbf{v}_{\hat{C}_1,\hat{C}_2}};\ \underbrace{\boxed{\mathcal{S}_{\mathbf{v}_{\hat{C}_1,\hat{C}_2}}(\mathbf{x},\{\hat{C}_1,\hat{C}_2\})}}_{\text{Scale trans. of } \mathcal{S}_2}\Big)$$

where $\mathbb{F}_p(t;\mathcal{S})$ denotes the CDF of a $\chi_p$ random variable truncated to the set $\mathcal{S}$.

# Numerical simulations

## Three dependence settings

(*a*) $\mathbf{U} = \mathbf{I}_n$ and $\mathbf{\Sigma}$ is the covariance matrix of an AR(1) model, i.e. $\Sigma_{ij} = \sigma^2 \rho^{|i-j|}$, with $\sigma = 1$ and $\rho = 0.5$.

(*b*) $\mathbf{U}$ is a compound symmetry covariance matrix, i.e. $\mathbf{U} = b + (a - b)\mathbf{I}_n$, with $a = 0.5$ and $b = 1$. $\mathbf{\Sigma}$ is a Toeplitz matrix, i.e. $\Sigma_{ij} = t(|i - j|)$, with $t(s) = 1 + 1/(1 + s)$ for $s \in \mathbb{N}$.

(*c*) $\mathbf{U}$ is the covariance matrix of an AR(1) model with $\sigma = 1$ and $\rho = 0.1$. $\mathbf{\Sigma}$ is a diagonal matrix with diagonal entries given by $\Sigma_{ii} = 1 + 1/i$.

# Numerical simulations

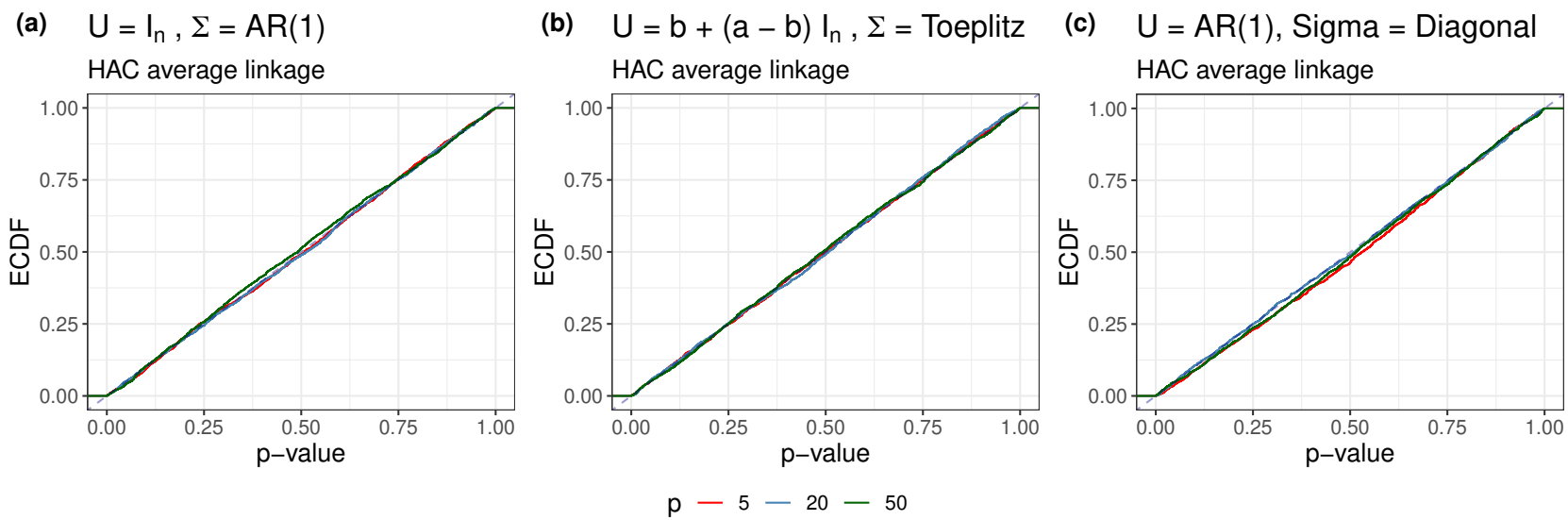
## Global null hypothesis

Let $n = 100$, $\boldsymbol{\mu} = \mathbf{O}_{n \times p}$, and set $\mathcal{C}$ to choose three clusters. Then, randomly select two groups and test for the difference of their means.

# Numerical simulations

## Global null hypothesis

Let $n = 100$, $\boldsymbol{\mu} = \mathbf{O}_{n \times p}$, and set $\mathcal{C}$ to choose three clusters. Then, randomly select two groups and test for the difference of their means.



# Numerical simulations

## Conditional power

Conditional power = probability of rejecting the null for a randomly selected pair of clusters when it holds.

Let $\boldsymbol{\mu}$ divide the $n = 50$ observations into three true clusters, for $\delta \in [4, 10.5]$ :

$$\mu_{ij} = \begin{cases} -\frac{\delta}{2} & \text{if } i \leq \lfloor \frac{n}{3} \rfloor, \\ \frac{\sqrt{3}\delta}{2} & \text{if } \lfloor \frac{n}{3} \rfloor < i \leq \lfloor \frac{2n}{3} \rfloor, \\ \frac{\delta}{2} & \text{otherwise.} \end{cases} \quad \forall\, i \in \{1, \ldots, n\},\ \forall\, j \in \{1, \ldots, p = 10\},$$

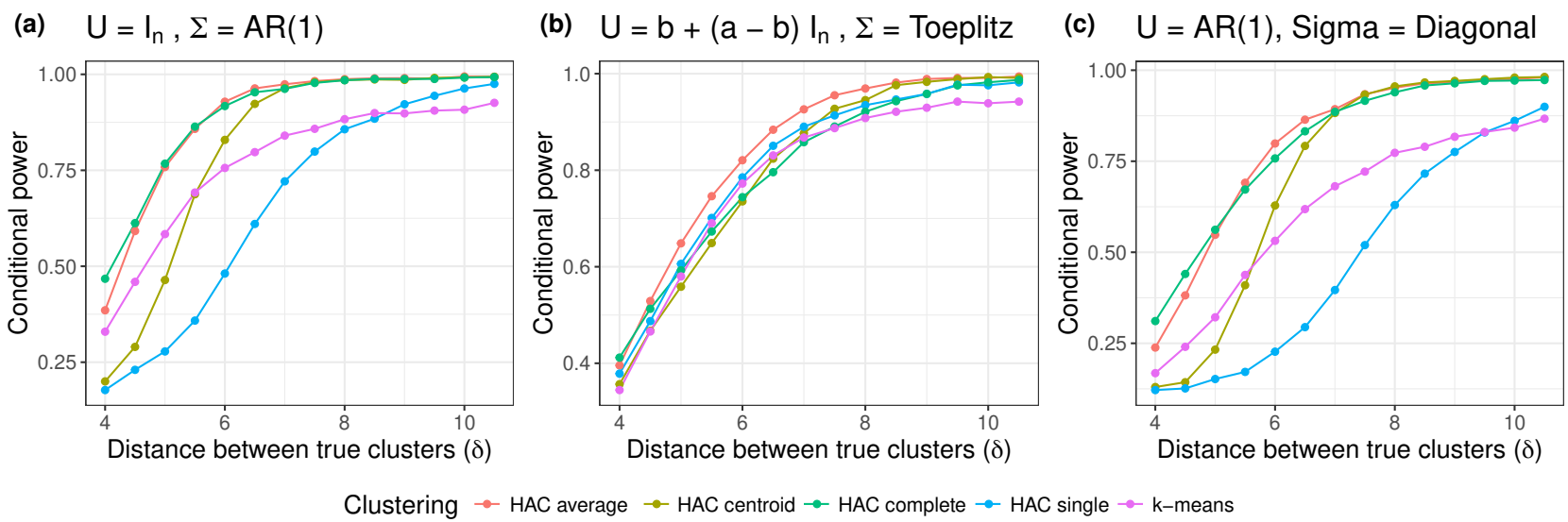
# Numerical simulations

## Conditional power

Conditional power = probability of rejecting the null for a randomly selected pair of clusters when it holds.

Let $\boldsymbol{\mu}$ divide the $n = 50$ observations into three true clusters, for $\delta \in [4, 10.5]$ :

$$\mu_{ij} = \begin{cases} -\frac{\delta}{2} & \text{if } i \leq \lfloor \frac{n}{3} \rfloor, \\ \frac{\sqrt{3}\delta}{2} & \text{if } \lfloor \frac{n}{3} \rfloor < i \leq \lfloor \frac{2n}{3} \rfloor, \\ \frac{\delta}{2} & \text{otherwise.} \end{cases} \quad \forall\, i \in \{1, \ldots, n\},\ \forall\, j \in \{1, \ldots, p = 10\},$$



# Estimation of unknown parameters

## Independence setting

Let $\mathbf{X}^{(n)} \sim \mathcal{MN}_{n\times p}(\boldsymbol{\mu}^{(n)}, \mathbf{I}_n, \sigma^2\mathbf{I}_p)$ and consider

$$\hat{p}(\mathbf{x}; \{\hat{C}_1, \hat{C}_2\}) = 1 - \mathbb{F}_p\left(\|\bar{x}_{\hat{C}_1} - \bar{x}_{\hat{C}_2}\|_2; \hat{\sigma}\sqrt{\frac{1}{|\hat{C}_1|} + \frac{1}{|\hat{C}_2|}}, \mathcal{S}_2(\mathbf{x}; \{\hat{C}_1, \hat{C}_2\})\right)$$

# Estimation of unknown parameters

## Independence setting

Let $\mathbf{X}^{(n)} \sim \mathcal{MN}_{n\times p}(\boldsymbol{\mu}^{(n)}, \mathbf{I}_n, \sigma^2\mathbf{I}_p)$ and consider

$$\hat{p}(\mathbf{x}; \{\hat{C}_1, \hat{C}_2\}) = 1 - \mathbb{F}_p\left(\|\bar{x}_{\hat{C}_1} - \bar{x}_{\hat{C}_2}\|_2; \hat{\sigma}\sqrt{\frac{1}{|\hat{C}_1|} + \frac{1}{|\hat{C}_2|}}, \mathcal{S}_2(\mathbf{x}; \{\hat{C}_1, \hat{C}_2\})\right)$$

Theorem 4 in Gao et al. 2022
If $\hat{\sigma}$ is an estimator of $\sigma$ such that

$$\lim_{n\to\infty} \mathbb{P}_{H_0^{\{\hat{C}_1^{(n)}, \hat{C}_2^{(n)}\}}}\left(\hat{\sigma}\left(\mathbf{X}^{(n)}\right) \geq \sigma \,\middle|\, \hat{C}_1^{(n)}, \hat{C}_2^{(n)} \in \mathcal{C}\left(\mathbf{X}^{(n)}\right)\right) = 1, \qquad (\sigma \text{ over-est})$$

then, for any $\alpha \in [0, 1]$, it holds

$$\limsup_{n\to\infty} \mathbb{P}_{H_0^{\{\hat{C}_1^{(n)}, \hat{C}_2^{(n)}\}}}\left(\hat{p}\left(\mathbf{X}^{(n)}; \left\{\hat{C}_1^{(n)}, \hat{C}_2^{(n)}\right\}\right) \leq \alpha \,\middle|\, \hat{C}_1^{(n)}, \hat{C}_2^{(n)} \in \mathcal{C}\left(\mathbf{X}^{(n)}\right)\right) \leq \alpha.$$

# Estimation of unknown parameters

## Independence setting

Let $\mathbf{X}^{(n)} \sim \mathcal{MN}_{n\times p}(\boldsymbol{\mu}^{(n)}, \mathbf{I}_n, \sigma^2\mathbf{I}_p)$ and consider

$$\hat{p}(\mathbf{x}; \{\hat{C}_1, \hat{C}_2\}) = 1 - \mathbb{F}_p\left(\|\bar{x}_{\hat{C}_1} - \bar{x}_{\hat{C}_2}\|_2; \hat{\sigma}\sqrt{\frac{1}{|\hat{C}_1|} + \frac{1}{|\hat{C}_2|}}, \mathcal{S}_2(\mathbf{x}; \{\hat{C}_1, \hat{C}_2\})\right)$$

Theorem 4 in Gao et al. 2022
If $\hat{\sigma}$ is an estimator of $\sigma$ such that

$$\lim_{n\to\infty} \mathbb{P}_{H_0^{\{\hat{C}_1^{(n)}, \hat{C}_2^{(n)}\}}}\left(\hat{\sigma}\left(\mathbf{X}^{(n)}\right) \geq \sigma \,\middle|\, \hat{C}_1^{(n)}, \hat{C}_2^{(n)} \in \mathcal{C}\left(\mathbf{X}^{(n)}\right)\right) = 1, \qquad (\sigma \text{ over-est})$$

then, for any $\alpha \in [0, 1]$, it holds

$$\limsup_{n\to\infty} \mathbb{P}_{H_0^{\{\hat{C}_1^{(n)}, \hat{C}_2^{(n)}\}}}\left(\hat{p}\left(\mathbf{X}^{(n)}; \left\{\hat{C}_1^{(n)}, \hat{C}_2^{(n)}\right\}\right) \leq \alpha \,\middle|\, \hat{C}_1^{(n)}, \hat{C}_2^{(n)} \in \mathcal{C}\left(\mathbf{X}^{(n)}\right)\right) \leq \alpha.$$

$\rightarrow$ Gao *et al.* propose an estimator $\hat{\sigma}$ that satisfies ($\sigma$ over-est) under mild assumptions on $\{\boldsymbol{\mu}^{(n)}\}_{n\in\mathbb{N}}$.

# Estimation of unknown parameters

## Arbitrary dependence setting

Let

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

Can we estimate both $\mathbf{U}$ and $\boldsymbol{\Sigma}$ ?

- Under the general model (dep), the scale matrices $\mathbf{U}$ and $\boldsymbol{\Sigma}$ are non-identifiable.
- We assume that one of the scale matrices is known, and assess the theoretical conditions that allow asymptotic control of the selective type I error when estimating the other one.
- Same reasoning for the estimation of $\mathbf{U}$ or $\boldsymbol{\Sigma}$ :

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

# Estimation of unknown parameters

## Arbitrary dependence setting

Let

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

Can we estimate both $\mathbf{U}$ and $\boldsymbol{\Sigma}$ ?

- Under the general model (dep), the scale matrices $\mathbf{U}$ and $\boldsymbol{\Sigma}$ are non-identifiable.
- We assume that one of the scale matrices is known, and assess the theoretical conditions that allow asymptotic control of the selective type I error when estimating the other one.
- Same reasoning for the estimation of $\mathbf{U}$ or $\boldsymbol{\Sigma}$ :

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

$\rightarrow$ How to extend the notion of over-estimation to matrices ?

# How to *over*-estimate a covariance matrix

We consider the natural extension of $\geq$ to the space of Hermitian matrices.

Loewner partial order $\succeq$

Let $A, B$ be two Hermitian matrices. $A \succeq B$ if and only if $A - B$ is positive semidefinite (PSD).

Remark : If $A = \hat{\sigma}\mathbf{I}_p$ and $B = \sigma\mathbf{I}_p$, the condition $A \succeq B$ becomes $\hat{\sigma} \geq \sigma$.

# How to *over*-estimate a covariance matrix

We consider the natural extension of $\geq$ to the space of Hermitian matrices.

### Loewner partial order $\succeq$

Let $A, B$ be two Hermitian matrices. $A \succeq B$ if and only if $A - B$ is positive semidefinite (PSD).

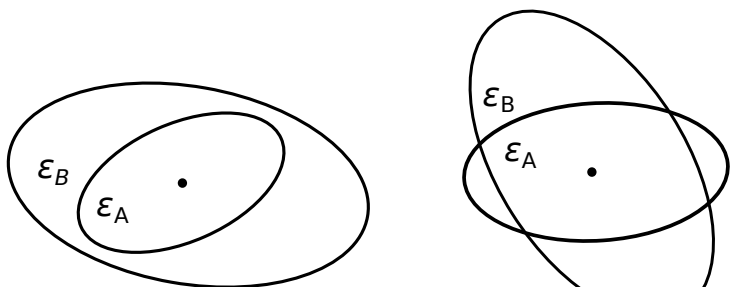
Remark : If $A = \hat{\sigma}\mathbf{I}_p$ and $B = \sigma\mathbf{I}_p$, the condition $A \succeq B$ becomes $\hat{\sigma} \geq \sigma$.

### Graphical interpretation

Every PSD matrix $A$ defines an ellipsoid $\mathcal{E}_A = \{x \in \mathbb{R}^d \ : \ x^T A x \leq 1\}$, where

- The eigenvectors of $A$ are the principal axes of $\mathcal{E}_A$,
- The eigenvalues of $A$ are the squared lengths of the principal axes of $\mathcal{E}_A$.

Then, it holds $\mathcal{E}_A \subset \mathcal{E}_B \Leftrightarrow A \preceq B$.




$A \preceq B$ $\quad$ $A \not\preceq B$

# Over-estimation of $\mathbf{\Sigma}$ for known $\mathbf{U}$

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

$$\hat{p}_{\hat{\mathbf{V}}_{\hat{C}_1,\hat{C}_2}}(\mathbf{x}; \{\hat{C}_1, \hat{C}_2\}) = 1 - \mathbb{F}_p\left( \|\bar{x}_{\hat{C}_1} - \bar{x}_{\hat{C}_2}\|_{\hat{\mathbf{V}}_{\hat{C}_1,\hat{C}_2}}; \mathcal{S}_{\hat{\mathbf{V}}_{\hat{C}_1,\hat{C}_2}}(\mathbf{x}, \{\hat{C}_1, \hat{C}_2\}) \right)$$

where $\hat{\mathbf{V}}_{\hat{C}_1,\hat{C}_2} = \mathbf{D}_{\hat{C}_1,\hat{C}_2}(\mathbf{U}_{\hat{C}_1,\hat{C}_2} \otimes \hat{\mathbf{\Sigma}}(\mathbf{x}))\mathbf{D}^T_{\hat{C}_1,\hat{C}_2}$.

# Over-estimation of $\mathbf{\Sigma}$ for known $\mathbf{U}$

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

$$\hat{p}_{\hat{\mathbf{V}}_{\hat{C}_1,\hat{C}_2}}(\mathbf{x}; \{\hat{C}_1, \hat{C}_2\}) = 1 - \mathbb{F}_p\left( \|\bar{x}_{\hat{C}_1} - \bar{x}_{\hat{C}_2}\|_{\hat{\mathbf{V}}_{\hat{C}_1,\hat{C}_2}}; \mathcal{S}_{\hat{\mathbf{V}}_{\hat{C}_1,\hat{C}_2}}(\mathbf{x}, \{\hat{C}_1, \hat{C}_2\}) \right)$$

where $\hat{\mathbf{V}}_{\hat{C}_1,\hat{C}_2} = \mathbf{D}_{\hat{C}_1,\hat{C}_2}(\mathbf{U}_{\hat{C}_1,\hat{C}_2} \otimes \hat{\mathbf{\Sigma}}(\mathbf{x}))\mathbf{D}^T_{\hat{C}_1,\hat{C}_2}$.

Theorem
If $\hat{\mathbf{\Sigma}}\left(\mathbf{X}^{(n)}\right)$ is a positive definite estimator of $\mathbf{\Sigma}$ such that

$$\lim_{n\to\infty} \mathbb{P}_{H_0^{\{\hat{C}_1^{(n)},\hat{C}_2^{(n)}\}}} \left( \hat{\mathbf{\Sigma}}\left(\mathbf{X}^{(n)}\right) \succeq \mathbf{\Sigma} \,\middle|\, \hat{C}_1^{(n)}, \hat{C}_2^{(n)} \in \mathcal{C}\left(\mathbf{X}^{(n)}\right) \right) = 1,$$

then, for any $\alpha \in [0,1]$, we have

$$\limsup_{n\to\infty} \mathbb{P}_{H_0^{\{\hat{C}_1^{(n)},\hat{C}_2^{(n)}\}}} \left( p_{\hat{\mathbf{V}}_{\hat{C}_1^{(n)},\hat{C}_2^{(n)}}}\left(\mathbf{X}^{(n)}; \left\{\hat{C}_1^{(n)}, \hat{C}_2^{(n)}\right\}\right) \le \alpha \,\middle|\, \hat{C}_1^{(n)}, \hat{C}_2^{(n)} \in \mathcal{C}\left(\mathbf{X}^{(n)}\right) \right) \le \alpha.$$

# Asymptotic over-estimator of $\mathbf{\Sigma}$

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

For a given estimator $\hat{\mathbf{\Sigma}}\left(\mathbf{X}^{(n)}\right)$ of $\mathbf{\Sigma}$, assessing whether $\hat{\mathbf{\Sigma}}\left(\mathbf{X}^{(n)}\right) \succeq \mathbf{\Sigma}$ asymptotically strongly depends on how the sequences $\{\boldsymbol{\mu}^{(n)}\}_{n\in\mathbb{N}}$ and $\{\mathbf{U}^{(n)}\}_{n\in\mathbb{N}}$ grow up to infinity.

# Asymptotic over-estimator of $\mathbf{\Sigma}$

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

For a given estimator $\hat{\mathbf{\Sigma}}\left(\mathbf{X}^{(n)}\right)$ of $\mathbf{\Sigma}$, assessing whether $\hat{\mathbf{\Sigma}}\left(\mathbf{X}^{(n)}\right) \succeq \mathbf{\Sigma}$ asymptotically strongly depends on how the sequences $\{\boldsymbol{\mu}^{(n)}\}_{n\in\mathbb{N}}$ and $\{\mathbf{U}^{(n)}\}_{n\in\mathbb{N}}$ grow up to infinity.

## Estimator candidate

$$\hat{\mathbf{\Sigma}} = \hat{\mathbf{\Sigma}}\left(\mathbf{X}\right) = \frac{1}{n-1}\left(\mathbf{X} - \bar{\mathbf{X}}\right)^T \mathbf{U}^{-1}\left(\mathbf{X} - \bar{\mathbf{X}}\right), \qquad \text{(estimator)}$$

where $\bar{\mathbf{X}}$ is a $n \times p$ matrix having as rows the mean across rows of $\mathbf{X}$.

# Asymptotic over-estimator of $\mathbf{\Sigma}$

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

For a given estimator $\hat{\mathbf{\Sigma}}\left(\mathbf{X}^{(n)}\right)$ of $\mathbf{\Sigma}$, assessing whether $\hat{\mathbf{\Sigma}}\left(\mathbf{X}^{(n)}\right) \succeq \mathbf{\Sigma}$ asymptotically strongly depends on how the sequences $\{\boldsymbol{\mu}^{(n)}\}_{n\in\mathbb{N}}$ and $\{\mathbf{U}^{(n)}\}_{n\in\mathbb{N}}$ grow up to infinity.

## Estimator candidate

$$\hat{\mathbf{\Sigma}} = \hat{\mathbf{\Sigma}}\left(\mathbf{X}\right) = \frac{1}{n-1}\left(\mathbf{X} - \bar{\mathbf{X}}\right)^T \mathbf{U}^{-1}\left(\mathbf{X} - \bar{\mathbf{X}}\right), \qquad \text{(estimator)}$$

where $\bar{\mathbf{X}}$ is a $n \times p$ matrix having as rows the mean across rows of $\mathbf{X}$.

$\rightarrow$ Assumptions on $\{\boldsymbol{\mu}^{(n)}\}_{n\in\mathbb{N}}$ and $\{\mathbf{U}^{(n)}\}_{n\in\mathbb{N}}$ to ensure that (estimator) a.s. asymptotically overestimates $\mathbf{\Sigma}$ ?

# Assumptions on $\boldsymbol{\mu}^{(n)}$

Assumptions 1 and 2 in Gao et al. 2022 *(Assumption 1)*

For all $n \in \mathbb{N}$, there are exactly $K^*$ distinct mean vectors among the first $n$ observations, i.e.

$$\left\{\mu_i^{(n)}\right\}_{i=1,\ldots,n} = \{\theta_1, \ldots, \theta_{K^*}\}.$$

Besides, the proportion of the first $n$ observations that have mean vector $\theta_k$ converges to $\pi_k > 0$, i.e.

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{\mu_i^{(n)} = \theta_k\} = \pi_k, \tag{as-1}$$

for all $k \in \{1, \ldots, K^*\}$, where $\sum_{k=1}^{K^*} \pi_k = 1$.

# Assumptions on $\boldsymbol{\mu}^{(n)}$

Assumptions 1 and 2 in Gao et al. 2022 *(Assumption 1)*

For all $n \in \mathbb{N}$, there are exactly $K^*$ distinct mean vectors among the first $n$ observations, i.e.

$$\left\{\mu_i^{(n)}\right\}_{i=1,\ldots,n} = \{\theta_1, \ldots, \theta_{K^*}\}.$$

Besides, the proportion of the first $n$ observations that have mean vector $\theta_k$ converges to $\pi_k > 0$, i.e.

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{\mu_i^{(n)} = \theta_k\} = \pi_k, \tag{as-1}$$

for all $k \in \{1, \ldots, K^*\}$, where $\sum_{k=1}^{K^*} \pi_k = 1$.

$\diamond$ If $\mathbf{U}^{(n)} = \mathbf{I}_n$, this is the only requirement to ensure asymp. over-estimation of $\boldsymbol{\Sigma}$.

# Assumptions on $\boldsymbol{\mu}^{(n)}$

Assumptions 1 and 2 in Gao et al. 2022 *(Assumption 1)*

For all $n \in \mathbb{N}$, there are exactly $K^*$ distinct mean vectors among the first $n$ observations, i.e.

$$\left\{\mu_i^{(n)}\right\}_{i=1,\dots,n} = \{\theta_1, \dots, \theta_{K^*}\}.$$

Besides, the proportion of the first $n$ observations that have mean vector $\theta_k$ converges to $\pi_k > 0$, i.e.

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{\mu_i^{(n)} = \theta_k\} = \pi_k, \qquad \text{(as-1)}$$

for all $k \in \{1, \dots, K^*\}$, where $\sum_{k=1}^{K^*} \pi_k = 1$.

◇ If $\mathbf{U}^{(n)} = \mathbf{I}_n$, this is the only requirement to ensure asymp. over-estimation of $\boldsymbol{\Sigma}$.
◇ For **general** $\mathbf{U}^{(n)}$, (as-1) turns into asking the quantities

$$\frac{1}{n} \sum_{l,s=1}^{n} \left(U^{(n)}\right)^{-1}_{ls} \mathbb{1}\{\mu_l^{(n)} = \theta_k\} \mathbb{1}\{\mu_s^{(n)} = \theta_{k'}\}$$

to **converge with explicit limit** as $n$ tends to infinity.

# One more assumption on $\boldsymbol{\mu}^{(n)}$ for non-diagonal $\mathbf{U}^{(n)}$

**Assumption on $\boldsymbol{\mu}^{(n)}$ for non-diagonal $\mathbf{U}^{(n)}$ *(Assumption 2)***

If $\mathbf{U}^{(n)}$ is non-diagonal for all $n \in \mathbb{N}$, for any $k, k' \in \{1, \ldots, K^*\}$, the proportion of the first $n$ observations at distance $r \geq 1$ in $\mathbf{X}^{(n)}$ having means $\theta_k$ and $\theta_{k'}$ converges, and its limit converges to $\pi_k \pi_{k'}$ when the lag $r$ tends to infinity. More precisely,

$$\lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n-r} \mathbb{1}\{\mu_i = \theta_k\} \mathbb{1}\{\mu_{i+r} = \theta_{k'}\} = \pi_{kk'}^r \underset{r \to \infty}{\longrightarrow} \pi_k \, \pi_{k'}. \qquad \text{(as-2)}$$

We are asking the proportion of pairs of observations having a given a pair of means to approach the product of individual proportions (as-1) when both observations are far away in $\mathbf{X}^{(n)}$.

# Assumptions on the sequence $\{\mathbf{U}^{(n)}\}_{n\in\mathbb{N}}$

**Assumption on $\{\mathbf{U}^{(n)}\}_{n\in\mathbb{N}}$ *(Assumption 3)***

Every superdiagonal of $\left(\mathbf{U}^{(n)}\right)^{-1}$ defines asymptotically a convergent sequence, whose limits sum up to a real value. More precisely, for any $i \in \mathbb{N}$ and any $r \geq 0$,

$$\lim_{n\to\infty} \left(U^{(n)}\right)^{-1}_{i\,i+r} = \Lambda_{i\,i+r}, \quad \text{where} \quad \lim_{i\to\infty} \Lambda_{i\,i+r} = \lambda_r \quad \text{and} \quad \sum_{r=0}^{\infty} \lambda_r = \lambda \in \mathbb{R}.$$

Moreover, for each $r \geq 0$, the sequence $\{\left(U^{(n)}\right)^{-1}_{i\,i+r}\}_{n\in\mathbb{N}}$ satisfies any of the following conditions :

- $(i)$ It is dominated by a summable sequence i.e. $\left|\left(U^{(n)}\right)^{-1}_{i\,i+r} - \Lambda_{i\,i+r}\right| \leq \alpha_i \ \forall\, n \in \mathbb{N}$, with $\{\alpha_i\}_{i=1}^{\infty} \in \ell_1$,
- $(ii)$ For each $i \in \mathbb{N}$, it is non-decreasing or non-increasing.

# Some admissible dependence models for $\{\mathbf{U}^{(n)}\}_{n\in\mathbb{N}}$

Remark 1 (Diagonal)

Let $\mathbf{U}^{(n)} = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$. If the sequence $\{\lambda_n\}_{n\in\mathbb{N}}$ is convergent, then the sequence $\{\mathbf{U}^{(n)}\}_{n\in\mathbb{N}}$ satisfies Assumption 3.

# Some admissible dependence models for $\{\mathbf{U}^{(n)}\}_{n\in\mathbb{N}}$

### Remark 1 (Diagonal)

Let $\mathbf{U}^{(n)} = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$. If the sequence $\{\lambda_n\}_{n\in\mathbb{N}}$ is convergent, then the sequence $\{\mathbf{U}^{(n)}\}_{n\in\mathbb{N}}$ satisfies Assumption 3.

### Remark 2 (Compound symmetry)

Let $a, b \in \mathbb{R}$ with $b \neq a \geq 0$. If $\mathbf{U}^{(n)} = b\mathbf{1}_{n\times n} + (a-b)\,\mathbf{I}_n$, where $\mathbf{1}_{n\times n}$ is a $n \times n$ matrix of ones, then $\{\mathbf{U}^{(n)}\}_{n\in\mathbb{N}}$ satisfies Assumption 3.

# Some admissible dependence models for $\{\mathbf{U}^{(n)}\}_{n\in\mathbb{N}}$

### Remark 1 (Diagonal)

Let $\mathbf{U}^{(n)} = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$. If the sequence $\{\lambda_n\}_{n\in\mathbb{N}}$ is convergent, then the sequence $\{\mathbf{U}^{(n)}\}_{n\in\mathbb{N}}$ satisfies Assumption 3.

### Remark 2 (Compound symmetry)

Let $a, b \in \mathbb{R}$ with $b \neq a \geq 0$. If $\mathbf{U}^{(n)} = b\mathbf{1}_{n\times n} + (a-b)\,\mathbf{I}_n$, where $\mathbf{1}_{n\times n}$ is a $n \times n$ matrix of ones, then $\{\mathbf{U}^{(n)}\}_{n\in\mathbb{N}}$ satisfies Assumption 3.

### Remark 3 (AR($P$))

Let $\mathbf{U}^{(n)}$ be the covariance matrix of an auto-regressive process of order $P \geq 1$ such that, if $P > 2$, $\beta_k \beta_{k'} \geq 0$ for all $k, k' \in \{1, \ldots, P\}$. Then, the sequence $\{\mathbf{U}^{(n)}\}_{n\in\mathbb{N}}$ satisfies Assumption 3.

# Estimation of $\mathbf{\Sigma}$ for known $\mathbf{U}$

## Final results

**Proposition**

Let $\mathbf{X}^{(n)} \sim \mathcal{MN}_{n\times p}(\boldsymbol{\mu}^{(n)}, \mathbf{U}^{(n)}, \mathbf{\Sigma})$, whose parameters $\boldsymbol{\mu}^{(n)}$, $\mathbf{U}^{(n)}$ satisfy Assumptions 1, 2 and 3 for some $K^* > 1$. Let $\hat{\mathbf{\Sigma}}$ be the estimator defined in (estimator). Then,

$$\lim_{n\to\infty} \mathbb{P}\left(\hat{\mathbf{\Sigma}}\left(\mathbf{X}^{(n)}\right) \succeq \mathbf{\Sigma}\right) = 1.$$

# Estimation of $\mathbf{\Sigma}$ for known $\mathbf{U}$

## Final results

**Proposition**

Let $\mathbf{X}^{(n)} \sim \mathcal{MN}_{n\times p}(\boldsymbol{\mu}^{(n)}, \mathbf{U}^{(n)}, \mathbf{\Sigma})$, whose parameters $\boldsymbol{\mu}^{(n)}$, $\mathbf{U}^{(n)}$ satisfy Assumptions 1, 2 and 3 for some $K^* > 1$. Let $\hat{\mathbf{\Sigma}}$ be the estimator defined in (estimator). Then,

$$\lim_{n\to\infty} \mathbb{P}\left(\hat{\mathbf{\Sigma}}\left(\mathbf{X}^{(n)}\right) \succeq \mathbf{\Sigma}\right) = 1.$$

**Proposition**

Let $\mathbf{X}^{(n)} \sim \mathcal{MN}_{n\times p}(\boldsymbol{\mu}^{(n)}, \mathbf{U}^{(n)}, \mathbf{\Sigma})$, whose parameters $\boldsymbol{\mu}^{(n)}$, $\mathbf{U}^{(n)}$ satisfy Assumptions 1, 2 and 3 for some $K^* > 1$. Let $\mathbf{x}^{(n)}$ be a realization of $\mathbf{X}^{(n)}$ and $\hat{C}_1^{(n)}$, $\hat{C}_2^{(n)}$ a pair of clusters estimated from $\mathbf{x}^{(n)}$. Let $\mathbf{Y}^{(n)}$ an independent and identically distributed copy of $\mathbf{X}^{(n)}$. Then,

$$\lim_{n\to\infty} \mathbb{P}_{H_0^{\{\hat{C}_1^{(n)}, \hat{C}_2^{(n)}\}}} \left(\hat{\mathbf{\Sigma}}\left(\mathbf{Y}^{(n)}\right) \succeq \mathbf{\Sigma} \,\middle|\, \hat{C}_1^{(n)}, \hat{C}_2^{(n)} \in \mathcal{C}\left(\mathbf{X}^{(n)}\right)\right) = 1.$$

# Numerical simulations

Let

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

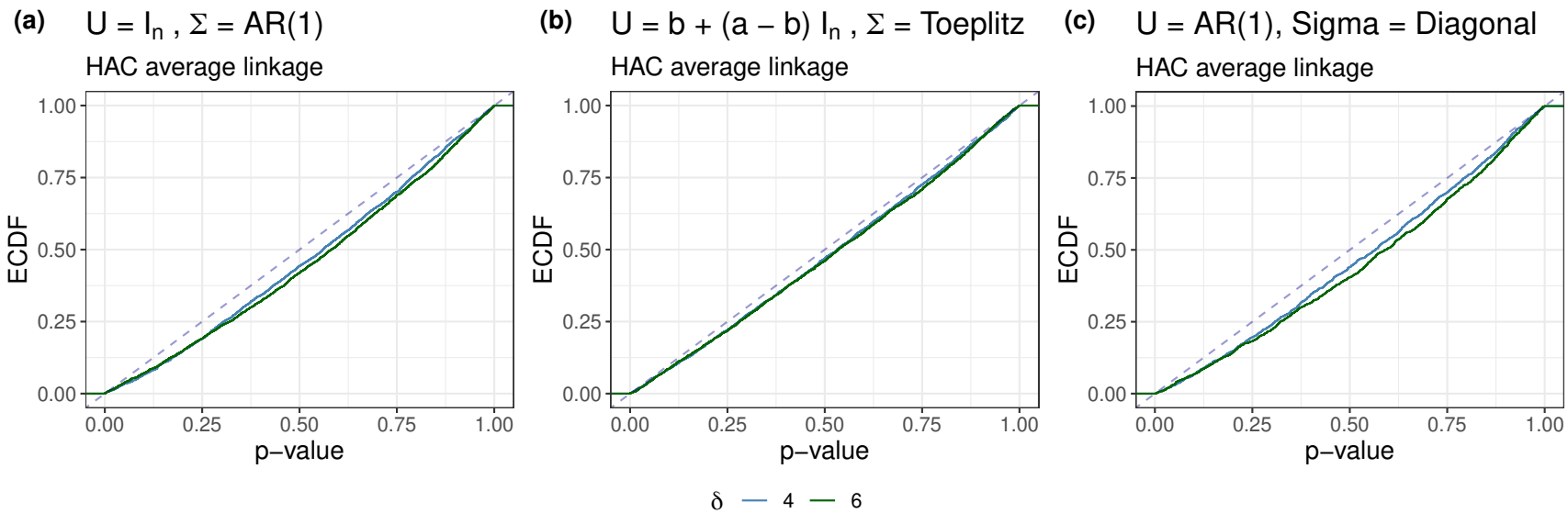
For $n = 500$ and $p = 10$, we simulated $K = 10000$ samples drawn from (dep) in settings $(a)$, $(b)$ and $(c)$ with $\boldsymbol{\mu}$ being divided into two clusters :

$$\mu_{ij} = \begin{cases} \frac{\delta}{j} & \text{if } i \leq \frac{n}{2}, \\ -\frac{\delta}{j} & \text{otherwise,} \end{cases} \qquad \forall\, i \in \{1, \ldots, n\},\ \forall\, j \in \{1, \ldots, p\},$$

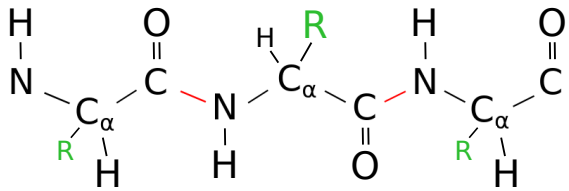
with $\delta \in \{4, 6\}$.

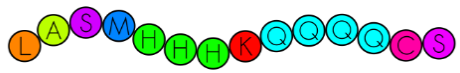
For HAC with average linkage we set $\mathcal{C}$ to chose three clusters. Then, we kept the samples for which (null) held when comparing two randomly selected clusters.

# Numerical simulations



# Proteins : sequence and conformation




Sequence

3D structure (conformation)

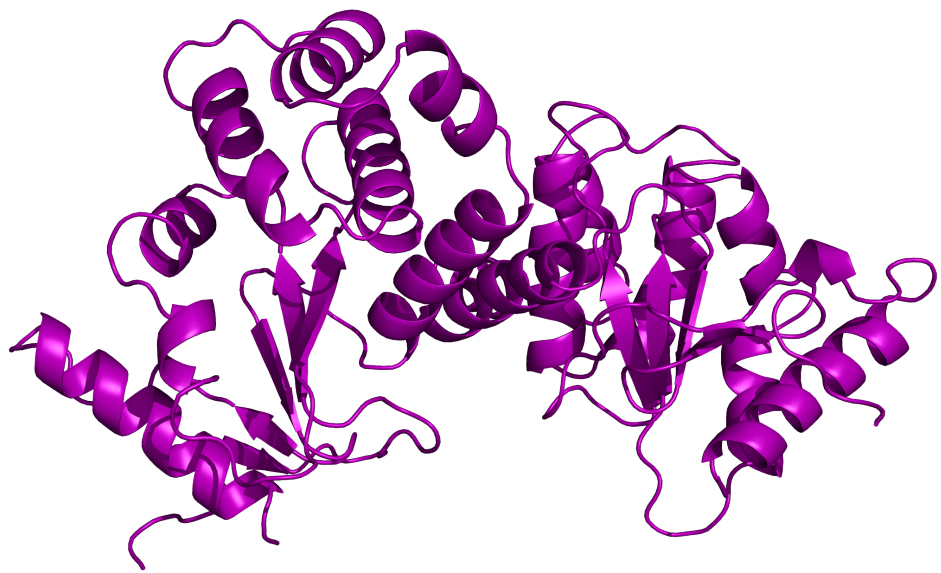
**Sequence $\overset{?}{\Leftrightarrow}$ 3D structure $\Leftrightarrow$ Function**

# Proteins : sequence and conformation

## The three dimensional structure may be stable...

Many proteins **fold** into their **native state**

Amino acid sequence $\Longleftrightarrow$ Well-defined 3D structure



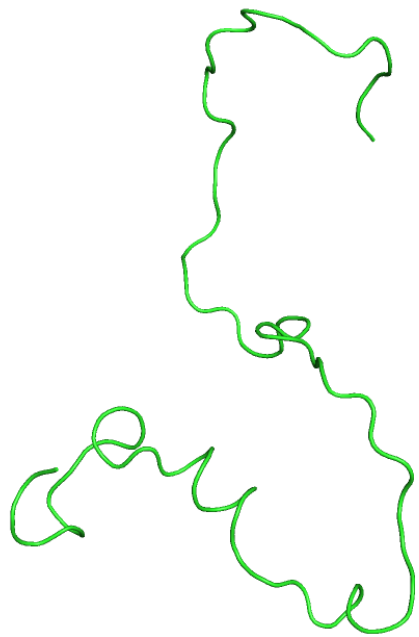
Structure : J.L., Guddat, L.W., Oxidized Dsba at 2.7 Anstromgs resolution, crystal form III (1998). https ://doi.org/10.2210/pdb1A2M/pdb.

# Proteins : sequence and conformation

## ...or not : Intrinsic Disorder in Proteins

### Intrinsically Disordered Proteins (IDP)

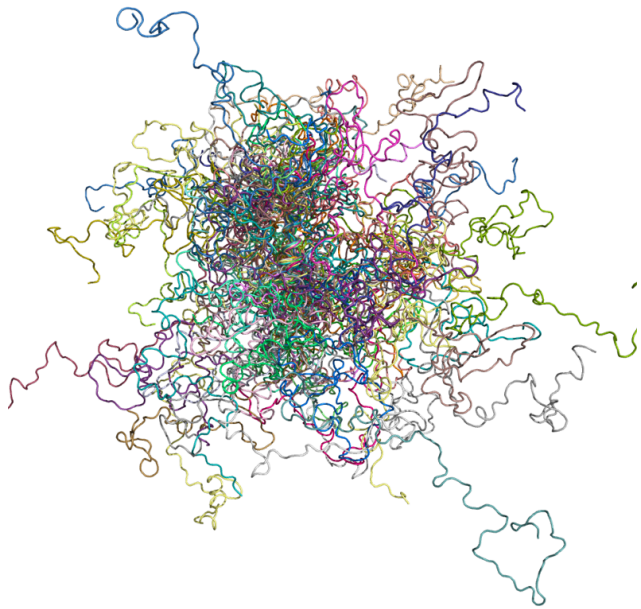
Lack of a native state : constant shape-changing and transitioning between different conformations



N-terminal region of Huntingtin (MD simulation). Elena-Real C.A. *et al.* Nat Struct Mol Biol. 2023.

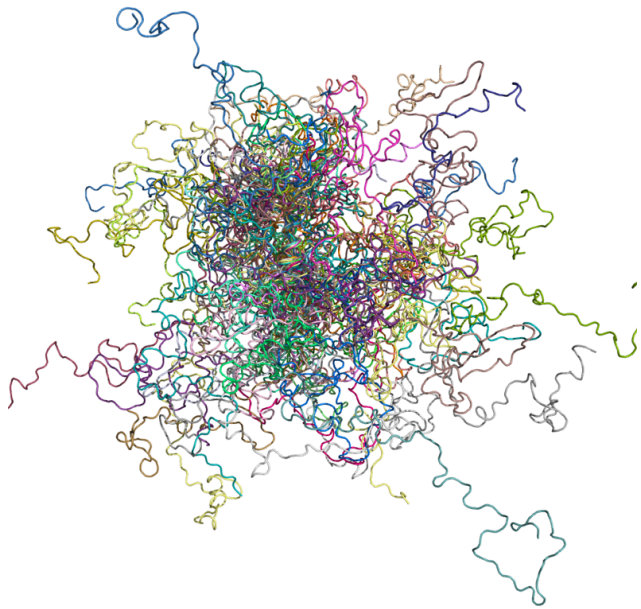
# Intrinsically Disordered Proteins

Amino acid sequence $\Longrightarrow$ **Conformational ensemble**



# Intrinsically Disordered Proteins

Amino acid sequence $\Longrightarrow$ **Conformational ensemble**



Classical approach to characterize conformational ensembles

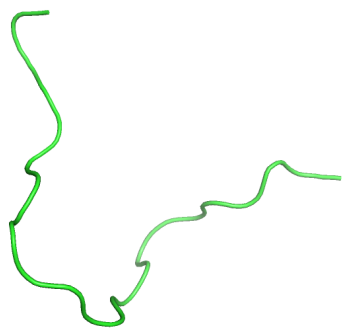
Clustering the set of states using pairwise Euclidean distances to feature conformations

Goal : post-clustering inference on protein data $\rightarrow$ need to admit arbitrary dependence structures $\mathbf{U}$ and $\mathbf{\Sigma}$.

# Hierarchical clustering of Hst5

Hst5 ensemble simulated with Flexible-Meccano (FM) [2] and filtered by SAXS data[3]

- $n = 2000$ conformations
- Featured by pairwise Euclidean distances of 24 amino acids $\Rightarrow p = 276$
- No temporal evolution in FM simulation : $\mathbf{U}^{(n)} = \mathbf{I}_n$
- $\mathbf{\Sigma}$ unknown to be estimated



## Strategy

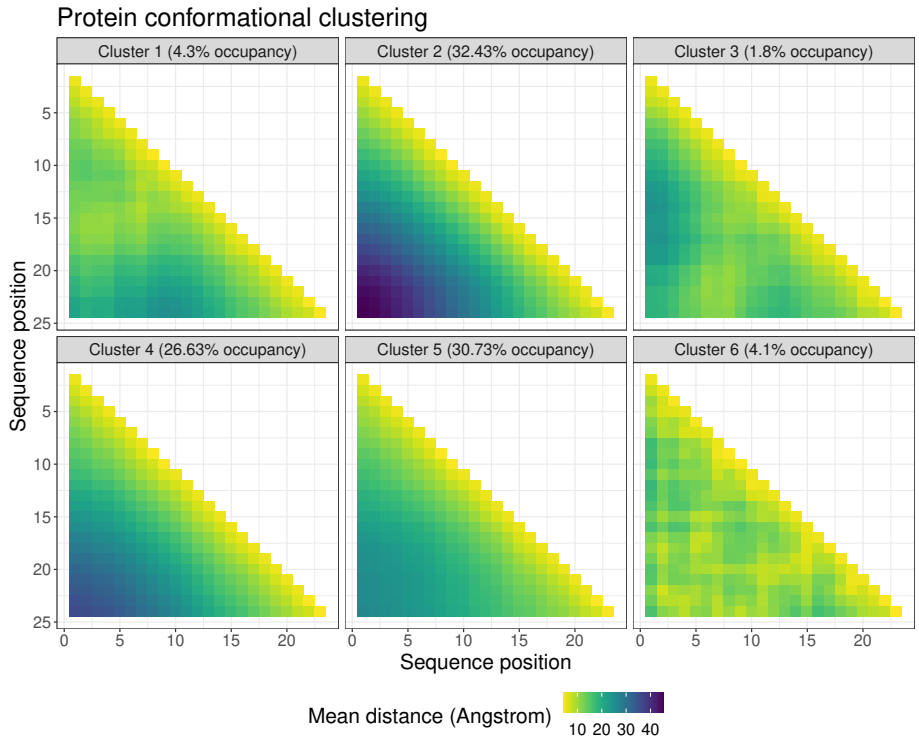
Hierarchical clustering with average linkage, find 6 clusters. Using R package `PCIdep` [4].

---

2. Ozenne *et al.* 2012, Bernadó *et al.* 2005. 3. Sagar *et al.* 2021.

4. https ://github.com/gonzalez-delgado/PCIdep

# Hierarchical clustering of Hst5




| **Cluster** | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| 2 | $2.187589 \cdot 10^{-4}$ | | | | |
| 3 | $3.039844 \cdot 10^{-11}$ | $1.41 \cdot 10^{-3}$ | | | |
| 4 | $1.070993 \cdot 10^{-10}$ | 0.300540 | $2.98464 \cdot 10^{-4}$ | | |
| 5 | $3.038979 \cdot 10^{-16}$ | 0.093018 | $6.015797 \cdot 10^{-5}$ | 0.105446 | |
| 6 | $1.729616 \cdot 10^{-6}$ | 0.010612 | $9.290826 \cdot 10^{-9}$ | $2.105 \cdot 10^{-3}$ | $5.624624 \cdot 10^{-5}$ |

**Thank you for your attention !**

R package `PCIdep` at https://github.com/gonzalez-delgado/PCIdep/

Preprint cooming soon...

https://www.math.univ-toulouse.fr/∼jgonzale/