ENSAI

ATPA Track
Academic Year 2024–2025

Methodological Project Report

Selective Inference for Hierarchical Clustering

Student: Jasmin Neveu

Supervisor: Javier González-Delgado

Analyzed article:
Lucy L. Gao, Jacob Bien, and Daniela Witten,
Selective Inference for Hierarchical Clustering,
Journal of the American Statistical Association,
119(545), 332–342, 2024.

Submission date: January 20, 2026

Contents

1 Introduction **2**

2 Classical versus selective testing **3**
2.1 Hypothesis testing and *p-values* 3
2.2 Examples motivating selective testing 6
2.2.1 Lasso 6
2.2.2 Publication bias 8
2.2.3 Clustering 8
2.3 Addressing selective testing 9
2.3.1 Simultaneous Inference 9
2.3.2 Sample splitting 10
2.3.3 Conditional Inference 10

3 Post clustering inference **11**
3.1 Notation and preliminaries 11
3.2 Gao et al. approach 11

A Proofs **12**
A.1 Proofs of Section 2 12
A.2 Proofs of Section 3 14

1 Introduction

Introduction text...

2 Classical versus selective testing

2.1 Hypothesis testing and *p-values*

General comments: This section is well-written and structured, but some work needs to be done. I have added some main comments about how to present some of the objects. We will discuss about that. Once this is done, we will speak about improving the flow by adding some text that helps the reader and creates a 'story'.

As we have a 20 pages limit we will probably have to move the proofs to the appendix (which is the usual practice in research articles). I have added an appendix at the end where you can move the proofs. Then, we will mention in the text that proofs are provided in the Appendix.

Minor comment: to write equations, use

$$2+2 \tag{1}$$

so that equation numbers appear in the text and equations can be referenced therein, using (1). If you don't want an equation to be numbered, because it maybe not be very relevant, or it corresponds to calculations inside a proof, use

$$1+1.$$

Also, Definitions and Remarks are numbered like 2.1, 2.2 but Propositions like 1,2,3... The same numbering should be used for all.

I think it is better is presented as follows (we will discuss next time about this). The main point is that I think is better to work directly on E (the topological space where the random variable takes values) for clarity. We will clarify next time.

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, $(E, \mathcal{E})$ a topological space and $\mathcal{T}$ the σ-algebra generated by $\mathcal{E}$. A *random variable* is a measurable function $X : (\Omega, \mathcal{F}) \to (E, \mathcal{T})$. The *(probability) distribution* of X is the mapping $P : \mathcal{T} \to [0,1]$ such that $P(O) = (\mathbb{P} \circ X^{-1})(O)$ for all $O \in \mathcal{T}$. We say that P is *supported* on E and denote by $\mathcal{M}_1(E)$ the set of all probability distributions supported on $\mathcal{E}$. From now on, we will set $E = \mathbb{R}$ and simply write $\mathcal{M}_1(\mathbb{R}) = \mathcal{M}_1$.

Probably the previous paragraph should be formulated more in detail, especially when defining $\mathcal{M}_1$. To discuss.

Definition 1 (Hypothesis, Definition 1.1 in [4])**.** *A hypothesis is a set of probability distributions in $\mathcal{M}_1$. A hypothesis is simple if it is a singleton, such as $\{P\}$ or $\{Q\}$. Otherwise it is composite.*

We define the alternative hypothesis to $\mathcal{H}_0$ as $\mathcal{M}_1 \setminus \mathcal{H}_0$.

Definition 2 (Test)**.** *A test is a function defined as*

$$\begin{array}{rcl} \pi : \Omega & \to & \{0,1\} \\ \omega & \mapsto & \pi(\omega). \end{array}$$

Definition 3 (Rejection rule)**.** *Let $\mathcal{H}_0 \subset \mathcal{M}_1$ be a hypothesis and let $\pi : \Omega \to \{0,1\}$ be a test. For $\omega \in \Omega$, we say that the test* rejects *$\mathcal{H}_0$ if $\pi(\omega) = 1$, and* does not reject *$\mathcal{H}_0$ if $\pi(\omega) = 0$.*

The test function is defined on Ω (it makes a decision based on an observation of the sample space, not based on a measurable set!). In the following definition you speak about rejection of $\mathcal{H}_0$ but this has not been defined. There has to be a link between the definition of test function (if defined and used) and the concept of rejecting/accepting a hypothesis. **Importantly**: You have defined a (simple) hypothesis as a distribution $\{P\}$. Then, what do we mean by *rejecting* a hypothesis? This should be defined (no need to define a test function for that if you don't want to). We need to define rejection of $\mathcal{H}_0$ before speaking about type I error. If this is not clear we will discuss next time.

Definition 4 (Type I error)**.** *We say that a test controls the type I error at level α if*

$$\mathbb{P}(\textit{Reject } \mathcal{H}_0) \leq \alpha, \quad \textit{for } \alpha \in (0,1).$$

We say that it controls the type I error exactly at level α if

$$\mathbb{P}(\textit{Reject } \mathcal{H}_0) = \alpha, \quad \textit{for } \alpha \in (0,1).$$

Note: It is a good practice to write $\forall\, u$ instead of $\forall u$. I modified this in the following equations.

Definition 5 (Stochastic dominance)**.** *Let X and Y be real-valued random variables. We say that Y stochastically dominates X, and write*

$$X \preceq_{\text{st}} Y,$$

if

$$\mathbb{P}(X \leq u) \geq \mathbb{P}(Y \leq u), \quad \forall\, u \in \mathbb{R}.$$

Note: Use $\mathcal{SU}(0,1)$ instead of $SU(0,1)$. I have also replaced it.

Definition 6 (Super-uniform random variable)**.** *Let X be a real-valued random variable. We say that X is super-uniform, and write $X \sim \mathcal{SU}(0,1)$, if X stochastically dominates a uniform random variable on $[0,1]$, that is, if*

$$\mathbb{P}(X \leq u) \leq u, \quad \forall\, u \in [0,1].$$

Proposition 1. *Let X and Y be real-valued random variables, with cumulative distribution functions F_X and F_Y, respectively. If $X \preceq_{\text{st}} Y$, then $F_X(Y) \sim SU(0,1)$.*

Remark 1. *If $X = Y$, then $F_X(X)$ is super-uniform on $[0,1]$. Moreover, if X has a continuous distribution function F_X, then*

$$F_X(X) \sim \mathcal{U}(0,1).$$

Note: Write p-value instead of p-value.

Definition 7 (p-value, Definition 1.1 in [4])**.** *Let $\mathcal{H}_0$ be a hypothesis. A p-value for $\mathcal{H}_0$ is a super-uniform random variable under $\mathcal{H}_0$.*

p-values are often used to build a test by defining the partition of the sample sapce using the rejection rule $\mathcal{R} = \mathbf{1}_{\{p \leq \alpha\}}$ for any $\alpha \in (0, 1)$.

Proposition 2. *Let $\mathcal{H}_0$ be a hypothesis, p a p-value for $\mathcal{H}_0$ and $\mathcal{R}$ the rejection rule defined by*

$$\mathcal{R} = \mathbf{1}_{\{p \leq \alpha\}}, \quad \alpha \in (0, 1).$$

Then, $\mathcal{R}$ controls the type I error at level α.

Here you use the term 'rejection rule' (which is fine), that is closely related to 'test' (defined above). This should be clarified: either use only test, either use only 'rejection rule', or (better) define 'rejection rule' as a particular type of test based on the p-value thresholding. After defining p-value, you can say that p-values are often used to build a test by defining the partition of the sample space using the rejection rule $\nVdash\{p \leq \alpha$, for any $\alpha \in (0, 1)$.

Importantly: in the previous proposition you say that the 'rejection rule controls the type I error' but the type I error control has been defined in Def. 2.3 for a 'test'. My previous comment should help clarify this.

Definition 8 (Test statistic)**.** *A test statistic is a measurable function*

$$T : \Omega \to \mathbb{R}$$

From now on, we will consider the case of unilateral tests. In this setting, we define the the p-value has the form:

$$p(X) = \mathbb{P}_{\mathcal{H}_0}(T(X) \geq t(x))$$

with T, t being test statistics what is T?

More generally, the p-value for unilateral test will be characterized using test statistics T. If it is a proposition it can't be a definition! The p-value has already been defined in Def. 2.6, so it can't be defined again. What we are doing here is *characterizing* the p-value for a unilateral test in terms of a statistic T.

Proposition 3. *Let T and T' be two test statistics. If T transforms X it can't be taking values from $\mathcal{F}$! Let X be a random variable and x a realization of X. Define*

$$p(x) = \mathbb{P}_{\mathcal{H}_0}\big(T'(X) \geq T(x)\big).$$

If

$$T'(X) \preceq_{\text{st}} T(X) \quad \textit{under } \mathcal{H}_0,$$

then p is a p-value for $\mathcal{H}_0$.

Remark 2. *If $T' = T$ and the distribution function F_T of $T(X)$ is continuous under $\mathcal{H}_0$, then*

$$p \overset{\mathcal{H}_0}{\sim} \mathcal{U}(0,1).$$

Here we should have a paragraph introduce what selective testing is. Just saying that in the classical setting the null hypothesis is independent of the data but in several settings $\mathcal{H}_0$ is chosen after seeing the data, what makes the classical testing approaches unsuitable. This is called selective testing, and it is motivated with some examples in the next section. The reader needs to have an idea of what we mean by selective testing before starting reading the examples.

In the classical setting, the null hypothesis is independent of the data but in several setting $\mathcal{H}_0$ is chosen after seeing the data, what makes the classical testing approaches unsuitable. This is called selective testing, and it is motivated with some examples in the next section.

2.2 Examples motivating selective testing

2.2.1 Lasso

To determine whether an explanatory variable helps explain a response variable through a linear regression model, a common practice is to test whether its associated coefficient is significantly different from zero. In the context of classical linear regression, this falls within the framework of non-selective inference.

In contrast, Lasso regression uses an ℓ_1-penalty to perform variable selection. Not all coefficients are tested—only those selected by the Lasso are considered. It is therefore impossible to define in advance which coefficients will be tested, since they depend on the outcome of the Lasso regression. This corresponds to a setting of selective inference.

let $x_1, \ldots, x_8$ be centered gaussian random variables with correlation matrix $(R_{i,j})_{1 \leq i,j \leq 8}$ defined by

$$R_{i,j} = 0.5^{|i-j|}, \quad 1 \leq i, j \leq 8.$$

The response is modeled as

$$y = \beta^\top x + 3\varepsilon, \quad \varepsilon \sim \mathcal{N}(0,1), \quad \beta = (3, 1.5, 0, 0, 2, 0, 0, 0)^\top.$$

We generated a sample of size $n = 100$ and obtained the following regularization path.

Say instead: to illustrate the unsuitability of classical inference in this context, we simulate samples... peform a LASSO for each sample, etc. You can add in a footnote that this is based on the code by Brian Staber.

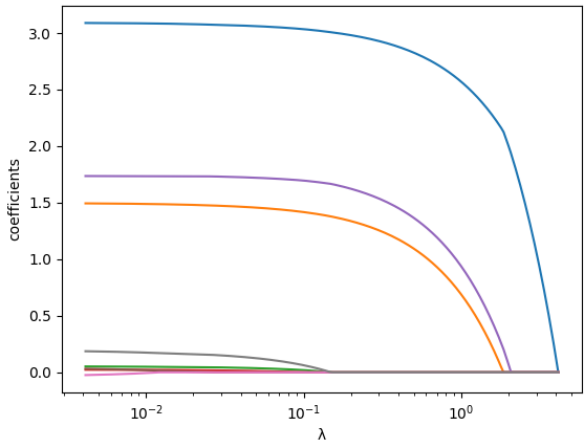

Figure 1: Regularization path of the coefficients obtained using the Lasso algorithm. Some coefficients are zero, and three converge toward the theoretical values (3,1.5,2).

Some explanation is needed about the next figure. We tested $\beta_3 =$ for each simulated sample and obtained the empirical p-value distribution depicted in Figure 2. p-values are not super-uniform, therefore... We tested $\beta_3 = 0$ for each simulated sample and obtained and obtained the empirical p-value distribution depitced in Figure 2. p-values are not super-uniform, therefore a selective inference approach should be used instead of naive testing after selection.

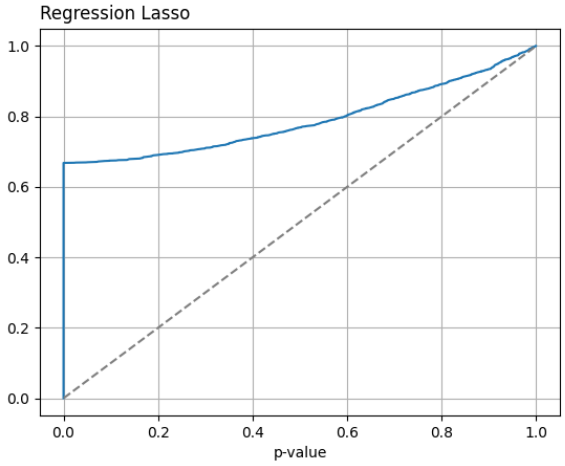

Figure 2: Cumulative distribution function of the p-values obtained from significance tests of the third coefficient estimated by a Lasso regression. The empirical distribution function was computed using $M = 2000$ simulations.

2.2.2 Publication bias

Most published studies gain publication due to their demonstrated significance. Testing statistical effects from such studies falls within selective inference, as these studies have undergone a selection process. If $Y_i \sim \mathcal{N}(\mu_i, 1)$ represents the effect size of a scientific study and only those with $|Y_i| > 1$ are published, denoted $\hat{I} = \{i : |Y_i| > 1\}$, then a naive level α test $H_{0,i} : \mu_i = 0$ for $i \in \hat{I}$ is invalid. Indeed, Fithian [2] demonstrates that the false positive rate among true nulls reaches approximately 0.16, far exceeding the nominal 0.05 level. Valid inference requires thresholding $|Y_i|$ at 2.41 rather than 1.96, the 0.95 quantile of the standard normal, imposing a more stringent criterion.

Leave the clustering example at the end.

2.2.3 Clustering

Another remarkable example of selective inference appears when evaluating the performance of clustering algorithms by testing for the equality of cluster mean. To illustrate the need of using appropriate tests in the context of selective inference, we simulated samples of gaussian random variables that were classified into $K = 3$ groups using hierarchical clustering and k-means

I would start maybe like: Another remarkable example of selective inference appears when evaluating the performace of clustering algorithms by testing for the equality of cluster means (or something like that).

Even if it is only you, avoid using 'I' and use 'we' or use passive tenses or forms like 'we can simulate', 'this can be illustrated by simulating', etc.

For each sample, equality of cluster means was tested using a classical z-test, for two randomly selected clusters. If the test controlled the type I error in this setting, the resulting p-value would be uniformly distributed under the null hypothesis. However, the 'clustering + post-selection testing' procedure leads to a deviation from the uniform distribution, demonstrating the lack of control of the type I error, as shown in Figure 3.

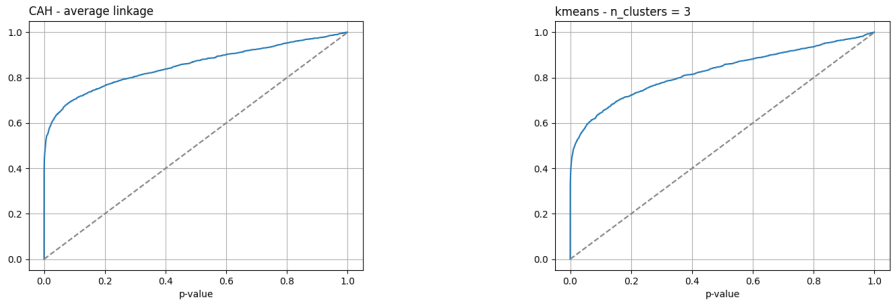

Figure 3: Cumulative distribution functions of the p-values obtained from tests of equality of means between clusters after hierarchical clustering (CAH) and k-means algorithms. The distribution functions were computed using $M = 2000$ simulations from a multivariate normal distribution with $\mu = 0_{n\times p}$ I guess you mean 0_p and $\Sigma = 0.98^{|i-j|}$ for $1 \leq i, j \leq n \times p$, I don't understand the $n \times p$ here. with $n = 100$ et $p = 5$. Mention that you set the clustering algorithms to choose $K = 3$ clusters, and that you tested the equality of cluster means for two randomly selected clusters (if that's the case).

Write k-means instead of *k-means*, CAH instead of *CAH* and p-value instead of p-value.

2.3 Addressing selective testing

Since the type I error is no longer controlled, alternative approaches are required. Accordingly, we present three methods as described by Yoav Benjamini [1].

2.3.1 Simultaneous Inference

This approach controls the family-wise error rate across all hypotheses:

$$\mathbb{P}(\text{At least 1 false positive among all hypotheses}) \leq \alpha.$$

This strategy proves highly conservative, ensuring that for every possible

set of hypotheses, the probability of at least one false positive remains below α.

2.3.2 Sample splitting

This approach consists in splitting the dataset into a training set X and a test set Y. The training set X is used to choose which hypotheses to test, denoted $H_0(X)$. Then, the tests are performed on the test set Y.

Although relatively simple to implement, this method raises several issues. First, statistical guarantees on the tests hold only if X and Y are independent, which is rarely the case in practice. In addition, comparing cluster means on the observations in Y requires assigning each test point to one of the clusters obtained from the clustering performed on X, a step that compromises validity. As discussed in [3], this strategy does not yield valid post-clustering inference in general.

2.3.3 Conditional Inference

This approach constitutes the most extensively studied framework for post-clustering inference. It controls the false positive rate conditional on hypothesis selection:

$$\mathbb{P}(\text{Reject } H_0(X) \mid H_0 \text{ selected}) \leq \alpha.$$

In the remainder of the article, we employ this method to develop a statistical procedure for testing equality of cluster means following a clustering algorithm.

3 Post clustering inference

3.1 Notation and preliminaries

Let $X \in \mathbb{R}^{n \times p}$ be the design matrix. A cluster is an element of a partition of the samples. We note C a clustering algorithm and $C_1 \in C(X)$ a cluster obtained by the algorithm C on X.

We note $\mu = (\mu_{i,j})_{i,j}$ such as $\mu_{i,j} = \mathbb{E}[X_{i,j}]$, with $X_{i,j}$ the element in row i and column j of the matrix X. For a subset G of $\{1, \dots, n\}$, we note $\overline{\mu}_G = \frac{1}{|G|} \sum_{i \in G} \mu_i \in \mathbb{R}^p$ and $\overline{X}_G = \frac{1}{|G|} \sum_{i \in G} X_i \in \mathbb{R}^p$

Definition 9 (Null Hypothesis)**.**

$$H_0^{\{C_1, C_2\}} : \overline{\mu}_{C_1(X)} = \overline{\mu}_{C_2(X)}$$

Definition 10 (Type I selective error for clustering)**.** *We say that a test controls the type I selective error for clustering at leval α if*

$$\mathbb{P}_{H_0^{\{C_1, C_2\}}}(\textit{Reject } H_0^{\{C_1, C_2\}} \mid C_1, C_2 \in C(X)) \leq \alpha, \quad \alpha \in (0, 1).$$

We say that it constrols exactly the type I selective error for clustering at level α if

$$\mathbb{P}_{H_0^{\{C_1, C_2\}}}(\textit{Reject } H_0^{\{C_1, C_2\}} \mid C_1, C_2 \in C(X)) = \alpha, \quad \alpha \in (0, 1).$$

To run a proper test, we need to control this error at level α. In the ideal, we would like to define a p-value as following:

$$p_{ideal}(x) = \mathbb{P}_{H_0^{\{C_1, C_2\}}}(T(X) \geq T(x) \mid C_1, C_2 \in C(X))$$

with T being a test statistic.

With this p-value, we can control the selective type I error for clustering.

Proposition 4. *The selection rule $\mathcal{R} = \mathbf{1}_{\{p_{ideal} \leq \alpha\}}, \quad \alpha \in (0, 1)$. controls the selective type I error for clustering at leval α*

However, p_{ideal} cannot be evaluated in pratice as it depends on parameters that are unknown. Thus, to address this issue, we need to add technical events to the conditioning set and consided:

$$p(x) = \mathbb{P}_{H_0^{\{C_1, C_2\}}}(T(X) \geq T(x) \mid C_1, C_2 \in C(X), E[X])$$

3.2 Gao et al. approach

A Proofs

A.1 Proofs of Section 2

Proof of Proposition 1. Let G_X denote the generalized inverse (quantile function) of F_X, defined for $u \in [0,1]$ as

$$G_X(u) = \inf\{x \in \mathbb{R} : F_X(x) \geq u\}.$$

By definition of the generalized inverse,

$$\{F_X(Y) \leq u\} = \{Y < G_X(u)\}.$$

Therefore,

$$\mathbb{P}\big(F_X(Y) \leq u\big) = \mathbb{P}\big(Y < G_X(u)\big) = F_Y\big(G_X(u)^-\big),$$

where $G_X(u)^-$ denotes the left limit at $G_X(u)$.

Since $X \preceq_{\text{st}} Y$, we have $F_Y \leq F_X$ pointwise, and thus

$$F_Y\big(G_X(u)^-\big) \leq F_X\big(G_X(u)^-\big).$$

By the defining property of the generalized inverse,

$$F_X\big(G_X(u)^-\big) \leq u.$$

Combining these inequalities yields

$$\mathbb{P}\big(F_X(Y) \leq u\big) \leq u, \quad \forall\, u \in [0,1],$$

which proves that $F_X(Y)$ is super-uniform.

□

Proof of Remark 1. By the proposition 1, we have that $F_X(X) \sim SU(0,1)$. If F_X is continuous, then $F_X(G_X(u)) = u$ for all $u \in [0,1]$, and hence

$$\mathbb{P}\big(F_X(X) \leq u\big) = F_X\big(G_X(u)\big) = u, \quad \forall\, u \in [0,1],$$

which concludes the proof. It's okay, but no need to repeat what we are proving at the end.

□

Proof of Proposition 2. By definition of the rejection rule,

$$\mathbb{P}_{\mathcal{H}_0}(\textit{Reject } \mathcal{H}_0) = \mathbb{P}_{\mathcal{H}_0}(p \leq \alpha).$$

Since p is a p-value for $\mathcal{H}_0$, it is super-uniform under $\mathcal{H}_0$, hence

$$\mathbb{P}_{\mathcal{H}_0}(p \leq \alpha) \leq \alpha.$$

This establishes control of the type I error at level α. □

Proof of Proposition 3. Let $F_{T'(X)}$ denote the distribution function of $T'(X)$ under $\mathcal{H}_0$. By Proposition 1, the stochastic dominance $T'(X) \preceq_{\text{st}} T(X)$ implies that

$$F_{T'(X)}\big(T(X)\big) \sim \text{SU}(0,1).$$

By definition,

$$p(x) = \mathbb{P}_{\mathcal{H}_0}\big(T'(X) \geq T(x)\big) = 1 - F_{T'(X)}\big(T(x)\big).$$

Let $u \in [0,1]$. Then

$$\begin{aligned}\mathbb{P}_{\mathcal{H}_0}\big(p(X) \leq u\big) &= \mathbb{P}_{\mathcal{H}_0}\big(1 - F_{T'(X)}(T(X)) \leq u\big) \\ &= \mathbb{P}_{\mathcal{H}_0}\big(F_{T'(X)}(T(X)) \geq 1 - u\big) \\ &= 1 - \mathbb{P}_{\mathcal{H}_0}\big(F_{T'(X)}(T(X)) \leq 1 - u\big)\,.\end{aligned}$$

Since $F_{T'(X)}(T(X))$ is super-uniform,

$$\mathbb{P}_{\mathcal{H}_0}\big(F_{T'(X)}(T(X)) \leq 1 - u\big) \leq 1 - u,$$

and therefore

$$\mathbb{P}_{\mathcal{H}_0}\big(p(X) \leq u\big) \leq u.$$

Thus, p is super-uniform under $\mathcal{H}_0$, and hence a p-value. □

Proof of Remark 2. When $T' = T$

$$p(X) = \mathbb{P}_{\mathcal{H}_0}\big(T(X') \geq T(X)\big) = 1 - F_T\big(T(X)\big),$$

By Remark 1, if F_T is continuous,

$$F_T\big(T(X)\big) \sim \mathcal{U}(0,1).$$

Consequently, for any $u \in [0,1]$,

$$\begin{aligned}\mathbb{P}_{\mathcal{H}_0}\big(p(X) \leq u\big) &= \mathbb{P}_{\mathcal{H}_0}\big(1 - F_T(T(X)) \leq u\big) \\ &= \mathbb{P}_{\mathcal{H}_0}\big(F_T(T(X)) \geq 1 - u\big) \\ &= 1 - \mathbb{P}_{\mathcal{H}_0}\big(F_T(T(X)) \leq 1 - u\big) \\ &= 1 - (1 - u) \\ &= u.\end{aligned}$$

Thus p is uniformly distributed on $[0,1]$ under $\mathcal{H}_0$. □

A.2 Proofs of Section 3

Proof of Proposition 4. By definition of the rejection rule,

$$\mathbb{P}_{H_0^{\{C_1,C_2\}}}\big(\text{Reject } H_0^{\{C_1,C_2\}} \mid C_1, C_2 \in C(X)\big) = \mathbb{P}_{H_0^{\{C_1,C_2\}}}\big(p_{\text{ideal}} \leq \alpha \mid C_1, C_2 \in C(X)\big)$$

Since p_{ideal} is a p-value for $\mathcal{H}_0^{\{C_1,C_2\}}$, it is super uniform under $\mathcal{H}_0^{\{C_1,C_2\}}$, hence

$$\mathbb{P}_{H_0^{\{C_1,C_2\}}}\big(\text{Reject } H_0^{\{C_1,C_2\}} \mid C_1, C_2 \in C(X)\big) \leq \alpha$$

This establishes control of the type I error at leval α. □

References

[1] Y. Benjamini. Selective Inference: The Silent Killer of Replicability. *Harvard Data Science Review*, 2(4), dec 16 2020. https://hdsr.mitpress.mit.edu/pub/l39rpgyc.

[2] W. Fithian, D. Sun, and J. Taylor. Optimal inference after model selection, 2017. arXiv:1410.2597.

[3] L. L. Gao, J. Bien, and D. Witten. Selective inference for hierarchical clustering. *Journal of the American Statistical Association*, 119(545):332–342, 2024.

[4] A. Ramdas and R. Wang. Hypothesis testing with e-values. *arXiv*, Oct. 2024.