Notes d'avancement du projet méthodologique de deuxième année

Jasmin NEVEU
Référent : Javier GONZALEZ-DELGADO

14 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, en général, de garanties théoriques fortes sur les partitions qu'elles génèrent dans les situations pratiques usuelles. Pour s'assurer que les résultats sont cohérents avec les attentes, il est très fréquent de réaliser des tests d'égalité 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 des tests et des comparaisons se fait après avoir observé les données (après le passage de l'algorithme de clustering), donc de manière adaptative.

Ce problème est similaire à celui de l'*overfitting* pour les algorithmes d'apprentissage supervisé. Si les données sont à la fois utilisées pour l'entraînement et l'évaluation, le modèle risque de ne pas généraliser et de trop *coller* aux données. Ici, faire un test sans prendre en compte la variabilité introduite par la procédure de clustering mène à une absence de maîtrise 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

Tout l'enjeu est donc de maîtriser cette erreur de type I. Pour cela on définie la *p-valeur*. La *p-valeur* s'interprète comme la probabilité de se tromper si on rejette H_0. Formellement elle est défini comme suit.

Soit,

— un espace d'observation $\mathcal{X}$ et une famille de lois $(P_\theta)_{\theta\in\Theta}$,

— une hypothèse nulle $H_0 \subset \Theta$,

— une statistique de test $T : \mathcal{X} \to \mathbb{R}$,

— une région de rejet correspondant à de grandes valeurs de T (test unilatéral à droite).

Pour un échantillon X de loi P_θ et une réalisation x, on note la statistique observée

$$t_{\text{obs}} = T(x).$$

Pour $\theta_0 \in H_0$, on note F_{T,θ_0} la fonction de répartition de T sous P_{θ_0}.

Définition formelle (test unilatéral à droite). La p-valeur associée à l'observation x est définie par

$$p(x) = \sup_{\theta\in H_0} P_\theta\big(T(X) \geq T(x)\big).$$

Test unilatéral à gauche. Lorsque la région de rejet correspond à de petites valeurs de T,

$$p(x) = \sup_{\theta\in H_0} P_\theta\big(T(X) \leq T(x)\big).$$

Test bilatéral. Dans le cas bilatéral, une définition classique est

$$p(x) = \sup_{\theta\in H_0} P_\theta\big(|T(X)| \geq |T(x)|\big),$$

c'est-à-dire la plus grande probabilité, sous H_0, d'observer une valeur de la statistique de test au moins aussi extrême que celle observée.

On rejette H_0 si la *p-valeur* $\leq \alpha$.

On remarque que la *p-valeur* s'écrit à partir d'une fonction de répartition d'une variable aléatoire, et est elle même une variable aléatoire. Une méthode pour savoir si un test est bien construit est donc d'observer que la *p-valeur* suit bien une loi uniforme.

Proposition 1. *Soit X une variable aléatoire continue, de fonction de répartition F définie sur I, strictement croissante. Alors $F(X)$ suit une loi uniforme sur $]0,1[$.*

Démonstration. On note G la fonction quantile, définie sur $]0,1[$ par

$$G(\omega) = \inf\{x \in \mathbb{R} \mid F(x) \geq \omega\}.$$

Si F est strictement croissante sur I, alors G est la réciproque de F sur $]0,1[$.

Pour $y \in]0,1[$, on a

$$\mathbb{P}\big(F(X) \leq y\big) = \mathbb{P}\big(X \leq G(y)\big) = F\big(G(y)\big) = y,$$

ce qui est bien la fonction de répartition de la loi uniforme $]0,1[$. □

3 Importance de l'inférence sélective

3.1 Clustering

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'égalité des moyennes (*z-test*) de clusters choisis de manière aléatoires. Si le test d'égalité des moyennes contrôlait correctement l'erreur de type I dans ce cadre, alors la *p-valeur* devrait suivre une loi uniforme sur $[0,1]$ lorsque l'hypothèse nulle est vrai. Or, en pratique, la procédure "clustering + test post-sélection" conduit à une déviation de cette loi uniforme, ce qui montre la non-maîtrise du risque de type I.

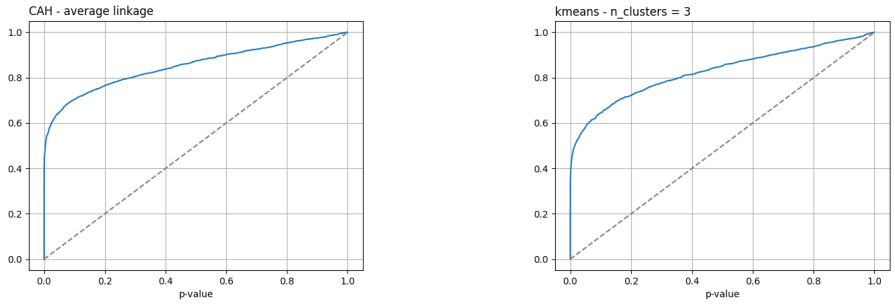

FIGURE 1 – Fonction de répartition des *p-valeur* obtenues par tests d'égalité des moyennes des clusters après un algorithme de *CAH* et de *kmeans*. Les fonctions de répartitions ont été calculées à partir de $M = 2000$ simulations d'une loi normale multivariée avec $\mu = 0_{n\times p}$ et $\Sigma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

3.2 Régression Lasso

Pour savoir si une variable explicative permet d'expliquer une varibale cible au travers d'un modèle de regression linéaire, on effectue un test de significativité du coefficient à associé à cette variable. On test s'il est significativement différent de 0. Dans le cadre d'une regression linéaire, on se situe bien dans le cadre de l'inférence classique. En revanche, la régression Lasso, la pénalisation L_1 permet de selectionner des variables. Ainsi tout les coefficients ne sont pas à tester, seuls ceux qui n'ont pas été selectionner le sont. On ne peut donc pas définir par avance quels coefficients tester car il dépendent de cette regression Lasso On se retrouve donc bien dans une situation d'inférence sélective. Lors du cours de *Programmation Algotithmique* encadré par Brian Staber, nous avons pu coder un algorithme de descente proximale. On a simulé un jeu de donnée de la sorte :

Soient $x_1, \dots, x_8$ des variables aléatoires gaussiennes centrées, de matrice de corrélation $(R_{ij})_{1 \leq i,j \leq 8}$ définie par $R_{ij} = 0.5^{|i-j|}$, pour $1 \leq i, j \leq 8$. La réponse est modélisée par $y = \beta^\top x + 3\varepsilon$, où $\varepsilon \sim \mathcal{N}(0,1)$ et $\beta = (3,\ 1.5,\ 0,\ 0,\ 2,\ 0,\ 0,\ 0)^\top$. On génère un échantillon de taille $n = 100$. On obtient le chemin de régularisation suivant

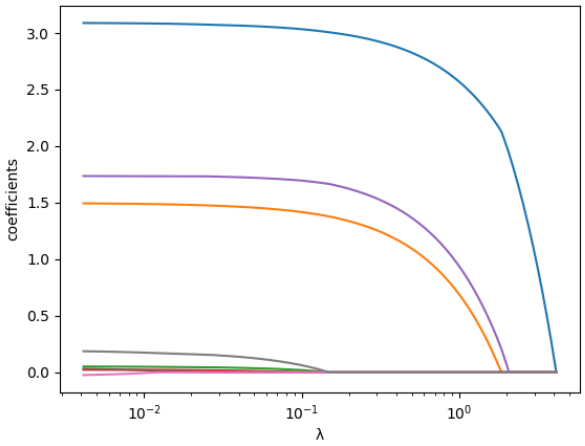

FIGURE 2 – Chemin de régularisation des coefficients obtenus par algorithme Lasso. On remarque que les certains coefficients sont nuls et que 3 autres s'approchent des valeurs théoriques (3,1.5,2).

J'ai chosi de tester la significativité du 3eme coefficient, qui en théorie est nul (j'aurais très bien pu choisir le 4, 6,7 ou 8eme). Si $|\hat{b_3}| > 1e^{-6}$ alors j'estime que la 3eme variable est selectionnée, dans ce cas je calcul la *p-valeur* associée au test de significativité. On obtient la fonction de répartition suivante de la *p-valeur*.

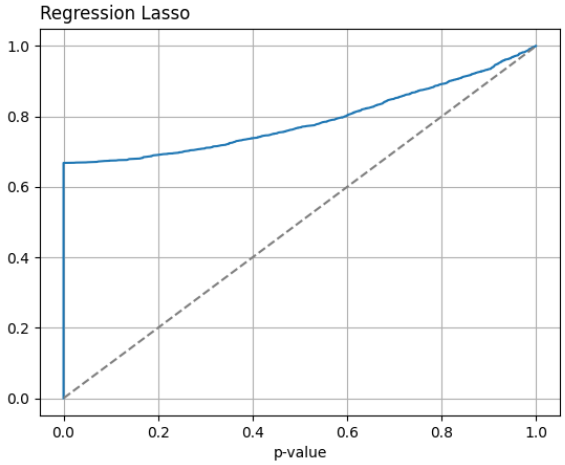

FIGURE 3 – Fonction de répartition des *p-valeur* obenutes par tests de significativité du 2eme coefficient obtenu par une regression Lasso. La fonction de répartition à été calculées à partir de $M = 2000$ simulations.