Notes d'avancement du projet méthodologique de deuxième année

Jasmin NEVEU
Référent : Javier GONZALEZ-DELGADO

13 décembre 2025

GitHub : link ↗
Article de recherche : Selective Inference for Hierarchical Clustering ↗

1 Introduction

Ce document a pour objectif de suivre l'avancement de mon projet méthodologique de deuxième année à l'ENSAI. Il contient des notes, des observations et les étapes réalisées tout au long de la progression du projet.

2 Compréhension du problème

Les méthodes de clustering ne disposent pas de garanties théoriques sur les sorties qu'elles génèrent. Pour s'assurer que les résultats soit cohérents aux attentes, il est très fréquent de réaliser des tests d'égalités des moyennes parmi les différents clusters. Cependant un simple z-test ou test de Wald mène à une inflation de l'erreur de type I. En effet, le choix du test se fait après avoir observé les données (après le passage de l'algorithme de clusterin g). Ce problème est similaire à celui de l'*overfitting* pour les algorithmes d'apprentissage supervisé. Si les données sont à la fois utilitées pour l'entrainement et l'évaluation, le modèle risque de ne pas généraliser et de trop *coller* aux données. Ici, faire un test sans prendre en compte l'aléa apporté par les algorithmes de clustering mène à une non maitrise du risque de type I.

Inférence classique	Inférence sélective
1. Hypothèse sur la loi de X	1. Hypothèse sur la loi de X
2. Choix de la question Q	2. Observation des données
3. Tests en utilisant $X \to H_0$	3. Choix de la question $Q(X)$
	4. Tests en utilisant $X \to H_0(X)$

TABLE 1 – Type d'inférence

3 Importance de l'inférence sélective

Pour montrer l'importance de réaliser des tests adaptés dans le cadre de l'inférence selective, j'ai simulé des tirages de variables aléatoires gaussiens sur lesquelles j'ai appliqué des méthodes de clustering (CAH et $kmeans$), j'ai ensuite réalisé des tests d'égalités des moyennes (z-$test$) de clusters choisi de manière aléatoires. Si le test controle correctement l'erreur de type I, alors la $pvalue$ du test doit suivre une loi uniforme sur $[0, 1]$. En effet,...

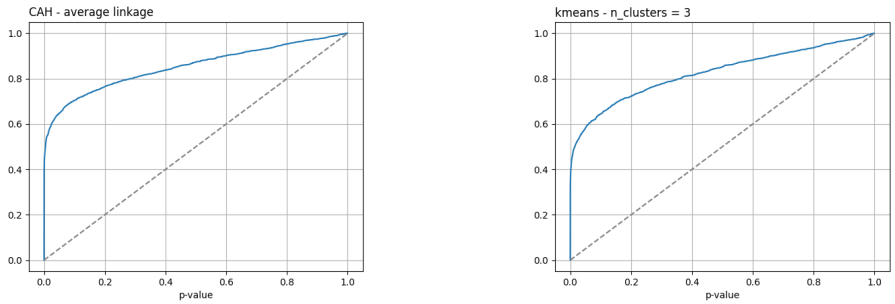

FIGURE 1 – Fonction de répartition des p-values obtenues par tests d'égalité des moyennes des clusters après un algorithme de CAH et de $kmeans$. Les fonctions de répartitions ont été calculées à partir de $M = 2000$ simulations d'une loi normale multivariée avec $\mu = 0_{n \times p}$ et $\Sigma = 0.98^{|i-j|}$ pour $1 \leq i, j \leq n \times p$. J'ai fixé $n = 100$ et $p = 5$.

On voit clairement qu'ici l'erreur de type I n'est pas controlé i.e $\mathbb{P}(... \leq ...) > \alpha$