Selective Inference for Hierarchical Clustering

Ganchao Wei

October 6, 2021

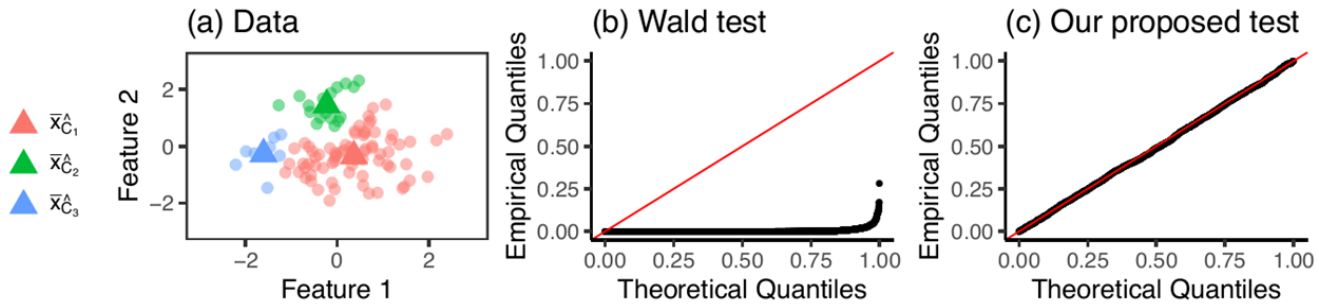
Overview

Introduction & Motivation

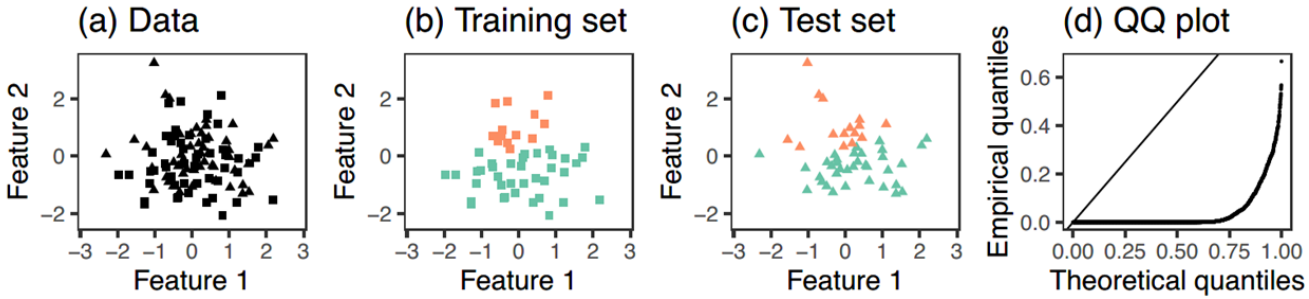
Goal: Test for a difference in means between groups.
But... The groups are usually defined via clustering $\Rightarrow$ inflate type I error if using the classical test.
Settings: $\boldsymbol{X} \sim MN_{n\times q}(\boldsymbol{\mu}, \boldsymbol{I}_n, \sigma^2\boldsymbol{I}_q)$, where $\boldsymbol{\mu}$ and $\sigma^2 > 0$ is known.
Define the population & empirical row mean of $\mathcal{G}$ as $\bar{\mu}_{\mathcal{G}} = \frac{1}{|\mathcal{G}|}\sum_{i\in\mathcal{G}}\mu_i$ and $\bar{X}_{\mathcal{G}} = \frac{1}{|\mathcal{G}|}\sum_{i\in\mathcal{G}}X_i$
Test: $H_0^{\{\hat{C}_1,\hat{C}_2\}} : \bar{\mu}_{\hat{C}_1} = \bar{\mu}_{\hat{C}_1}$ vs. $H_1^{\{\hat{C}_1,\hat{C}_2\}} : \bar{\mu}_{\hat{C}_1} \neq \bar{\mu}_{\hat{C}_1}$
Traditional Wald Test: $\mathbb{P}_{H_0^{\{\hat{C}_1,\hat{C}_2\}}}(||\bar{X}_{\hat{C}_1} - \bar{X}_{\hat{C}_2}||_2 \geq ||\bar{x}_{\hat{C}_1} - \bar{x}_{\hat{C}_2}||_2)$, which can be calculated using $(\sigma\sqrt{\frac{1}{\hat{C}_1} + \frac{1}{\hat{C}_2}})\cdot\chi_q$.
This is problematic! Even when there's no signal, clustering is maximizing the differences between clusters. In other words, we use the data to select the null $\Rightarrow$ null distribution of the Wald test statistic is not proportional to a χ_q distribution $\Rightarrow$ too aggressive (inflate α).

Introduction & Motivation

2000 simulated data: $\mu = 0_{100\times 2}$ and $\sigma^2 = 1$. Average linkage hierarchical clustering.



Maybe data splitting can remedy that? Do clustering in training set and then do 3-NN in test set.



It still doesn't work...Use the data to select null hypothesis, and $H_0^{\{\hat{C}_1, \hat{C}_2\}}$ is a function of test observations.

Introduction & Motivation

In this paper, they develped a selective inference framework to test for a difference in means after clustering.
key idea: define p-value that conditions on the event $\{\hat{C}_1, \hat{C}_2 \in C(\boldsymbol{X})\}$
Compared to previous research:

- Don't need resampling & provide exact finite-sample inference (if σ is known).
- No need for sample splitting, and allows inference on all the data.

Selective Inference: framework for test

Conditional p-value:

$\mathbb{P}_{H_0^{\{\hat{C}_1,\hat{C}_2\}}}(||\bar{X}_{\hat{C}_1} - \bar{X}_{\hat{C}_2}||_2 \geq ||\bar{x}_{\hat{C}_1} - \bar{x}_{\hat{C}_2}||_2 | \hat{C}_1, \hat{C}_2 \in C(\boldsymbol{X}))$

$H_0^{\{\hat{C}_1,\hat{C}_2\}}$ controls the **selective type I error rate for clustering** if

$\mathbb{P}_{H_0^{\{\hat{C}_1,\hat{C}_2\}}}(\textit{reject } H_0^{\{\hat{C}_1,\hat{C}_2\}} \textit{ at level } \alpha | \hat{C}_1, \hat{C}_2 \in C(\boldsymbol{X})) \leq \alpha$

We can redefine the p-value as :

$$p(\mathbf{x}; \{\hat{\mathcal{C}}_1, \hat{\mathcal{C}}_2\}) = \mathbb{P}_{H_0^{\{\hat{\mathcal{C}}_1,\hat{\mathcal{C}}_2\}}}\Big(\|\bar{X}_{\hat{\mathcal{C}}_1} - \bar{X}_{\hat{\mathcal{C}}_2}\|_2 \geq \|\bar{x}_{\hat{\mathcal{C}}_1} - \bar{x}_{\hat{\mathcal{C}}_2}\|_2 \,\Big|\, \hat{\mathcal{C}}_1, \hat{\mathcal{C}}_2 \in \mathcal{C}(\mathbf{X}), \boldsymbol{\pi}^{\perp}_{\nu(\hat{\mathcal{C}}_1,\hat{\mathcal{C}}_2)}\mathbf{X} = \boldsymbol{\pi}^{\perp}_{\nu(\hat{\mathcal{C}}_1,\hat{\mathcal{C}}_2)}\mathbf{x}, \operatorname{dir}\left(\bar{X}_{\hat{\mathcal{C}}_1} - \bar{X}_{\hat{\mathcal{C}}_2}\right) = \operatorname{dir}\left(\bar{x}_{\hat{\mathcal{C}}_1} - \bar{x}_{\hat{\mathcal{C}}_2}\right)\Big), \quad (8)$$

, where $\boldsymbol{\pi}^{\perp}_{\nu}$ projects onto the orthogonal complement of the vector ν and $\nu(\hat{C}_1, \hat{C}_2)_i = 1\{i \in \hat{C}_1\}|\hat{C}_1| - 1\{i \in \hat{C}_2\}|\hat{C}_2|$

Selective Inference: framework for test

Theorem 1. *For any realization* $\mathbf{x}$ *from* (1) *and for any non-overlapping groups of observations* $\mathcal{G}_1, \mathcal{G}_2 \subseteq \{1, 2, \ldots, n\}$,

$$p(\mathbf{x}; \{\mathcal{G}_1, \mathcal{G}_2\}) = 1 - \mathbb{F}\left(\|\bar{x}_{\mathcal{G}_1} - \bar{x}_{\mathcal{G}_2}\|_2; \sigma\sqrt{\frac{1}{|\mathcal{G}_1|} + \frac{1}{|\mathcal{G}_2|}}, \mathcal{S}(\mathbf{x}; \{\mathcal{G}_1, \mathcal{G}_2\}) \right), \tag{9}$$

where $p(\cdot;\cdot)$ *is defined in* (8), $\mathbb{F}(t; c, \mathcal{S})$ *denotes the cumulative distribution function of a* $c \cdot \chi_q$ *random variable truncated to the set* $\mathcal{S}$, *and*

$$\mathcal{S}(\mathbf{x}; \{\mathcal{G}_1, \mathcal{G}_2\}) = \left\{ \phi \geq 0 : \mathcal{G}_1, \mathcal{G}_2 \in \mathcal{C}\left(\boldsymbol{\pi}^{\perp}_{\nu(\mathcal{G}_1, \mathcal{G}_2)} \mathbf{x} + \left(\frac{\phi}{\frac{1}{|\mathcal{G}_1|} + \frac{1}{|\mathcal{G}_2|}} \right) \nu(\mathcal{G}_1, \mathcal{G}_2) dir(\bar{x}_{\mathcal{G}_1} - \bar{x}_{\mathcal{G}_2})^T \right) \right\}. \tag{10}$$

Furthermore, if $H_0^{\{\mathcal{G}_1, \mathcal{G}_2\}}$ *is true, then* $p(\mathbf{X}; \{\mathcal{G}_1, \mathcal{G}_2\}) \mid \mathcal{G}_1, \mathcal{G}_2 \in \mathcal{C}(\mathbf{X}) \sim \text{Uniform}(0, 1)$, *i.e.*

$$\mathbb{P}_{H_0^{\{\mathcal{G}_1, \mathcal{G}_2\}}} \left(p(\mathbf{X}; \{\mathcal{G}_1, \mathcal{G}_2\}) \leq \alpha \;\middle|\; \mathcal{G}_1, \mathcal{G}_2 \in \mathcal{C}(\mathbf{X}) \right) = \alpha, \quad \forall\ 0 \leq \alpha \leq 1. \tag{11}$$

That is, rejecting $H_0^{\{\mathcal{G}_1, \mathcal{G}_2\}}$ *whenever* $p(\mathbf{x}; \{\mathcal{G}_1, \mathcal{G}_2\})$ *is below* α *controls the selective type I error rate (Definition 1) at level* α.

Selective Inference: framework for test

So, to compute p-value, it suffices to characterize $\hat{S} = S(\boldsymbol{x}; \{\hat{C}_1, \hat{C}_2\}) = \{\phi \geq 0 : \hat{C}_1, \hat{C}_2 \in C(\boldsymbol{x}'(\phi))\}$, where

$$\mathbf{x}'(\phi) = \boldsymbol{\pi}_{\hat{\nu}}^{\perp}\mathbf{x} + \left(\frac{\phi}{1/|\hat{\mathcal{C}}_1| + 1/|\hat{\mathcal{C}}_2|}\right) \hat{\nu}\ \mathrm{dir}(\bar{x}_{\hat{\mathcal{C}}_1} - \bar{x}_{\hat{\mathcal{C}}_2})^T, \quad \hat{\nu} = \nu(\hat{\mathcal{C}}_1, \hat{\mathcal{C}}_2)$$

Not very intuitive... Since $\boldsymbol{x}'\hat{\nu} = \bar{x}_{\hat{\mathcal{C}}_1} - \bar{x}_{\hat{\mathcal{C}}_2}$, the ith row of $\boldsymbol{x}'(\phi)$ is:

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

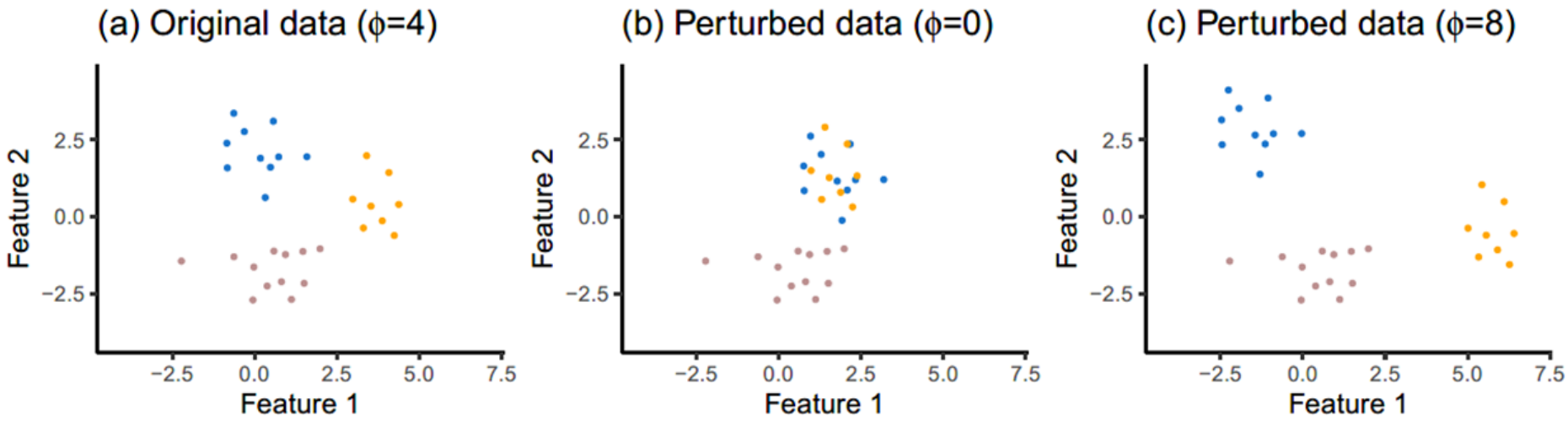
So, that means observations in $\hat{C}_1$ and $\hat{C}_2$ will:

- be pulled apart, if $\phi > ||\bar{x}_{\hat{\mathcal{C}}_1} - \bar{x}_{\hat{\mathcal{C}}_2}||_2$.
- keep the unchanged, if $\phi = ||\bar{x}_{\hat{\mathcal{C}}_1} - \bar{x}_{\hat{\mathcal{C}}_2}||_2$
- be pushed together, if $0 \leq \phi < ||\bar{x}_{\hat{\mathcal{C}}_1} - \bar{x}_{\hat{\mathcal{C}}_2}||_2$

Selective Inference: framework for test

$\boldsymbol{x}'(\phi)$ = perturbed version of $\boldsymbol{x}$:observations in $\hat{C}_1$ and $\hat{C}_2$ will

- be pulled apart, if $\phi > ||\bar{x}_{\hat{C}_1} - \bar{x}_{\hat{C}_2}||_2$.
- keep the unchanged, if $\phi = ||\bar{x}_{\hat{C}_1} - \bar{x}_{\hat{C}_2}||_2$
- be pushed together, if $0 \leq \phi < ||\bar{x}_{\hat{C}_1} - \bar{x}_{\hat{C}_2}||_2$



In (b), the algorithm no longer estimates these clusters. In this example, $\hat{S} = [2.8, \infty)$

Furthermore, $\hat{S}$ describes the set of ϕ, s.t. the algorithm preserves the results after perturbation.

Computing $\hat{S}$ for hierachical clustering

By previous definition, we can simply calculate the p-value by Monte Carlo (later). But we can make use of properties in hierarchical clustering (e.g. dissimilarity & linkage) to save computation time for $\hat{S}$.
Let

$$\left\{\mathcal{W}_1^{(t)}(\mathbf{x}), \mathcal{W}_2^{(t)}(\mathbf{x})\right\} = \underset{\mathcal{G},\mathcal{G}' \in \mathcal{C}^{(t)}(\mathbf{x}), \mathcal{G} \neq \mathcal{G}'}{\arg\min} d(\mathcal{G}, \mathcal{G}'; \mathbf{x})$$

be the 'winning pair' at step t. Then there's a Lemma about the height and merges, after perturbation (next page)

Computing $\hat{S}$ for hierachical clustering

Lemma 1. *Suppose that* $\mathcal{C} = \mathcal{C}^{(n-K+1)}$, *i.e. we perform hierarchical clustering to obtain* K *clusters. Then,*

$$d\left(\mathcal{W}_1^{(t)}(\mathbf{x}), \mathcal{W}_2^{(t)}(\mathbf{x}); \mathbf{x}'(\phi)\right) = d\left(\mathcal{W}_1^{(t)}(\mathbf{x}), \mathcal{W}_2^{(t)}(\mathbf{x}); \mathbf{x}\right), \ \forall\, \phi \geq 0, \forall\, t = 1, \ldots, n-K, \quad (15)$$

where $\left(\mathcal{W}_1^{(t)}(\mathbf{x}), \mathcal{W}_2^{(t)}(\mathbf{x})\right)$ *is the "winning pair" of clusters that merged at the* t^{th} *step of the hierarchical clustering algorithm applied to* $\mathbf{x}$. *Furthermore, for any* $\phi \geq 0$,

$$\hat{\mathcal{C}}_1, \hat{\mathcal{C}}_2 \in \mathcal{C}(\mathbf{x}'(\phi)) \quad \text{if and only if} \quad \mathcal{C}^{(t)}(\mathbf{x}'(\phi)) = \mathcal{C}^{(t)}(\mathbf{x}) \ \forall\, t = 1, \ldots, n-K+1. \quad (16)$$

In short, the height and merges in the first $n - K$ steps keep the same after perturbation.

Computing $\hat{S}$ for hierachical clustering

If we further define the "losing pairs" as:

$$\mathcal{L}(\mathbf{x}) = \bigcup_{t=1}^{n-K} \left\{ \{\mathcal{G}, \mathcal{G}'\} : \mathcal{G}, \mathcal{G}' \in \mathcal{C}^{(t)}(\mathbf{x}), \mathcal{G} \neq \mathcal{G}', \{\mathcal{G}, \mathcal{G}'\} \neq \left\{ \mathcal{W}_1^{(t)}(\mathbf{x}), \mathcal{W}_2^{(t)}(\mathbf{x}) \right\} \right\}$$

, and lower & upper bound of life time:

$$l_{\mathcal{G},\mathcal{G}'}(\mathbf{x}) \equiv \min \left\{ 1 \leq t \leq n-K : \mathcal{G}, \mathcal{G}' \in \mathcal{C}^{(t)}(\mathbf{x}), \{\mathcal{G}, \mathcal{G}'\} \neq \{\mathcal{W}_1^{(t)}(\mathbf{x}), \mathcal{W}_2^{(t)}(\mathbf{x})\} \right\}$$

$$u_{\mathcal{G},\mathcal{G}'}(\mathbf{x}) \equiv \max \left\{ 1 \leq t \leq n-K : \mathcal{G}, \mathcal{G}' \in \mathcal{C}^{(t)}(\mathbf{x}), \{\mathcal{G}, \mathcal{G}'\} \neq \{\mathcal{W}_1^{(t)}(\mathbf{x}), \mathcal{W}_2^{(t)}(\mathbf{x})\} \right\}$$

Then, we can calculate $\hat{S}$ as shown in the next page.

Computing $\hat{S}$ for hierachical clustering

Theorem 2. *Suppose that $\mathcal{C} = \mathcal{C}^{(n-K+1)}$, i.e. we perform hierarchical clustering to obtain K clusters. Then, for $\hat{\mathcal{S}}$ defined in (12),*

$$\hat{\mathcal{S}} = \bigcap_{\{\mathcal{G},\mathcal{G}'\}\in\mathcal{L}(\mathbf{x})} \left\{ \phi \geq 0 : d(\mathcal{G},\mathcal{G}';\mathbf{x}'(\phi)) > \max_{l_{\mathcal{G},\mathcal{G}'}(\mathbf{x})\leq t\leq u_{\mathcal{G},\mathcal{G}'}(\mathbf{x})} d\left(\mathcal{W}_1^{(t)}(\mathbf{x}), \mathcal{W}_2^{(t)}(\mathbf{x}); \mathbf{x}\right) \right\}, \quad (18)$$

where $\left\{\mathcal{W}_1^{(t)}(\mathbf{x}), \mathcal{W}_2^{(t)}(\mathbf{x})\right\}$ is the pair of clusters that merged at the t^{th} step of the hierarchical clustering algorithm applied to $\mathbf{x}$, $\mathcal{L}(\mathbf{x})$ is defined in (17) to be the set of "losing pairs" of clusters in $\mathbf{x}$, and $[l_{\mathcal{G},\mathcal{G}'}(\mathbf{x}), u_{\mathcal{G},\mathcal{G}'}(\mathbf{x})]$ is the lifetime of such a pair of clusters in $\mathbf{x}$. Furthermore, (18) is the intersection of $\mathcal{O}(n^2)$ sets.

They further show details about when and how these sets can be efficiently computed.
In particular, by specializing to squared Euclidean distance & a certain class of linkages, each of these sets is defined by a single quadratic inequality and the coefficients can be efficiently computed.

Extentions

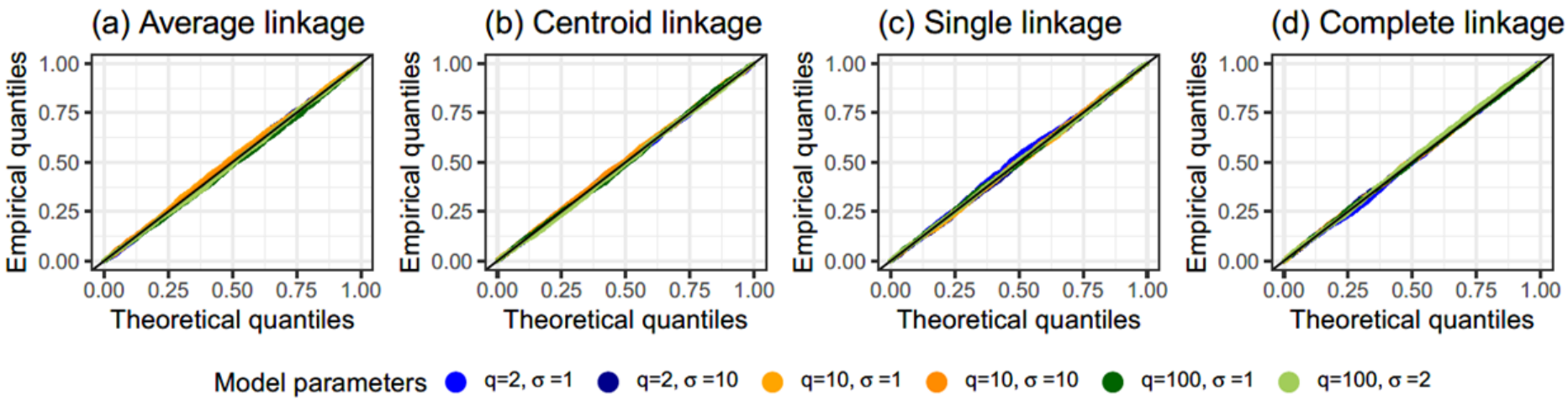
Extension 1: Monte Carlo approximation to p-value
The previous results only apply for hierarchical clustering, excluding the complete linkage. So we have to estimate p-value by Monte Carlo:

$$p(\mathbf{x}; \{\hat{\mathcal{C}}_1, \hat{\mathcal{C}}_2\}) = \mathbb{E}\left[\mathbb{1}\left\{\phi \geq \|\bar{x}_{\hat{\mathcal{C}}_1} - \bar{x}_{\hat{\mathcal{C}}_2}\|_2, \hat{\mathcal{C}}_1, \hat{\mathcal{C}}_2 \in \mathcal{C}(\mathbf{x}'(\phi))\right\}\right] / \mathbb{E}\left[\mathbb{1}\left\{\hat{\mathcal{C}}_1, \hat{\mathcal{C}}_2 \in \mathcal{C}(\mathbf{x}'(\phi))\right\}\right]$$

for $\phi \sim (\sqrt{\frac{1}{\hat{\mathcal{C}}_1} + \frac{1}{\hat{\mathcal{C}}_2}}) \cdot \chi_q$
To sample efficiently, we can further implement importance sampling.
Extension 2: Non-spherical covariance matrix: when
$\boldsymbol{X} \sim MN_{n\times q}(\boldsymbol{\mu}, \boldsymbol{I}_n, \boldsymbol{\Sigma})$
Extension 3: Unknown variance: just plug in the estimate of σ.

Simulation Results: Uniform p-values

$\boldsymbol{X} \sim MN_{n\times q}(\boldsymbol{\mu}, \boldsymbol{I}_n, \sigma^2 \boldsymbol{I}_q)$, with $\boldsymbol{\mu} = 0_{n\times q}$
Simulate 2000 data sets for $n = 150$, $\sigma \in \{1, 2, 10\}$ and $q \in \{2, 10, 100\}$



Conditional power and detection probability

They further check the conditional power and detection probability. Still under the setting $\boldsymbol{X} \sim MN_{n\times q}(\boldsymbol{\mu}, \boldsymbol{I}_n, \sigma^2\boldsymbol{I}_q)$, with $n = 30$. But now there are 3 equidistant clusters,

$$\mu_1 = \cdots = \mu_{\frac{n}{3}} = \begin{bmatrix} -\delta/2 \\ 0_{q-1} \end{bmatrix}, \mu_{\frac{n}{3}+1} = \cdots = \mu_{\frac{2n}{3}} = \begin{bmatrix} 0_{q-1} \\ \sqrt{3}\delta/2 \end{bmatrix}, \mu_{\frac{2n}{3}+1} = \cdots = \mu_n = \begin{bmatrix} \delta/2 \\ 0_{q-1} \end{bmatrix}$$

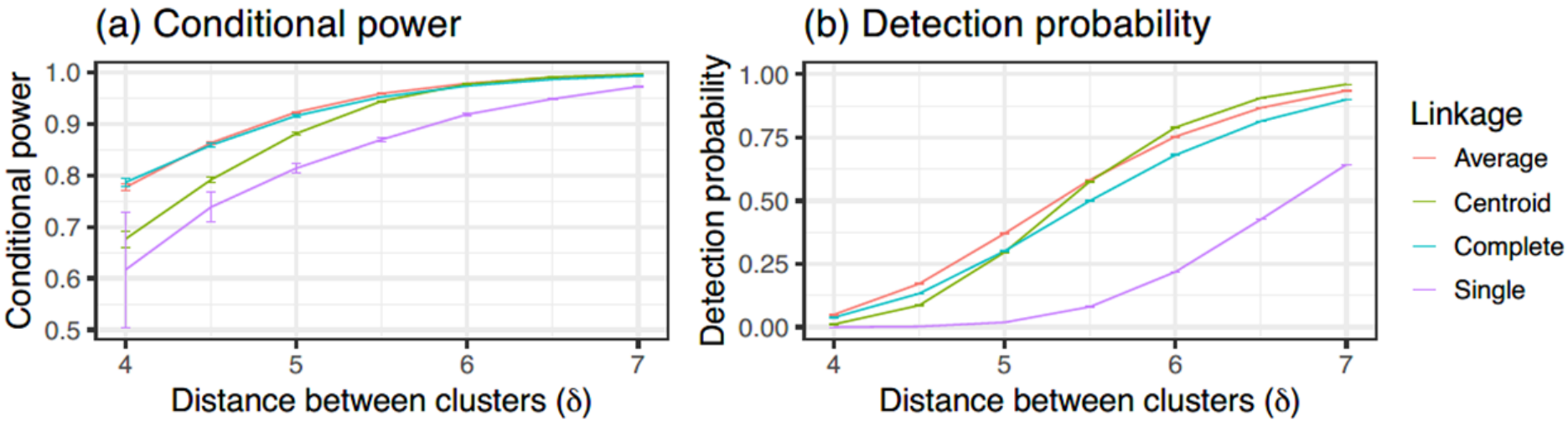
They simulated 300,000 dat sets for $\sigma = 1$, $q = 10$ and 7 evenly-spaced values of $\delta \in [4, 7]$. The significance level is $\alpha = 0.05$. The conditional power is defined as:

$$\frac{\#\text{ data sets where we reject } H_0^{\{\hat{\mathcal{C}}_1, \hat{\mathcal{C}}_2\}}, \text{ and } \hat{\mathcal{C}}_1 \text{ and } \hat{\mathcal{C}}_2 \text{ are true clusters}}{\#\text{ data sets where } \hat{\mathcal{C}}_1 \text{ and } \hat{\mathcal{C}}_2 \text{ are true clusters}}$$

The detection probability is defined as:

$$\frac{\#\text{ data sets where } \hat{\mathcal{C}}_1 \text{ and } \hat{\mathcal{C}}_2 \text{ are true clusters}}{300,000}$$

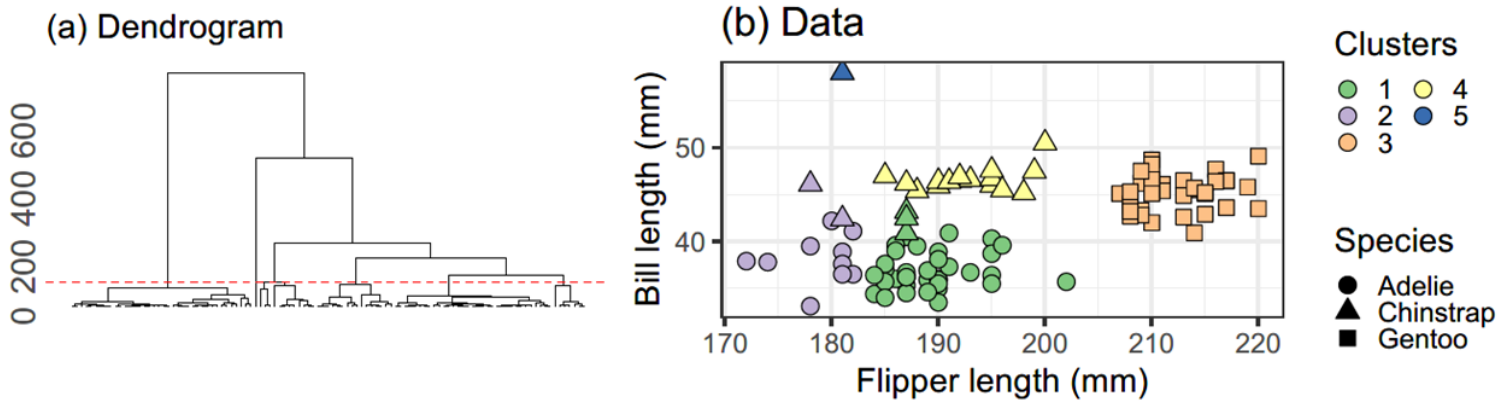
Conditional power and detection probability



- conditional power and detection probability increases as distance between clusters increase
- Conditional power: average and complete linkage highest; single linkage lowest.
- Detection probabilities: average, centroid & complete $\gg$ single

Application 1: Palmer penguins

First, estimate σ from a separated data: bill length and flipper length of 58 female penguins in 2009.
Then do average linkage hierarchical clustering with squared Euclidean distance to 107 penguins in 2007-2008. The features are still bill length and flipper length.



Cluster pairs	$(1,2)$	$(1,3)$	$(1,4)$	$(2,3)$	$(2,4)$	$(3,4)$
Test statistic	10.1	25.0	10.1	33.8	17.1	18.9
Our p-value	0.591	1.70×10^{-14}	0.714	0.070	0.291	2.10×10^{-6}
Wald p-value	0.00383	$< 10^{-307}$	0.00101	$< 10^{-307}$	4.29×10^{-5}	1.58×10^{-11}

Application 2: Single-cell RNA sequencing data

After pre-processing, there are 2 datasets (all with 500 genes):

- "no clusters": randomly sampling 600 memory T cells.
- "clusters": randomly sampling 200 each of memory T cells, B cells and monocytes.

Apply ward-linkage hierarchical clustering with squared Euclidean distance to data. The Σ in $\boldsymbol{X} \sim MN_{n\times q}(\boldsymbol{\mu}, \boldsymbol{I}_n, \Sigma)$ is estimated by principal complement thresholding ('POET') to the left out data set:

- "no clusters": 3 pseudo-clusters, containing 64, 428 and 108 cells.
- "clusters": nearly recover the true clusters.

corrected p-values make sense:

	"No clusters"			"Clusters"l		
Cluster pairs	(1, 2)	(1, 3)	(2, 3)	(1, 2)	(1, 3)	(2, 3)
Test statistic	4.05	4.76	2.96	3.04	4.27	4.38
Our p-value	0.20	0.27	0.70	4.60×10^{-28}	3.20×10^{-82}	1.13×10^{-73}
Wald p-value	$< 10^{-307}$	$< 10^{-307}$	$< 10^{-307}$	$< 10^{-307}$	$< 10^{-307}$	$< 10^{-307}$