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 14, 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

$$1+1.$$

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.

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 $\mathbb{P} \in \mathcal{M}_1$

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

Remark 2.1. *We define the alternative hypothesis to $\mathcal{H}_0$ as $\mathcal{M}_1 \setminus \mathcal{H}_0$. If you are defining something it must be inside a definition (not a remark). Add this sentence to the previous defintion.*

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

$$\begin{aligned} \pi : \Omega\underline{\mathcal{F}} &\to \{0,1\} \\ \omega\underline{x} &\mapsto \pi(\omega\underline{x}). \end{aligned}$$

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

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

We say that it controls ~~exactly~~ 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 2.4 (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 $SU(0,1)$. I have also replaced it.

Definition 2.5 (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 $F_X(Y) \sim SU(0,1)$.*

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

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

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 $X \preceq_{\mathrm{st}} Y$, 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. □

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

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

which concludes the proof.~~Therefore $F_X(X) \sim \mathcal{U}(0,1)$ when X is continuously distributed.~~ It's okay, but no need to repeat what we are proving at the end. □

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

Definition 2.6 (p-value, Definition 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 $\mathcal{H} : 0$ be a hypothesis, p a p-value for $\mathcal{H}_0$ and $\mathcal{R}$ ~~be~~ the rejection rule defined by*

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

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

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.

Proof. By definition of the rejection rule,

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

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. □

In this work/From now on,~~For the rest of the document~~, we will ~~only~~ consider the case of unilateral tests. In this setting~~Usually~~, we define the the p-value has the form~~*p-value* by~~:

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

what is T? More~~But more~~ generally, we will define *p-value* with the following proposition. 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, i.e. transformations of X*

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

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

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

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

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

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.

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~~one performs a significance test on the coefficient associated with that variable, testing~~ whether it associated coefficient it~~is~~ significantly different from zero. In the context of classical linear regression, this falls within the framework of non-selective~~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 corresponds to a setting of~~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 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.

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

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.

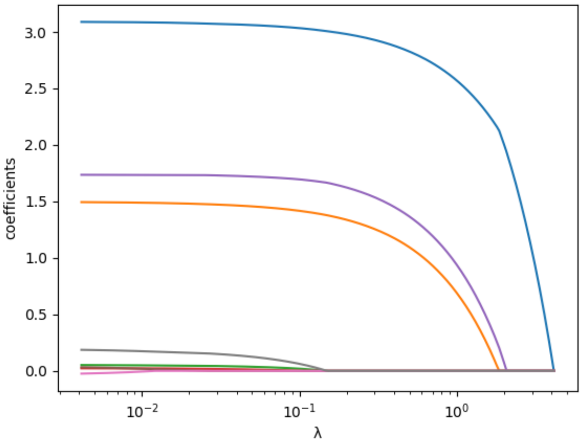

Figure 1: Regularization path of the coefficients obtained using the Lasso algorithm. Some coefficients are zero, and three converge toward the theoretical values (3,1.5,2).

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...

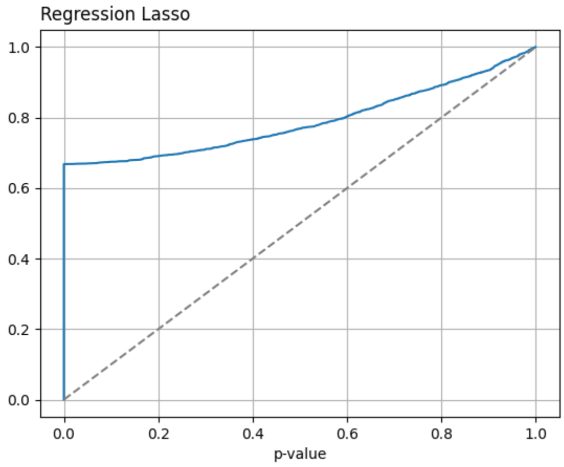

Figure 2: Cumulative distribution function of the p-values obtained from significance tests of the third coefficient estimated by a Lasso regression. The empirical distribution function was computed using $M = 2000$ simulations.

2.2.2 Publication bias

[1] Leave the clustering example at the end.

2.2.3 Clustering

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). To illustrate the need~~importance~~ of using appropriate tests in the context of selective inference, I simulated samples of gaussian random variables that were classified into $K = 3$ groups using~~to which I applied clustering methods~~ (hierarchical clustering and k-means~~*k-means*~~).

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, the I then performed tests of equality of cluster means was tested using a classical z-test, for two randomly selected clusters.~~(*z-tests*) on clusters chosen in a data-dependent manner.~~ If the test ~~of equality of means~~ ~~correctly~~ controlled the type I error in this setting, the resulting p-value would be uniformly distributed under the null.~~follow a uniform distribution on ([0,1]) when the null hypothesis is true.~~ However,~~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, as shown in Figure 3.

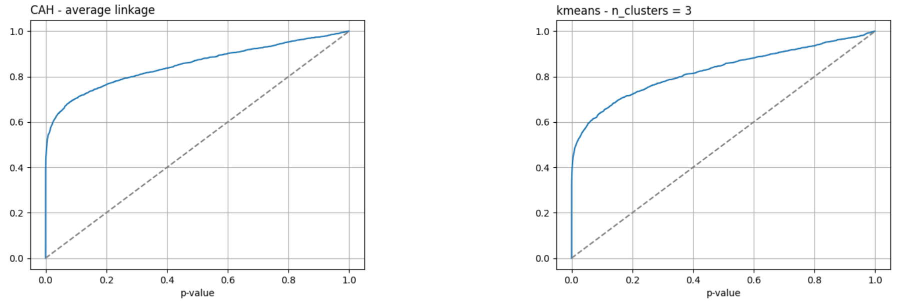

Figure 3: Cumulative distribution functions of the p-values obtained from tests of equality of means between clusters after hierarchical clustering (*CAH*) and *k-means* algorithms. The distribution functions were computed using $M = 2000$ simulations from a multivariate normal distribution with $\mu = 0_{n\times p}$ I guess you mean 0_p et $\Sigma = 0.98^{|i-j|}$ pour $1 \leq i, j \leq n \times p$, I don't understand the $n \times p$ here. avec $n = 100$ 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

3 Post clustering inference

A Proofs

A.1 Proofs of Section 2

Here use:

Proof of Proposition 1. □

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.