ENSAI

ATPA Track
Academic Year 2024–2025

Methodological Project Report

Selective Inference for Hierarchical Clustering

Student: Jasmin Neveu

Supervisor: Javier Gonzáles-Delgado

Analyzed article:
Lucy L. Gao, Jacob Bien, and Daniela Witten,
Selective Inference for Hierarchical Clustering,
arXiv: 2012.02936v3 [stat.ME] 31 Oct 2022

Submission date: January 11, 2026

Contents

1 Introduction

Introduction text...

2 Classical versus selective testing

2.1 Hypothesis testing and *p-values*

Let $(\Omega, \mathcal{F})$ and $(E, \mathcal{E})$ be a measurable spaces. We denote by $\mathcal{M}_1$ the set of all probability measures on $(\Omega, \mathcal{F})$. A random variable with values in E is a measurable function $X : (\Omega, \mathcal{F}, \mathbb{P}) \to (E, \mathcal{E})$, with $\mathbb{P} \in \mathcal{M}_1$

Definition 2.1 (Hypothesis, 1.1 in [2])**.** *A hypothesis is a set of probability measures in $\mathcal{M}_1$. A hypothesis is simple if it is a singleton, such as $\{\mathbb{P}\}$ or $\{\mathbb{Q}\}$. Otherwise it is composite.*

Remark 2.1. *We define the alternative hypothesis to $\mathcal{H}_0$ as $\mathcal{M}_1 \setminus \mathcal{H}_0$.*

Definition 2.2 (Test)**.** *A test is a function defined as*

$$\begin{array}{rcl} \pi : \mathcal{F} & \to & \{0,1\} \\ x & \mapsto & \pi(x). \end{array}$$

Definition 2.3 (Type I error)**.** *We say that a test controls the type I error at level α if*

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

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

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

Definition 2.4 (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

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

Definition 2.5 (Super uniform)**.** *Let X be a real-valued random variable. We say that X is super uniform, and write $X \sim SU(0,1)$, if X stochastically dominates a uniform random variable on $[0,1]$, that is,*

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

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

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

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

By definition of the generalized inverse,

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

Therefore,

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

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

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

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

By the defining property of the generalized inverse,

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

Combining these inequalities yields

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

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

Remark 2.2. *If $X = Y$, then $F_X(X)$ is super uniform on $[0,1]$. Moreover, if X has a continuous distribution function F_X, then*

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

Proof. 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

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

Therefore $F_X(X) \sim \mathcal{U}(0,1)$ when X is continuously distributed. □

Definition 2.6 (p-value, 1.1 in [2]). *Let $\mathcal{H}_0$ be a hypothesis. A p-value for $\mathcal{H}_0$ is a super uniform random variable under $\mathcal{H}_0$.*

Proposition 2. *Let $\mathcal{R}$ be the rejection rule defined by*

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

Let p be a p-value for $\mathcal{H}_0$. Then $\mathcal{R}$ controls the type I error at level α.

Proof. By definition of the rejection rule,

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

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

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

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

For the rest of the document, we will only consider unilateral test. Usually, we define the *p-value* by

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

But more generally, we will define *p-value* with the following proposition.

Proposition 3. *Let T and T' be two test statistics, i.e. transformations of X*

$$T : \mathcal{F} \to \mathbb{R}, \qquad T' : \mathcal{F} \to \mathbb{R}.$$

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

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

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

Proof. Under $\mathcal{H}_0$, let $F_{T'(X)}$ denote the distribution function of $T'(X)$. 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. □

Remark 2.3. *If $T' = T$ and the distribution function F_T of $T(X)$ is continuous under $\mathcal{H}_0$, then*

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

Proof. 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 2.2, 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 $\mathcal{H}_0$. □

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, one performs a significance test on the coefficient associated with that variable, testing whether it is significantly different from zero. In the context of classical linear regression, this falls within the framework of standard 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 creates a situation of selective inference.

During the Algorithmic Programming course supervised by Brian Staber, we implemented a proximal gradient descent algorithm. We simulated a dataset as follows: let $x_1, \dots, x_8$ be centered gaussian random variables with correlation matrix $(R_{ij})_{1 \leq i,j \leq 8}$ defined by

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

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.

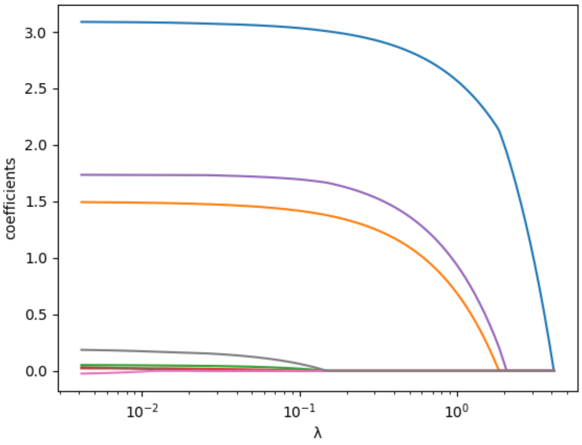

Figure 1: Regularization path of the coefficients obtained using the Lasso algorithm. Some coefficients are zero, and three converge toward the theoretical values (3,1.5,2).

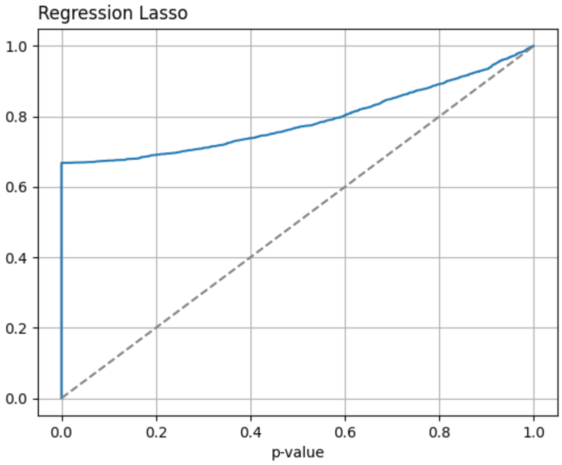

Figure 2: Cumulative distribution function of the p-values obtained from significance tests of the third coefficient estimated by a Lasso regression. The distribution function was computed using $M = 2000$ simulations.

2.2.2 Clustering

To illustrate the importance of using appropriate tests in the context of selective inference, I simulated samples of gaussian random variables to which I applied clustering methods (hierarchical clustering and *k-means*). I then performed tests of equality of means (*z-tests*) on clusters chosen in a datadependent manner. If the test of equality of means correctly controlled the type I error in this setting, the resulting p-value would follow a uniform distribution on ([0,1]) when the null hypothesis is true. In practice, 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.

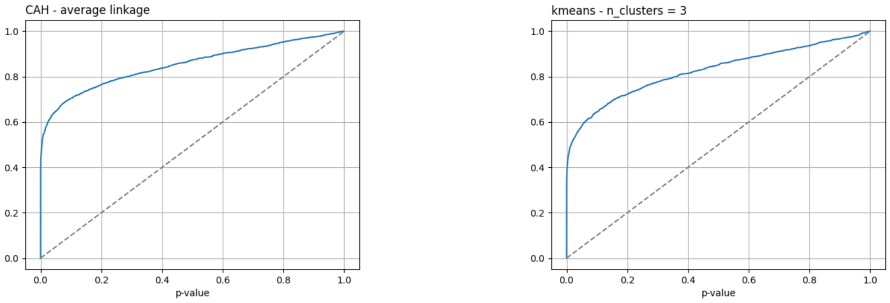

Figure 3: Cumulative distribution functions of the p-values obtained from tests of equality of means between clusters after hierarchical clustering (CAH) and *k-means* algorithms. The distribution functions were computed using $M = 2000$ simulations from a multivariate normal distribution with $\mu = 0_{n\times p}$ et $\Sigma = 0.98^{|i-j|}$ pour $1 \leq i, j \leq n \times p$, avec $n = 100$ et $p = 5$.

2.2.3 Publication bias

[1]

2.3 Addressing selective testing

3 Post clustering inference

References

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

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