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 26, 2026

Contents

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

$$1 + 1.$$

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 $\mathcal{H}_0$ is a set of probability distributions in $\mathcal{M}_1$. A hypothesis is simple if it is a singleton, such as $\{P\}$ or $\{Q\}$, and composite otherwise~~. Otherwise it is composite~~. The complementary set~~We define the alternative hypothesis to $\mathcal{H}_0$ as~~ $\mathcal{M}_1 \setminus \mathcal{H}_0$ is called its alternative hypothesis.*

Definition 2 (Test). *A test for $\mathcal{H}_0$ is a binary partition of the sample space, defined by a mapping~~function defined as~~*

$$\begin{aligned} \pi_{\mathcal{H}_0} : \Omega &\to \{0, 1\} \\ \omega &\mapsto \pi_{\mathcal{H}_0}(\omega). \end{aligned} \tag{2}$$

For $\omega \in \Omega$, we say that the test rejects $\mathcal{H}_0$ based on ω if $\pi_{\mathcal{H}_0}(\omega) = 1$, and does not reject $\mathcal{H}_0$ based on ω if $\pi_{\mathcal{H}_0}(\omega) = 0$.

I have moved the 'rejection' definition inside the test one. I think this is clearer. Then, no need to speak of *rejection rule* but only of *test* (equivalent concepts). I know I spoke about rejection rule but I think this is clearer.

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 $\pi_{\mathcal{H}}(\omega) = 1$, and does not reject $\mathcal{H}_0$ if $\pi(\omega) = 0$.*~~

In what follows, we will be writing $\mathbb{P}(\text{Reject } \mathcal{H}_0)$, which is a standard and easy-to-read form. Before starting using that, you should write a note explaning that this is a notation that you will be adopting from now on, meaning:

$$\mathbb{P}\left(\text{Reject } \mathcal{H}_0\right) = \mathbb{P}\left(\{\omega \in \Omega \,:\, \pi_{\mathcal{H}_0}(\omega) = 1\}\right),$$

according to your previous definition. You can add that we make the dependence on π implicit. In short: it is okay to use shortcuts as 'Reject H0' but we always need to formally specify what we mean by them.

Don't use italics for *Reject* in equations.

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

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

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

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

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 $SU(0,1)$!

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 $X \preceq_{\mathrm{st}} Y$, 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 $\alpha \in (0, 1)$.

Not amazing to start a sentence with the word p-value. Also you can use $\mathbb{1}$ for the indicator function.

The previous paragraph is okay, but (following my previous comment) I think we should only speak about *test* and avoid *rejection rule*. So you can say that we often build tests using p-values as $\pi = \mathbb{1}\{p \leq \alpha\}$. So in the following proposition you can also replace $\mathcal{R}$ by π.

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.

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, to define a test based on the information given by a real-valued random variable X, p-values are often built in the form~~we define the the p-value has the form~~:

$$p(x)\underline{p(X)} = \mathbb{P}_{\mathcal{H}_0}(T(X) \geq T(x)\underline{t(x)}), \tag{3}$$

where T is a test statistic and x is a realization of X.~~with T, t being test statistics~~

More generally, the p-value for unilateral test will be characterized using test statistics T.

The previous sentence does not clearly justify why do we propose the following proposition. I would recall the following: 1. The p-value (3) has a uniform distribution under $\mathcal{H}_0$. 2. However, we have defined p-values as being SU, which is a more general family of distributions. 3. To adapt the form of the p-value to that setting, we adopt the construction of the following proposition.

Proposition 3. *Let T and T' be two test statistics. 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

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

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*

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

In the classical setting, the null hypothesis is independent of the data. However, in many practical applications~~ but in several setting~~ $\mathcal{H}_0$ is chosen *after seeing the data*. In this setting~~, what makes~~ the classical testing approaches built for type I error control are unsuitable. Instead, statistical guarantees need to be provided via *selective inference*. In particular, the theory of statistical testing of data-driven null hypotheses is known as *selective testing*[2]. The next section provides some examples that motivate this framework.~~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 cite the Lasso paper. As only the coefficients selected by the Lasso can be tested, the null hypothesis depends~~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, and is therefore data-driven~~.~~ This corresponds to a setting of selective inference, where the classical control of type I error fails~~.~~ To illustrate so, consider a centered Gaussian vector $X = (X_1, \ldots, X_8)$ with correlation matrix $(R_{i,j})_{1 \leq i,j \leq 8}$, defined by Standard notation: upper case for random variables and lower case for their realizations. Also, Gaussian is written in upper case.

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

The response is modeled as

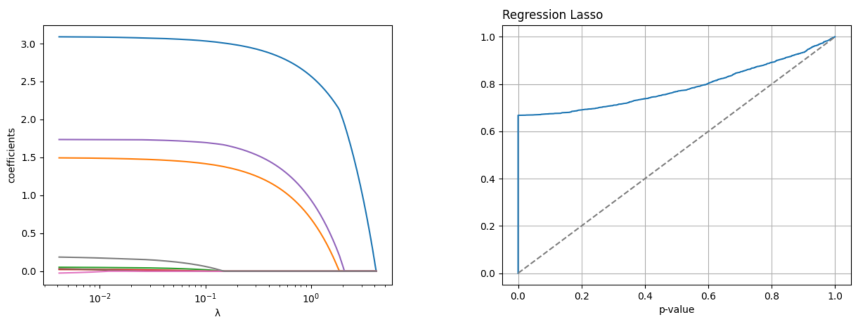
$$Y = \beta^\top X + 3\varepsilon, \quad \varepsilon \sim \mathcal{N}(0,1), \quad \beta = (3, 1.5, 0, 0, 2, 0, 0, 0)^\top.$$

Is ε one-dimensional?

We generated $n = 100$ independent realizations of (X, Y) and implemented the Lasso algorithm, obtaining~~a sample of size $n = 100$ and obtained~~ the ~~following~~ regularization path presented in Figure 1(a). Then, we tested $\beta_3 = 0$. After repeating this pipeline $M = 2000$ times, we obtained~~for each simulated sample and obtained and obtained~~ the empirical p-value distribution depicted in Figure 1(b). p-values are not super-uniform, therefore a selective inference approach should be used instead of naive testing after selection. Avoid starting a sentence with the word p-value.

Advice: use two panels in the same figure. I have modified it but you can undo if you don't like.

I have modified a bit the text above but it is still not very clear. The simulation needs to be explained more clearly. One question: the coefficient β_3 is always selected across the M simulations?



(a) (b)

Figure 1: (a) Regularization path of the coefficients obtained using the Lasso algorithm. This is for a single realization of (X, Y), right? Some coefficients are zero, and three converge toward the theoretical values (3,1.5,2). (b) Empirical cumulative distribution function (ECDF) of the p-values obtained from testing $\beta_3 = 0$~~significance tests of the third~~ after selecting the coefficient~~ coefficient estimated~~ by a Lasso regression. The ECDF~~empirical distribution function~~ was computed from~~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 $Y_i \sim \mathcal{N}(\mu_i, 1)$

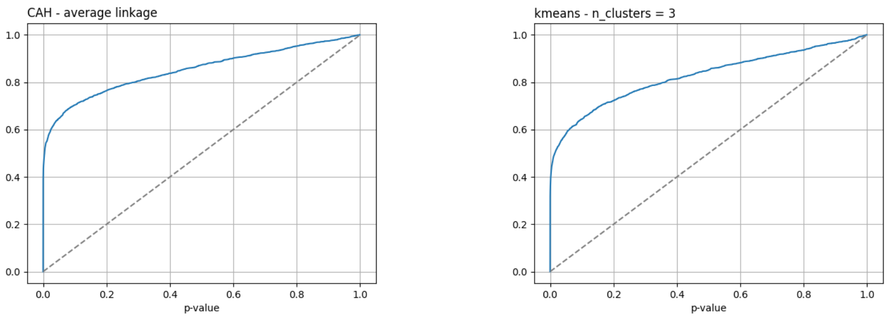
represents the effect size of a scientific study and only those with $|Y_i| > 1$ are published, denoted $\hat{I} = \{i : |Y_i| > 1\}$, then a naive level α test $H_{0,i} : \mu_i = 0$ for $i \in \hat{I}$ is invalid. Indeed, Fithian [2] demonstrates that the false positive rate among true nulls reaches 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.

To discuss together!

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 means. To illustrate the need of using appropriate tests in the context of selective inference, we simulated samples of Gaussian~~gaussian~~ random variables Not clear! Gaussian random vectors defined how? And how many samples? that were classified into $K = 3$ groups using hierarchical clustering and k-means.

For each sample, the 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 (super-)uniformly distributed under the null hypothesis. However, the 'clustering + post-selection testing' procedure leads to a deviation from super-uniformity~~the uniform distribution, demonstrating the lack of control of the type I error~~, as shown in Figure 2.



(a) (b)

Figure 2: Cumulative distribution functions of the p-values obtained from tests of equality of means between clusters after (a) hierarchical agglomerative clustering (HAC)~~(CAH)~~ and (b) 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).

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.

Why new page?

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*

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

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

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

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 $T'(X) \preceq_{\mathrm{st}} T(X)$ implies that

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

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. $\square$

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.