Inférence Post-Clustering

Nicolas Enjalbert-Courrech [1,a] & Cathy Maugis-Rabusseau [1,b] & Pierre Neuvial [1,c]

[1] *Institut de Mathématiques de Toulouse; UMR5219 Université de Toulouse; CNRS,*
[a] *UPS, F-31062 Toulouse Cedex 9, France France*
nicolas.enjalbert-courrech@math.univ-toulouse.fr
[b] *INSA, F-31077 Toulouse, France*
cathy.maugis@insa-toulouse.fr
[c] *UPS, F-31062 Toulouse Cedex 9, France France*
pierre.neuvial@math.univ-toulouse.fr

Résumé. On s'intéresse au problème de "double-dipping" c'est-à-dire à l'utilisation du même jeu de données pour faire d'abord un clustering des observations puis un test statistique dont l'hypothèse nulle dépend des classes obtenues à l'étape précédente. Dans un premier temps, nous effectuons un état de l'art des nombreuses approches récemment proposées pour ce problème, en les regroupant en deux catégories : les méthodes de partitionnement de l'information et les approches conditionnelles. Ensuite, nous proposons une comparaison numérique afin d'évaluer leur performance en termes de contrôle du risque de première espèce, de puissance statistique, et de temps de calcul.

Mots-clés. Tests d'hypothèse, clustering, double-dipping, test post-sélection.

Abstract. This work tackles the problem of "double-dipping," which refers to the use of the same dataset to first perform clustering of observations and then conduct a statistical test, where the null hypothesis depends on the clusters obtained in the previous step. Initially, we conduct a review of the numerous approaches recently proposed for this problem, grouping them into two categories: information partitioning methods and conditional approaches. Then, we propose a numerical comparison to evaluate their performance in terms of controlling the Type I error rate, statistical power, and computation time.

Keywords. Hypothesis testing, clustering, double-dipping, selective inference.

1 Introduction

Dans ce travail, nous nous intéressons au problème de "double-dipping" c'est-à-dire à l'utilisation des mêmes données pour faire 1) un clustering des individus puis 2) un test basé sur les classes obtenues. Nous nous plaçons dans un cadre Gaussien où pour chaque individu $i \in \{1, \dots, n\}$, $X_i \sim \mathcal{N}_p(\mu_i, \Sigma)$ et les X_i sont indépendants. On note par la suite $\mathbf{X} = (X_i)_{i=1,\dots,n}$ la matrice de taille $n \times p$ regroupant les vecteurs X_i. Soit $C(\mathbf{X}) = \{C_1(\mathbf{X}), ..., C_K(\mathbf{X})\}$ le clustering en K classes obtenues

par une méthode de clustering C sur $\mathbf{X}$. Dans ce travail, on s'intéresse à la question de tester s'il y a une différence entre deux classes $C_k(\mathbf{X})$ et $C_{k'}(\mathbf{X})$. On considère donc l'hypothèse nulle :

$$\mathcal{H}_0^\eta : \eta(C_k(\mathbf{X}), C_{k'}(\mathbf{X}))^T \boldsymbol{\mu} = 0, \tag{1}$$

avec $\boldsymbol{\mu} = (\mu_i)_{i=1,\dots,n} \in \mathbb{R}^{n\times p}$ et le vecteur de contraste $\eta(C_k(\mathbf{X}), C_{k'}(\mathbf{X})) \in \mathbb{R}^n$. Par exemple, pour le test de comparaison des moyennes des classes C_k et $C_{k'}$, le vecteur de contraste s'écrit pour chaque individu i,

$$\eta_i(C_k, C_{k'}) = \left(\frac{\mathbb{1}_{i\in C_k}}{|C_k|} - \frac{\mathbb{1}_{i\in C_{k'}}}{|C_{k'}|} \right) \tag{2}$$

avec $|C_k|$ le cardinal de la classe C_k.

Lorsque le vecteur de contraste η est fixé a priori et ne dépend pas du jeu de données observé, l'hypothèse nulle n'est pas aléatoire et les tests classiques de comparaison de moyennes entre deux classes peuvent être appliqués. Par exemple, la p-valeur $p(\mathbf{x})$ associée au test de Hotelling est donnée par

$$p(\mathbf{x}) = \mathbb{P}_{\mathcal{H}_0^\eta} \left(\|\eta^T \mathbf{X}\|_2 \geq \|\eta^T \mathbf{x}\|_2 \right). \tag{3}$$

Dans le cadre étudié ici, le vecteur $\eta = \eta(C_k(\mathbf{X}), C_{k'}(\mathbf{X}))$ dépend des données, et la procédure de test "naïve" associée à la p-valeur dans (3) ne contrôle plus le risque de première espèce [Gao et al., 2022].

Les méthodes récemment proposées pour répondre à cette problématique de test post-clustering peuvent être classées en deux catégories : les méthodes de partitionnement de l'information et les approches conditionnelles. L'objectif de ce travail est de faire un état de l'art de ces méthodes et de comparer ces méthodes par des simulations numériques.

2 Partitionnement de l'information

L'idée de la première catégorie des méthodes étudiées est de reprendre le principe de partitionnement des données pour les méthodes d'apprentissage supervisé proposé par Cox [1975]. Le but est d'obtenir deux sous-échantillons indépendants, l'un pour construire/ajuster le modèle et l'autre pour la procédure de test. Dans le cas de l'inférence post-clustering, Zhang et al. [2019] ont développé une procédure de test qui utilise le *data splitting* mais cette procédure ne contrôle pas le risque de première espèce. En effet, reporter l'information du clustering (construit à partir du premier sous-échantillon) afin de labéliser les individus du deuxième sous-échantillon avant le test, transpose une information provenant du premier échantillon sur le deuxième. Comme le montre Gao et al. [2022], l'hypothèse testée reste aléatoire du fait du lien entre le vecteur de contraste et les données utilisées pour faire le test.

Afin de contrebalancer le transfert d'information pour la labélisation du jeu de test, Leiner et al. [2023] et Neufeld et al. [2023a] proposent de nouvelles méthodes de séparation des données. Au lieu de partitionner le jeu de données $\mathbf{X}$ en deux sous-ensembles d'observations de taille n_1 et n_2

avec $n_1 + n_2 = n$, ces méthodes partitionnent l'information pour chaque observation, construisant ainsi deux jeux de données indépendants ou conditionnellement indépendants $\mathbf{X}^{(1)}$ et $\mathbf{X}^{(2)}$ de taille n. Le clustering appliqué sur $\mathbf{X}^{(1)}$ permet d'obtenir la classification des n individus, et d'effectuer la procédure de test sur $\mathbf{X}^{(2)}$ indépendant du vecteur de contraste.

Leiner et al. [2023] proposent une procédure appelée *data fission* qui vise à créer deux nouveaux jeux de données $\mathbf{X}^{(1)}$ et $\mathbf{X}^{(2)}$ tels que $\mathbf{X} = h(\mathbf{X}^{(1)}, \mathbf{X}^{(2)})$, où les lois de $\mathbf{X}^{(1)}$ et $\mathbf{X}^{(2)}|\mathbf{X}^{(1)}$ sont connues et $h()$ est une fonction connue. Pour la loi gaussienne multivariée $X_i \sim \mathcal{N}(\mu_i, \Sigma)$, la procédure définit $Z_i \sim \mathcal{N}(0, \Sigma)$ indépendante de X_i, et $X_i^{(1)} = X_i + \tau Z_i \sim \mathcal{N}(\mu_i, (1 + \tau^2)\Sigma)$ et $X_i^{(2)} = X_i - \tau^{-1} Z_i \sim \mathcal{N}(\mu_i, (1 + \tau^{-2})\Sigma)$. Le paramètre de fission $\tau > 0$ permet de définir le partage d'information entre $\mathbf{X}^{(1)}$ et $\mathbf{X}^{(2)}$.

Neufeld et al. [2023a] proposent la procédure de *data thinning* qui, pour certaines lois, permet de décomposer $\mathbf{X}$ en deux jeux de données $\mathbf{X}^{(1)}$ et $\mathbf{X}^{(2)}$ indépendants tels que $\mathbf{X}^{(1)} + \mathbf{X}^{(2)} = \mathbf{X}$. Pour une loi gaussienne multivariée, la procédure propose de générer $X_i^{(1)}|X_i = x_i \sim \mathcal{N}_p(\epsilon x_i, \epsilon(1 - \epsilon)\Sigma)$, où $\epsilon \in [0, 1]$ est un paramètre de partage de l'information, puis $X_i^{(2)} = X_i - X_i^{(1)}$. Ainsi la procédure donne deux jeux de données $\mathbf{X}^{(1)}$ et $\mathbf{X}^{(2)}$ de lois connues et indépendants. Le clustering peut ainsi être obtenu à partir de $\mathbf{X}^{(1)}$ tel que $C_k, C_{k'} \in C(\mathbf{X}^{(1)})$, et $\eta(C_k, C_{k'})$ est considéré comme fixé pour la procédure de test sur $\mathbf{X}^{(2)}$. La loi sous l'hypothèse nulle de la p-valeur

$$p(\mathbf{x}) = \mathbb{P}_{\mathcal{H}_0^\eta}\left(||\eta(\mathbf{x}^{(1)})^T \mathbf{X}^{(2)}||_2 \geq ||\eta(\mathbf{x}^{(1)})^T \mathbf{x}^{(2)}||_2\right) \tag{4}$$

est ainsi connue.

Néanmoins, ces deux méthodes demandent de connaître la vraie matrice de covariance Σ. De plus, le paramètre ϵ (resp. τ), qui pilote le partage d'information, a besoin d'être calibré dans la procédure de *data thinning* (resp. *data fission*).

3 Approche conditionnelle

3.1 Définition d'un test conditionnel

Une seconde famille de méthodes permettant de résoudre ce problème d'inférence post-clustering consiste à prendre en compte explicitement l'action de clustering dans le calcul de la p-valeur. Celle-ci est inspirée de la littérature récente sur l'inférence post-sélection de modèle [Fithian et al., 2014], où il s'agit de prendre en compte une étape de sélection de variables. Par exemple pour le test d'Hotelling considéré dans (3), on souhaite conditionner par l'évènement $\{C_k(\mathbf{x}), C_{k'}(\mathbf{x}) \in C(\mathbf{X})\}$ et définir une p-valeur conditionnelle comme :

$$\begin{aligned}
&\mathbb{P}_{\mathcal{H}_0^\eta}\Big(||\eta(C_k(\mathbf{X}), C_{k'}(\mathbf{X}))^T \mathbf{X}||_2 \geq ||\eta(C_k(\mathbf{X}), C_{k'}(\mathbf{X}))^T \mathbf{x}||_2 \;\Big|\; C_k(\mathbf{x}), C_{k'}(\mathbf{x}) \in C(\mathbf{X})\Big) \\
&= \mathbb{P}_{\mathcal{H}_0^\eta}\Big(||\eta(C_k(\mathbf{x}), C_{k'}(\mathbf{x}))^T \mathbf{X}||_2 \geq ||\eta(C_k(\mathbf{x}), C_{k'}(\mathbf{x}))^T \mathbf{x}||_2 \;\Big|\; C_k(\mathbf{x}), C_{k'}(\mathbf{x}) \in C(\mathbf{X})\Big).
\end{aligned} \tag{5}$$

Ce conditionnement dans (5) permet de fixer le vecteur de contraste, qui ne dépend que du résultat du clustering. Néanmoins cette p-valeur n'étant pas accessible, il faut choisir un conditionnement plus fort donnant accès à la loi de la statistique de test sous l'hypothèse nulle.

Gao et al. [2022] s'intéressent au problème de comparaison de moyenne entre classes (voir équation (2)) sous l'hypothèse $\Sigma = \sigma^2 I_p$. Afin de pouvoir expliciter la p-valeur, ils sur-conditionnent par rapport à (5) en s'appuyant sur la décomposition $\mathbf{X} = \pi_\eta^\perp \mathbf{X} + \frac{\|\eta^T \mathbf{X}\|_2}{\|\eta\|_2^2}\, \eta\, dir(\eta^T \mathbf{X})$ où $\pi_\eta^\perp$ est la projection sur l'orthogonal de la droite engendrée par le vecteur de contraste η. En fixant $\pi_\eta^\perp \mathbf{X}$ et $dir(\eta^T \mathbf{X})$, l'aspect aléatoire dans la décomposition de $\mathbf{X}$ est uniquement porté par la statistique de test $\|\eta^T \mathbf{X}\|_2$. La p-valeur obtenue par ce sur-conditionnement peut alors s'écrire :

$$p(\mathbf{x}, \{C_k, C_{k'}\}) = 1 - \mathbb{F}(\|\eta^T \mathbf{x}\|_2; \sigma\|\eta\|_2, S(\mathbf{x}, \{C_k; C_{k'}\})) \tag{6}$$

où $\mathbb{F}(t; c, S)$ désigne la fonction de répartition de la loi $c.\chi_p$ tronquée à un ensemble S, et

$$S(\mathbf{x}, \{C_k, C_{k'}\}) = \{\phi \geq 0 : C_k, C_{k'} \in C(\tilde{\mathbf{x}}(\phi))\} \tag{7}$$

est l'ensemble des $\phi > 0$ tels que les classes C_k et $C_{k'}$ sont préservées par la procédure de clustering appliquée aux données perturbées $\tilde{\mathbf{x}}(\phi) = \pi_\eta^\perp \mathbf{x} + \frac{\phi}{\|\eta\|_2^2}\, \eta\, dir(\eta^T \mathbf{x})$.

3.2 Mise en pratique de la méthode

Gao et al. [2022] obtiennent une caractérisation explicite de S dans le cas de la classification ascendante hiérarchique pour certaines mesures d'agrégation, et arrivent ainsi à un calcul explicite de la p-valeur associée. Chen and Witten [2023] ont étendu cette procédure dans le cadre de la classification par K-means. Pour réussir à expliciter l'ensemble S, ils s'appuient sur les propriétés des K-means et en conditionnant par un événement plus fort imposant le maintien de la partition des individus à chaque itération de l'algorithme des K-means, appliqué aux données perturbées. Néanmoins, ce sur-conditionnement risque de faire perdre de la puissance au test statistique final. Lorsque cet ensemble S ne peut être explicité, une procédure de Monte-Carlo par Importance Sampling est utilisée pour approcher la p-valeur. Bien que ces résultats aient été obtenus spécifiquement dans le cas de la comparaison de deux classes (voir équation (2)), nous avons montré qu'ils se généralisent sans difficulté à n'importe quel vecteur de contraste ne dépendant que des classes $C_k(\mathbf{X})$ et $C_{k'}(\mathbf{X})$.

3.3 Extensions pour relaxer la variance sphérique connue

Dans un premier temps, Gao et al. [2022] et Chen and Witten [2023] établissent leurs résultats pour une matrice de covariance sphérique connue $\Sigma = \sigma^2 I_p$. Ils proposent ensuite d'étendre les résultats au cas d'une variance générale connue en transformant les données afin de les rendre sphérique. Si la matrice de covariance n'est pas connue, Gao et al. [2022] et Chen and Witten [2023] ont étudié théoriquement l'impact de l'estimation de la variance dans le cas sphérique sur le contrôle du risque de première espèce. Estimer la variance en ignorant la structure de classe conduit à une surestimation de la variance. Gao et al. [2022] ont montré qu'une telle surestimation maintient le contrôle du risque de première espèce, au prix d'une perte de puissance. González-Delgado et al. [2023] ont proposé une généralisation de la définition du test conditionnel de Gao et al. [2022], et ont étendu le résultat de surestimation ci-dessus à une covariance Σ quelconque, et pour toute structure de dépendance entre individus.

Afin de contourner le problème difficile d'estimation de la variance, Yun and Foygel Barber [2023] proposent un test conditionnel dans le cas d'une variance sphérique inconnue. Pour aborder ce problème, ils considèrent une hypothèse nulle plus forte, demandant l'égalité de toutes les vraies moyennes μ_i de tous les individus $i \in C_k(\mathbf{X}) \cup C_{k'}(\mathbf{X})$. Une nouvelle statistique de test est proposée, qui prend en compte la dispersion intra-classe des individus :

$$R(\mathbf{X}) = (|C_k(\mathbf{X})| + |C_{k'}(\mathbf{X})| - 2) \frac{\|\mathcal{P}_0\mathbf{X}\|_F^2}{\|\mathcal{P}_1\mathbf{X}\|_F^2} \tag{8}$$

où $\mathcal{P}_0\mathbf{X}$ capture la différence des moyennes (comme dans Gao et al. [2022]) et $\mathcal{P}_1\mathbf{X}$ capture l'inertie intra-classe entre les deux classes considérées. Dans le même esprit que Gao et al. [2022], $\mathbf{X}$ est décomposé en fonction de $\mathcal{P}_0\mathbf{X}$, $\mathcal{P}_1\mathbf{X}$ et $(I_n - \mathcal{P}_0 - \mathcal{P}_1)\mathbf{X}$. La p-valeur est alors sur-conditionnée en fixant les valeurs des éléments suivants :

$$\|\mathcal{P}_0\mathbf{X}\|_F^2 + \|\mathcal{P}_1\mathbf{X}\|_F^2, \frac{\mathcal{P}_0\mathbf{X}}{\|\mathcal{P}_0\mathbf{X}\|_F}, \frac{\mathcal{P}_1\mathbf{X}}{\|\mathcal{P}_1\mathbf{X}\|_F}, (I_n - \mathcal{P}_0 - \mathcal{P}_1)\mathbf{X}. \tag{9}$$

Ainsi cette p-valeur sur-conditionnée peut être calculée à partir d'un quantile d'une loi de Fisher tronquée à un nouvel ensemble S préservant le clustering sur des données perturbées. Dans le cas d'un clustering à 2 classes, pour les K-means et Classifications Ascendantes Hiérarchiques , le nouvel ensemble S est inclus dans l'ensemble S explicite, proposé par Chen and Witten [2023] et Gao et al. [2022] respectivement. Dans le cas d'un clustering à plus de 2 classes ou une autre méthode de clustering, la p-valeur est estimée par Importance Sampling.

4 Comparaison numérique des performances des méthodes

Un grand nombre de publications récentes sont apparues sur ce sujet [Gao et al., 2022, Chen and Witten, 2023, González-Delgado et al., 2023, Yun and Foygel Barber, 2023, Leiner et al., 2023, Neufeld et al., 2023a, Dharamshi et al., 2023], complété par des publications d'application des méthodes à des problématiques biologiques [Neufeld et al., 2024, 2023b] ou des publications adaptant les méthodes à des tests de comparaison de classe par variable [Hivert et al., 2024, Chen and Gao, 2023].

Dans ce contexte, nous avons souhaiter proposer une comparaison quantitative des différentes méthodes au travers de simulations numériques. Dans un premier temps, l'étude que nous avons menée permet de vérifier si ces méthodes contrôlent bien le risque de première espèce. Une fois les méthodes défaillantes mises de côté, l'étude propose une analyse de la puissance statistique des méthodes toujours en compétition. Il s'agit en particulier de mesurer l'impact sur la puissance statistique et sur le temps de calcul d'une méthode d'inférence conditionnelle avec une expression explicite de la p-valeur, comparée à une estimation par Importance Sampling.

L'objectif de telles expériences numériques est de 1) comprendre et trouver le meilleur compromis entre la puissance statistique et le temps de calcul des méthodes étudiées et 2) prendre du recul sur la possibilité d'étendre le calcul explicite à d'autres méthodes de clustering. En particulier, une perspective naturelle de ce travail est l'étude du clustering par mélanges gaussiens. En effet, cette méthode de clustering permet d'estimer la matrice de covariance de chaque classes, et d'obtenir

une probabilité d'appartenance de chaque individu à une classe. Il serait intéressant d'exploiter ces deux informations dans le cadre de l'inférence post clustering.

References

Lucy L Gao, Jacob Bien, and Daniela Witten. Selective inference for hierarchical clustering. *Journal of the American Statistical Association*, pages 1–11, 2022.

David R Cox. A note on data-splitting for the evaluation of significance levels. *Biometrika*, 62(2): 441–444, 1975.

Jesse M Zhang, Govinda M Kamath, and N Tse David. Valid post-clustering differential analysis for single-cell rna-seq. *Cell systems*, 9(4):383–392, 2019.

James Leiner, Boyan Duan, Larry Wasserman, and Aaditya Ramdas. Data fission: splitting a single data point. *Journal of the American Statistical Association*, pages 1–12, 2023.

Anna Neufeld, Ameer Dharamshi, Lucy L Gao, and Daniela Witten. Data thinning for convolution-closed distributions. *arXiv preprint arXiv:2301.07276*, 2023a.

William Fithian, Dennis Sun, and Jonathan Taylor. Optimal inference after model selection. *arXiv preprint arXiv:1410.2597*, 2014.

Yiqun T Chen and Daniela M Witten. Selective inference for k-means clustering. *Journal of Machine Learning Research*, 24(152):1–41, 2023.

Javier González-Delgado, Juan Cortés, and Pierre Neuvial. Post-clustering inference under dependency. *arXiv preprint arXiv:2310.11822*, 2023.

Young-Joo Yun and Rina Foygel Barber. Selective inference for clustering with unknown variance. *Electronic Journal of Statistics*, 17(2):1923–1946, 2023.

Ameer Dharamshi, Anna Neufeld, Keshav Motwani, Lucy L Gao, Daniela Witten, and Jacob Bien. Generalized data thinning using sufficient statistics. *arXiv preprint arXiv:2303.12931*, 2023.

Anna Neufeld, Lucy L Gao, Joshua Popp, Alexis Battle, and Daniela Witten. Inference after latent variable estimation for single-cell rna sequencing data. *Biostatistics*, 25(1):270–287, 2024.

Anna Neufeld, Joshua Popp, Lucy L Gao, Alexis Battle, and Daniela Witten. Negative binomial count splitting for single-cell rna sequencing data. *arXiv preprint arXiv:2307.12985*, 2023b.

Benjamin Hivert, Denis Agniel, Rodolphe Thiébaut, and Boris P Hejblum. Post-clustering difference testing: valid inference and practical considerations with applications to ecological and biological data. *Computational Statistics & Data Analysis*, page 107916, 2024.

Yiqun T Chen and Lucy L Gao. Testing for a difference in means of a single feature after clustering. *arXiv preprint arXiv:2311.16375*, 2023.