RÉPUBLIQUE FRANÇAISE
Liberté Égalité Fraternité

Université de Toulouse

Doctorat de l'Université de Toulouse

préparé à l'Université Toulouse III - Paul Sabatier

Inférence post-sélection pour l'analyse des données transcriptomiques

---

Thèse présentée et soutenue, le 18 décembre 2024 par

Nicolas ENJALBERT COURRECH

École doctorale
EDMITT - Ecole Doctorale Mathématiques, Informatique et Télécommunications de Toulouse

Spécialité
Mathématiques et Applications

Unité de recherche
IMT : Institut de Mathématiques de Toulouse

Thèse dirigée par
Pierre NEUVIAL et Cathy MAUGIS-RABUSSEAU

Composition du jury
M. Boris HEJBLUM, Rapporteur, Inserm Nouvelle-Aquitaine
M. David CAUSEUR, Rapporteur, Institut Agro Rennes Angers
Mme Béatrice LAURENT-BONNEAU, Examinatrice, INSA Toulouse
Mme Guillemette MAROT, Examinatrice, Université de Lille
M. Pierre NEUVIAL, Directeur de thèse, CNRS Occitanie Ouest
Mme Cathy MAUGIS-RABUSSEAU, Co-directrice de thèse, INSA Toulouse

À mes parents, Christian et Nathalie
À mon frère, Thomas

Remerciements

Tout d'abord, je tiens à exprimer ma profonde gratitude à mes directeurs de thèse, Pierre Neuvial et Cathy Maugis-Rabusseau, pour leur confiance, leur engagement et leur patience tout au long de ces années. Votre soutien indéfectible, vos conseils avisés et votre disponibilité ont été essentiels pour mener à bien ce travail. Je garderai précieusement les enseignements reçus tout au long de ce parcours.

Je remercie également mes rapporteurs, David Causeur et Boris Hejblum, pour le temps consacré à la relecture attentive de ce manuscrit et pour leurs retours constructifs. Un merci tout particulier à Boris pour nos discussions enrichissantes sur l'inférence post-clustering, qui ont permis de mieux cerner les subtilités et les défis de cette problématique. Je souhaite également remercier Béatrice Laurent-Bonneau et Guillemette Marot d'avoir accepté de faire partie de mon jury et de m'avoir donné l'opportunité de présenter ce travail.

Je tiens à exprimer ma reconnaissance à l'ensemble des chercheurs avec qui j'ai eu l'occasion de collaborer à l'Institut de Mathématiques de Toulouse, que ce soit dans le cadre de l'enseignement, de l'organisation de conférences ou au sein de l'école doctorale MITT. Un merci tout particulier à Sébastien et Guillaume, dont l'enthousiasme pour la vulgarisation scientifique a été contagieux.

Ces trois ans de thèse n'auraient pu aussi bien se dérouler sans les amitiés que j'ai pu me créer parmi les doctorants, ATER et post-docs du laboratoire. Soso, je ne te remercierai jamais assez pour ton soutien, nos discussions sans fin et nos. Antho, Alex, nos jetons du Biergarten sont la preuve des bons moments passés ensemble autour d'une bière ou d'un bon jeu de société. Merci pour toutes les fois où vous avez dû répéter les règles parce que je ne vous écoutais pas, oupsi. Jojo, pour toutes nos "petites" discussions de deux heures en fin de journée, merci. Pauline, merci pour ta gentillesse à tout égard, nos papotages qui m'ont permis de relativiser. Biensûr, je veux remercier tous les doctorants que j'ai eu la chance de croiser durant mon parcours : ceux de ma promo avec qui j'ai passé beaucoup de temps, Étienne, Erwanne, Armand ; ceux qui m'ont montré la voie à suivre au labo, Laëtitia, Javi, Clément ; et ceux qui arrivent après moi, Candice, Benjamin, Flo, Angel, bon courage pour la suite. En particulier, je pense à mes co-bureaux Mitja, Jianyu, Anirban, qui m'ont rappelé les subtilités de la grammaire française.

Je ne pourrais pas non plus oublier mes amis, qui m'ont soutenu tout au long de ce parcours. Merci Ma Pa (et Eva) ma jumelle spirituelle, pour tous nos moments passés et à venir. Merci les coupains de la Savoie (ou de la sangria, je ne sais jamais) - Maddie, Loann, Mika, Anne, Gaëlle, Alice et Jérémy - pour toutes nos soirées autour d'un verre, d'une raclette ou d'un time bomb. En remontant plus loin, je pense aux copains du lycée qui me supportent depuis 10 ans sans en découdre. Mariana, merci d'avoir été le soutien indéfectible particulièrement durant ces trois ans. Pierre, Camille, je ne peux rêver de meilleurs amis que vous. Merci pour la confiance que vous me portez et bon vent en Martinique. Cindy, Marine, Inès, les coupines de toujours. Je pense aussi à tous ceux d'Aurignac qui, pour l'instant d'un week-end, une journée ou même une heure, savent me faire évader dans ce coin de paradis.

Enfin, mon parcours de vie a été guidé par les personnes me connaissant depuis toujours : ma famille.Merci à chacun d'entre vous d'avoir essayé de comprendre mon travail, même lorsque ce n'était pas clair pour moi-même. Papa, Maman, merci de m'avoir poussée aussi loin dans mes études et d'avoir toujours cru en moi. Vous m'avez porté jusqu'ici et inconditionnellement encouragé dans tout ce que j'ai entrepris. Jamais je ne pourrais assez vous remercier. Mon frère Toto, Sarah, merci pour tout, tout ce que nous avons vécu depuis toujours, nos vacances et week-ends, nos soirées, nos fous-rires et tout ce qui fait que nous serons toujours inséparables. Ludo, merci pour tout le soutien que tu m'as apporté, la patience que tu as eue, et toutes les attentions que tu as su me porter.

Contents

Chapitre 1

Introduction

1.1 Motivations biologiques des questions statistiques étudiées

1.1.1 Données transcriptomiques : mesure de l'activité des gènes

L'information génique, stockée dans l'ADN, s'exprime à travers la transcription, l'ARN messager jouant le rôle d'intermédiaire dans le réseau d'information. Le développement des puces à ADN dans les années 1990 a eu un impact majeur sur les études génomiques et transcriptomiques (Lowe et al., 2017). Les données transcriptomiques donnent une mesure quantitative du niveau d'activité (ou niveau d'expression) des gènes. Ces technologies ont évolué jusqu'à nos jours, permettant aux chercheurs d'acquérir une meilleure compréhension de l'impact des changements de l'activité transcriptionnelle sur différents tissus, phénotypes ou pathologies. Ces avancées ont, par exemple, permis la collecte de données transcriptomiques dans le cadre du projet américain *the Cancer Genome Atlas Program* (TGCA et al., 2014), qui vise à caractériser les altérations génomiques dans des tissus provenant de patients atteints de divers types de cancer afin d'améliorer leur traitement. Ces données sont utilisées, par exemple, par Thorsson et al. (2018) pour cartographier le paysage immunitaire du cancer, identifiant différentes sous-populations de cellules immunitaires associées à la réponse immunitaire anti-tumorale. Trois grandes évolutions technologiques pour l'obtention de données transcriptomiques sont présentées (Lowe et al., 2017).

Les puces à ADN (ou *microarrays*) sont les premières technologies ayant permis de capturer l'information génique à l'échelle du génome entier (Heller, 2002). Cette capture se fait par fluorescence, permettant d'obtenir un niveau quantitatif d'expression des gènes. Cette expression est traditionnellement modélisée par une distribution gaussienne. Ces données se présentent comme une matrice où une ligne de la matrice représente l'expression de chaque gène pour un individu biologique.

Le séquençage à haut débit, aussi nommé séquençage de nouvelle génération (NGS), constitue une évolution majeure dans les études transcriptomiques, en fournissant un niveau de détail plus poussé que les puces à ADN. Cette technique permet une couverture plus précise du génome ainsi qu'une meilleure détection des gènes faiblement exprimés. Les échantillons d'ARN sont découpés en courtes séquences (action de séquençage), qui sont ensuite alignées et assemblées afin d'obtenir l'expression de génome entier. Les données obtenues fournissent des données de comptages, c'est-à-dire des entiers positifs, qui ont la particularité de présenter une grande hétérogénéité et montrent une inflation de zéros. Cette sur-représentation des zéros est attribuée à plusieurs causes supposées, notamment la non-expression d'un gène pour un individu donné ou un problème technique lors du séquençage. Ces données ne peuvent pas être modélisées par une distribution gaussienne, contrairement aux données issues des puces à ADN. La distribution binomiale négative est donc couramment utilisée pour ces données. Les premières méthodes de séquençage à haut-débit, dites *bulk RNA sequencing* (**bulk RNAseq**), permettent de séquencer un mélange cellulaire d'un échantillon biologique étudié afin d'obtenir une mesure d'expression de chaque gène pour un individu biologique. Ces données se présentent de nouveau comme une matrice où une ligne de la matrice représente l'expression de chaque gène pour un individu biologique. Par la suite, l'évolution technologique

a permis de séquencer séparément chaque cellule d'un échantillon biologique (**scRNAseq** pour *single-cell RNA sequencing*).

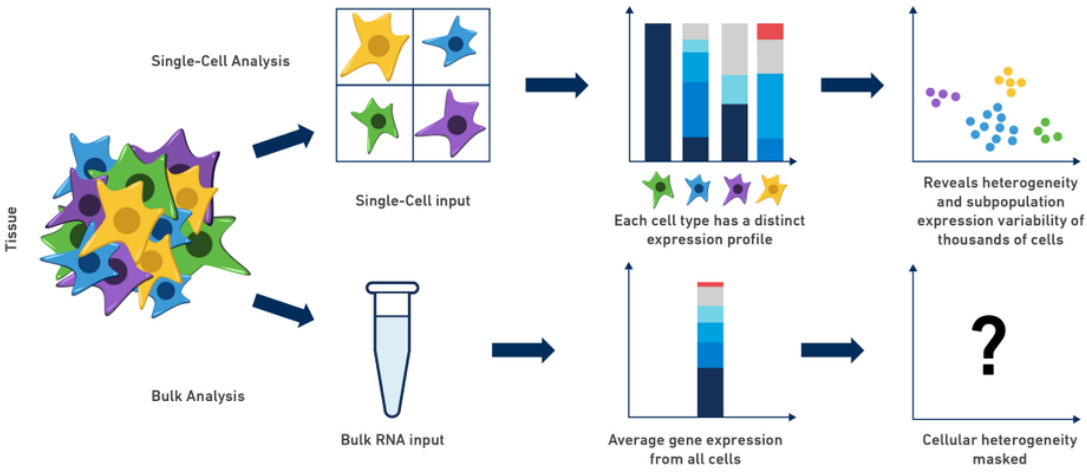

FIGURE 1.1 – Schéma comparant les données bulk RNAseq et single-cell RNAseq. Le séquençage en cellule unique permet de capturer l'hétérogénéité génique d'un tissu, tandis que le bulk RNAseq ne fournit qu'un profil global de ce tissu. (*Source : site web de 10X Genomics*)

Le gain majeur de cette dernière avancée est la granularité supplémentaire de l'expression génique acquise. Cela permet de prendre ainsi en compte l'hétérogénéité des cellules (voir la Figure 1.1 pour la différence entre ces deux types de séquençage). Désormais, pour un individu biologique, une matrice d'expression est obtenue au lieu d'un vecteur (comme dans le cas des données bulk RNAseq). Les lignes de cette matrice correspondent à l'expression génique des cellules. En contrepartie, le séquençage se fait sur une plus petite quantité de matériel biologique (du tissu contenant des milliers de cellules, le séquençage est fait sur une cellule), ce qui diminue le nombre de séquences disponibles par cellule. Cela contribue à exacerber l'inflation des zéros, rendant les données plus éparses (environ 80% des données sont des zéros, principalement en raison de problèmes techniques liés à la détection de l'expression génique).

Ces technologies apportent une diversité de données, mais avec une caractéristique commune : la grande dimension, c'est-à-dire, le nombre de variables (ici les gènes) est plus grand que le nombre d'individus statistique (ici les individus biologiques). Les êtres vivants possèdent un grand nombre de gènes, allant d'environ 13 000 gènes codant pour la mouche, à environ 20 000 pour l'Homme, et jusqu'à environ 100 000 pour le blé. Les expérimentations biologiques ainsi que l'utilisation des techniques de séquençage ont un coût élevé, ne permettant d'avoir que peu de réplicats par condition biologique. Cet effet est exacerbé dans le cadre des données scNRAseq où il est courant de n'avoir qu'un seul réplicat par condition biologique. Les données transcriptomiques obtenues sont donc toutes de grande dimension, étudiant un grand nombre de variables pour un nombre limité d'individus.

1.1.2 Les questions statistiques apportées par les données transcriptomiques

Ce nouveau paradigme des données transcriptomiques a ouvert de nouvelles questions à la frontière de la bio-informatique et de la statistique, telles que la normalisation des données (Bacher and Kendziorski, 2016; Lun et al., 2016; Hafemeister and Satija, 2019; Mortazavi et al., 2008), la gestion des données manquantes (Andrews and Hemberg, 2017; Qiu, 2020), la correction des biais techniques et du plan d'expérience (Haghverdi et al., 2018; Stuart et al.,

2019) ou bien l'analyse différentielle de ces données (Hatfield et al., 2003; Trapnell et al., 2013; Anders and Huber, 2010; Soneson and Robinson, 2018; Korthauer et al., 2016). Cette dernière question est abordée dans ce manuscrit.

1.1.2.1 L'analyse d'expression différentielle pour les données provenant de puces à ADN ou bulk RNAseq.

Plusieurs axes d'analyse peuvent être envisagés dans le cadre de l'analyse des données transcriptomiques. Un exemple de données est une partie des données TCGA étudiant le carcinome urothélial de la vessie [1] (BLCA) (TGCA et al., 2014). Une des questions qui nous intéresse dans cette étude est de déterminer quelles sont les différences d'expression génique entre les stades II et III de ce cancer. Cette compréhension permettra de mieux saisir les évolutions de ce cancer et ainsi de proposer des traitements appropriés aux patients. Les stades II et III sont les deux conditions biologiques comparées dans cette étude. Les deux conditions comportent respectivement 130 et 140 réplicats biologiques (patients atteints de cancer). Cette question vise ainsi à détecter, parmi les 12 534 gènes, ceux qui s'expriment significativement plus dans un stade que dans l'autre. Cette analyse est appelée *Analyse d'expression différentielle*.

La procédure statistique pour répondre à cette question consiste à effectuer un test par gène afin de comparer les moyennes entre les deux conditions biologiques. Les tests usuellement utilisés sont le test de Student (Student, 1908) et son équivalent non paramétrique, le test de la somme des rangs de Wilcoxon (1945). Pour s'adapter aux données de puce à ADN, Smyth (2004) propose la méthode limma, qui a été ensuite adaptée aux données bulk RNAseq par l'extension voom de Law et al. (2014). Pour ces dernières, les méthodes DEseq2 (Love et al., 2014) et edgeR (Robinson et al., 2010) ont été développées pour prendre en compte la nature des données de comptage. Ces trois méthodes font partie des plus utilisées pour l'analyse différentielle sur les données provenant de puces à ADN et de bulk RNAseq (Jeffery et al., 2006; Soneson and Delorenzi, 2013; Conesa et al., 2016).

Autant de tests sont effectués qu'il y a de gènes. Une correction de tests multiples doit être appliquée afin de contrôler le nombre de faux positifs (Dudoit et al., 2008). Ici, cela se définit comme le nombre de gènes détectés à tort comme différentiellement exprimés. Ainsi, les biologistes obtiennent une liste de gènes d'intérêt définis comme différentiellement exprimés. Cependant, les gènes sont souvent regroupés en ensembles ayant une fonction précise (pathways). Une analyse supplémentaire consiste alors à vérifier si certains ensembles de gènes sont particulièrement représentés dans cette liste d'intérêt, ce qui est appelée l'analyse d'enrichissement. Des listes d'ensembles de gènes ont été établies, telles que la *Gene Ontology* (GO) (Gene Ontology, 2015) ou la *Kyoto Encyclopedia of Genes and Genomes* (KEGG) (Kanehisa and Goto, 2000). Deux stratégies d'analyse d'analyse de ces listes coexistent (Ebrahimpoor et al., 2020; Goeman and Bühlmann, 2007). La première, l'analyse d'enrichissement, consiste à identifier, parmi la liste des gènes d'intérêt, les ensembles de gènes qui sont statistiquement sur-représentés. La seconde stratégie, l'analyse d'ensembles de gènes, teste directement la significativité des ensembles de gènes sans avoir besoin d'une liste de gènes d'intérêt prédéfinie.

1.1.2.2 L'identification des gènes marqueurs pour l'analyse de données single-cell RNAseq

La granularité plus fine des données scRNAseq ajoute une dimension d'analyse supplémentaire. De nouvelles questions d'analyse se posent comme l'intégration des conditions biologiques étudiées (Butler et al., 2018; Stuart et al., 2019; Hao et al., 2021) ou bien l'identification

1. Tumeur maligne de la paroi de la vessie, courante, causant environ 150 000 décès par an dans le monde.

des types cellulaires. Ces questions font partie des onze grands challenges de l'analyse des données scRNAseq (Lähnemann et al., 2020). L'outil standard pour l'analyse de ces données est la librairie R `Seurat` (Hao et al., 2024). Les premières étapes du pipeline d'analyse consistent à prétraiter les données, intégrer les différents réplicats pouvant provenir de plusieurs conditions biologiques, normaliser les données, réduire la dimension (usuellement l'Analyse en Composantes Principales (PCA, Pearson, 1901), le t-Distributed Stochastic Neighbor Embedding (t-SNE, Maaten and Hinton, 2008), et l'Uniform Manifold Approximation and Projection (UMAP, Becht et al., 2018)).

Le pipeline d'analyse vise ensuite à identifier les types cellulaires présents dans les tissus étudiés. Cette procédure se découpe en deux étapes. Une étape de *clustering* (classification non supervisée), dans laquelle les cellules sont regroupées pour former des groupes homogènes est suivie d'une étape de test de comparaison entre les groupes de gènes estimés lors du clustering. Cette procédure de test est appelée *inférence post-clustering*. L'analyse se poursuit par une correction des tests multiples, ce qui permet d'obtenir une liste de gènes, dits gènes marqueurs, qui distinguent les groupes cellulaires estimés. L'annotation des types cellulaires se fait en comparant les gènes marqueurs pour un groupe de cellules avec des données connues issues de la littérature scientifique, abordant ainsi une autre question statistique. Dans ce manuscrit, et pour l'analyse des données scRNAseq, nous nous concentrerons sur la question de l'identification des gènes marqueurs.

`Seurat` utilise l'algorithme de Louvain (Blondel et al., 2008) pour obtenir des groupes de cellules, tandis que le test de Wilcoxon est utilisé pour la comparaison entre les groupes de cellules obtenus. À l'issue de cette procédure, même après correction des tests multiples, un grand nombre de gènes apparaissent comme gènes marqueurs. Afin de réduire cette liste, un filtrage est effectué sur des descripteurs statistiques. Comme nous le verrons en détail dans la Section 1.3.3, comparer deux groupes estimés à partir des mêmes données utilisées pour le test invalide l'analyse statistique standard.

Une fois l'analyse effectuée par condition, nous sommes souvent amené à comparer des conditions biologiques. Nous nous plaçons ici dans le même contexte que la Section 1.1.2.1. Dans le cadre des données scRNAseq, les individus biologiques sont constitués de milliers de cellules plus ou moins indépendantes et, surtout, de distributions différentes selon leur type cellulaire. Ainsi, plusieurs stratégies coexistent. Une première stratégie consiste à agréger par réplicat biologique les cellules afin d'obtenir des données proches de celles du bulk RNAseq (nommé pseudo-bulk), permettant ainsi d'utiliser les méthodes standards de l'analyse différentielle pour les données bulk RNAseq. Une deuxième stratégie consiste à considérer les cellules comme des réplicats biologiques et à utiliser des techniques d'inférence développées pour les données bulk RNAseq.

Squair et al. (2021) montrent que ces deux stratégies présentent un nombre élevé de faux positifs, dû en partie au faible nombre de réplicats biologiques, conséquence du coût élevé du séquençage en cellule unique (ce qui affecte particulièrement la première stratégie), mais aussi à l'hétérogénéité de l'expression cellulaire (ce qui affecte davantage la deuxième stratégie). Cette question d'analyse d'expression différentielle fait elle aussi partie des onze grands challenges liés à ces données (Lähnemann et al., 2020). Des méthodes d'analyse ont été développées (Gauthier et al., 2021; Tiberi et al., 2023; Korthauer et al., 2016; Nabavi et al., 2016; Wang and Nabavi, 2018) pour répondre à cette question, en essayant de prendre en compte la hiérarchie des données (les cellules proviennent d'individus, qui sont des réplicats biologiques d'une condition) et la variabilité cellulaire. Cette question d'analyse différentielle pour les données scRNAseq ne sera pas abordée dans ce manuscrit.

Il est important de noter la différence de vocabulaire : les gènes marqueurs sont ceux obtenus en comparant des groupes de cellules (estimés sur le jeu de données) afin d'identifier les types cellulaires. D'un autre côté, les gènes différentiellement exprimés sont ceux obtenus en

comparant des conditions biologiques (définies et connues a priori par l'expérience biologique). Dans ce manuscrit, la question des gènes différentiellement exprimés concerne l'analyse des données provenant des puces à ADN et du bulk RNAseq, tandis que les gènes marqueurs concernent l'analyse des données scRNAseq.

1.2 Tests multiples dans le cadre de l'analyse différentielle

Soit $\mathbf{X}$ une matrice d'expression provenant des puces à ADN ou des données bulk RNAseq de taille $n \times m$, avec m gènes mesurés pour n individus, répartis dans au moins deux groupes correspondant aux conditions biologiques étudiées. Dans le cadre des données BLCA décrites dans la Section 1.1.2.1, le jeu de données se compose de $m = 12\,534$ gènes et $n = 270$ individus répartis en deux groupes (correspondant aux deux conditions biologiques étudiées), avec $n_1 = 130$ et $n_2 = 140$ réplicats biologiques.

1.2.1 Conduire un test statistique

Un test statistique a pour but de déterminer si une hypothèse précise (dite hypothèse nulle, $\mathcal{H}_0$) faite sur une caractéristique d'une population est "compatible" avec les observations faites sur un échantillon tiré de cette population (Cox, 2006). Souvent, $\mathcal{H}_0$ est définie comme l'absence de différence dans le cas de comparaison de moyennes de deux populations. Cette hypothèse s'oppose à une hypothèse alternative $\mathcal{H}_1$ qui est complémentaire et moins précise que $\mathcal{H}_0$ comme par exemple, l'existence d'une différence entre les deux moyennes.

Une statistique de test $\mathcal{T}(\mathbf{X})$ (calculée sur l'échantillon $\mathbf{X} = \{X_1, ..., X_n\}$) est choisie de telle sorte que son comportement soit différent sous $\mathcal{H}_0$ et $\mathcal{H}_1$, et que sa distribution sous l'hypothèse nulle soit connue.

L'étape suivante consiste à calculer la p-valeur, c'est-à-dire la probabilité d'observer une valeur aussi extrême que $\mathcal{T}(\mathbf{x})$ (avec l'échantillon observé $\mathbf{x} = \{x_1, ..., x_n\}$) dans la distribution de $\mathcal{T}(\mathbf{X})$ sous $\mathcal{H}_0$. Ainsi, l'hypothèse nulle est rejetée au profit de l'hypothèse alternative si la p-valeur est inférieure à un certain seuil, appelé niveau de signification (ou risque de première espèce), et généralement noté α.

Par construction, les p-valeurs sous $\mathcal{H}_0$ sont uniformes. Dans ce manuscrit, afin de contrôler la validité d'un test donné, nous comparerons la fonction de répartition empirique (ecdf pour "empirical cumulative distribution function") des p-valeurs obtenues sous $\mathcal{H}_0$ à la fonction de répartition théorique (cdf) de la distribution uniforme.

La Figure 1.2 montre un exemple des ecdf de trois tests. Si l'ecdf des p-valeurs est superposée (courbe bleue) sur la distribution théorique uniforme (ligne noire) alors le test est valide. Si l'ecdf des p-valeurs est au-dessus de la ligne noire alors le test n'est pas valide car les p-valeurs sont stochastiquement plus petites. Dans le cas contraire, si un test fournit des p-valeurs trop grandes, le test reste valide mais est trop conservateur (courbe verte), c'est-à-dire que son niveau effectif est inférieur au niveau de signification cible α.

Au vu de ces définitions, un test "prudent" ne rejetant jamais $\mathcal{H}_0$ est valide. Or ce n'est pas un comportement attendu d'un test afin de conclure sur $\mathcal{H}_0$. La puissance statistique mesure la probabilité de rejeter $\mathcal{H}_0$ quand $\mathcal{H}_1$ est vrai. Empiriquement, nous estimons la puissance statistique par la proportion d'échantillons (simulés sous $\mathcal{H}_1$) dont le test a rejeté $\mathcal{H}_0$, c'est-à-dire la proportion de p-valeurs inférieures au niveau α fixé a priori.

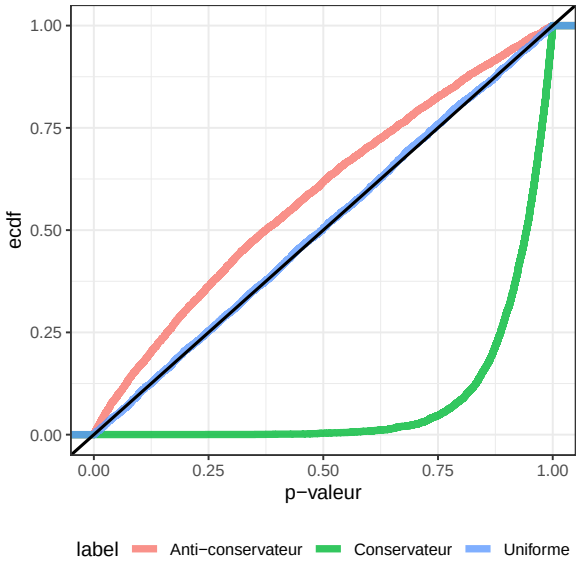

FIGURE 1.2 – Exemple d'évaluation de l'erreur de première espèce : ecdf des p-valeurs provenant d'échantillons simulés sous $\mathcal{H}_0$. La ligne noire représente la fonction de répartition théorique de la distribution uniforme. La courbe bleue représente les tests valident dont les p-valeurs suivent une distribution uniforme. La courbe verte est un exemple de p-valeurs obtenues avec un test conservateur. La courbe rouge représente un exemple de test anti-conservateur. Les p-valeurs sont stochastiquement plus petites que la distribution uniforme ainsi le test n'est pas valide.

1.2.2 Les tests statistiques dans le cadre de l'analyse d'expression différentielle

Dans le cadre de l'analyse d'expression différentielle des données transcriptomiques, la première étape est de faire un test statistique sur la différence d'expression moyenne entre deux groupes, pour chaque gène $j \in [|m|] := \{1, ..., m\}$. L'hypothèse nulle se formule alors comme suit : $\mathcal{H}_0 : \mu_1^{[j]} = \mu_2^{[j]}$ contre l'hypothèse alternative bilatérale $\mathcal{H}_1 : \mu_1^{[j]} \neq \mu_2^{[j]}$, avec $\mu_1^{[j]}$ et $\mu_2^{[j]}$ les moyennes des groupes 1 et 2 pour le gène j.

Plusieurs tests sont utilisés dans le cadre de l'analyse d'expression différentielle. Deux tests non spécifiques aux données transcriptomiques sont couramment employés pour cette problématique. Le test de Welch (1947) est une adaptation du test de Student (1908) de comparaison de moyennes de deux échantillons supposés indépendants et gaussiens, au cas où les variances des deux échantillons sont supposées différentes. Ce test est adapté aux modélisations des données de puces à ADN. Le deuxième test non spécifique aux données transcriptomiques est le test de la somme des rangs de Wilcoxon (1945), également connu sous le nom de test de Mann and Whitney (1947). Il teste si la somme des rangs des deux groupes est égale, permettant ainsi d'évaluer de manière non paramétrique l'égalité des deux distributions. Ce test est non paramétrique, c'est-à-dire que la distribution des données est inconnue (ou non supposée), permettant son application à la fois aux données de puces à ADN et particulièrement aux données bulk RNAseq où l'hypothèse d'une distribution gaussienne n'est pas pertinente.

Le deuxième type de tests est spécifique aux données transcriptomiques (Soneson and Delorenzi, 2013). Dans un premier temps, Smyth (2004) propose la méthode limma pour les données de puces à ADN, disponible dans le package R `limma` (Ritchie et al., 2015). Cette

méthode repose sur un modèle linéaire par gène, et les paramètres sont testés via une statistique bayésienne empirique. Law et al. (2014) ont ensuite proposé la méthode limma-voom, une adaptation aux données bulk RNAseq. Le module voom applique une transformation logarithmique aux données de comptage pour les rendre plus proches d'une distribution gaussienne. De plus, une estimation de la relation moyenne-variance permet de pondérer les tests afin de compenser la surdispersion propre aux données bulk RNAseq. Les deux autres méthodes notables sont celles proposées dans les packages R `DEseq2` (Love et al., 2014) et `edgeR` (Robinson et al., 2010). Ces deux méthodes ont été développées spécifiquement pour les données bulk RNAseq afin de prendre en compte leur nature (données de comptage et sur-dispersion, entre autres). Les deux approches sont similaires en modélisant les données via des distributions binomiales négatives. Dans un premier temps, le paramètre de surdispersion est estimé par un estimateur bayésien pour edgeR et par une modélisation moyenne-variance, proche de celle utilisée dans voom, pour DEseq2. Un modèle linéaire généralisé pour données binomiales négatives est ensuite ajusté. Les coefficients de ce modèle sont testés via le test exact de Fisher (pour edgeR) ou le test de Wald (pour DEseq2) pour la comparaison de deux groupes, et le test du rapport de vraisemblance est utilisé pour des designs plus complexes.

1.2.3 Tests multiples

En sortie de la procédure de test, une p-valeur est obtenue pour chaque gène. Pour l'ensemble des m gènes testés, considérons l'ensemble des p-valeurs $\{p_1, ..., p_m\}$. Définissons les quantités suivantes : soit $\mathscr{H}_0 = \{j \in [|m|], \mathcal{H}_0^{[j]} \text{ est vraie}\}$ l'ensemble des hypothèses nulles vraies, et $\mathscr{H}_1$ son complémentaire, représentant l'ensemble des hypothèses alternatives vraies (ainsi $\mathscr{H}_0 \cap \mathscr{H}_1 = \emptyset$ et $|\mathscr{H}_0 \cup \mathscr{H}_1| = m$). Rappelons que le but de l'analyse d'expression différentielle est d'obtenir une liste de gènes dits différentiellement exprimés (gènes **DE**), c'est-à-dire pour lesquels il existe une différence significative entre les deux groupes comparés. Pour un ensemble de gènes S donné, la valeur $\mathrm{FP}(S) := |S \cap \mathscr{H}_0|$ correspond alors au nombre (inconnu) de faux positifs dans l'ensemble S, c'est-à-dire le nombre de gènes rejetés à tort. Une solution naïve à ce problème serait de déclarer comme différentiellement exprimés les gènes dont les p-valeurs sont inférieures au niveau de significativité α, c'est-à-dire $R := \{j \in [|m|], p_j < \alpha\}$. Il est bien connu dans la littérature qu'une telle procédure ne contrôle pas le nombre de faux positifs (Tukey, 1953; Holm, 1979). L'erreur de type I, qui est contrôlée dans le cadre d'un test unique, n'a pas de sens dans le contexte des tests multiples. Ainsi, d'autres risques ont été définis dans le cadre des tests multiples.

Contrôle du Family-Wise Error Rate (FWER). Une mesure de risque naturelle dans le cadre des tests multiples est le contrôle du *Family-Wise Error Rate* (FWER) (Tukey, 1953). Ce risque est défini comme la probabilité d'avoir au moins un faux rejet :

$$\mathrm{FWER}(R) = \mathbb{P}(\mathrm{FP}(R) > 0) \tag{1.1}$$

Une procédure de test multiple produisant R contrôle le FWER si $\mathrm{FWER}(R) \leq \alpha$. Un exemple de procédure qui contrôle le FWER est celle de Bonferroni (Dunn, 1961), définie par

$$R^{\mathrm{Bf}} = \left\{ j \in [|m|], p_j < \frac{\alpha}{m} \right\}. \tag{1.2}$$

Cette procédure fournit un ensemble de rejets R^{Bf} qui contrôle le FWER, tel que :

$$\mathrm{FWER}(R^{\mathrm{Bf}}) = \mathbb{P}\left(\mathrm{FP}(R^{\mathrm{Bf}}) > 0\right) = \mathbb{P}\left(\exists j \in \mathscr{H}_0, p_j \leq \frac{\alpha}{m}\right) \leq \frac{|\mathscr{H}_0|}{m}\alpha \leq \alpha. \tag{1.3}$$

Contrôle du False Discovery Rate (FDR). Benjamini and Hochberg (1995) ont proposé une autre mesure de risque : le taux de faux positifs, ou *False Discovery Rate* (FDR). Cette mesure est particulièrement pertinente dans le contexte génomique, où la correction de Bonferroni est trop conservatrice. Le FDR d'un ensemble de rejets R est défini par

$$\mathrm{FDR}(R) = \mathbb{E}\left(\mathrm{FDP}(R)\right) \tag{1.4}$$

$$\text{avec } \mathrm{FDP}(R) = \frac{\mathrm{FP}(R)}{|R| \vee 1} \tag{1.5}$$

où le FDP (*False Discoverie Proportion*) correspond à la proportion de faux positifs de l'ensemble R et où le symbole $\vee$ représente le maximum entre deux quantités. Une procédure de test multiple qui contrôle le FDR est une procédure qui fournit un ensemble R tel que $\mathrm{FDR}(R) \leq \alpha$. Une procédure classique pour contrôler le FDR est celle de Benjamini and Hochberg (1995, BH). Soient $p_{(1)} \leq ... \leq p_{(m)}$ les p-valeurs ordonnées. La procédure BH définit l'ensemble de rejet

$$R^{\mathrm{BH}} = \left\{ j \in [|m|], p_{(j)} < \frac{j^{\mathrm{BH}}(\alpha)}{m}\alpha \right\}, \tag{1.6}$$

$$\text{avec } j^{\mathrm{BH}}(\alpha) = \max\left\{ j' \in [|m|], p_{(j')} < \frac{j'}{m}\alpha \right\}.$$

Dépendance entre les p-valeurs. Benjamini and Hochberg (1995) ont montré que leur procédure contrôle le FDR uniquement dans le cas d'indépendance entre les p-valeurs. En pratique, cette hypothèse n'est pas forcément vraie. Pour un type de dépendance positive appelé PRDS (*Positive Regression Dependence on a Subset*), Benjamini and Yekutieli (2001) ont montré que la procédure BH contrôle toujours le FDR. Bien que la validité de l'hypothèse PRDS pour les études d'expression différentielle n'ait pas été formellement prouvée, elle est généralement acceptée comme une hypothèse raisonnable dans ce domaine et dans les études génomiques en général (Goeman and Solari, 2014). Toutefois, l'utilisation pratique et l'interprétation du FDR en génomique se heurtent à deux obstacles majeurs.

Obstacle I : le FDR d'un sous-ensemble de gènes différentiellement exprimés n'est pas contrôlé. Supposons que nous ayons obtenu une liste R de gènes DE par une procédure de contrôle FDR appliquée au niveau q. Comme indiqué par Goeman and Solari (2011), la déclaration $\mathrm{FDR}(R) \leq q$ s'applique uniquement à la liste R, et aucune autre déduction statistiquement valide ne peut généralement être faite sur d'autres listes de gènes. Toutefois, une pratique courante consiste à modifier manuellement cette liste en ajoutant ou en retirant des gènes, sur la base d'une connaissance externe ou a priori (telle que la connaissance des ensembles de gènes, voir Section 1.1.2.1). Un exemple typique est celui des volcano plots (Cui and Churchill, 2003). Le volcano plot est une représentation graphique couramment utilisée pour les résultats d'expression différentielle dans le cadre d'études génomiques. Chaque gène est représenté par un point correspondant aux coordonnées suivantes :

- une estimation de la "taille d'effet" (ou bien du *fold change*) qui est généralement quantifiée par la différence entre les expressions géniques moyennes (à l'échelle logarithmique) des deux groupes comparés. Plus la valeur du log fold-change est grande en valeur absolue, plus il y a une différence constatée entre les deux groupes ;
- une mesure de la signification pour le test d'expression différentielle utilisé. En général, la p-valeur est l'indicateur utilisé, transformé en échelle $-\log_{10}$ afin de mettre en avant les p-valeurs faibles (dont le test semble significativement exprimer une différence entre les groupes).

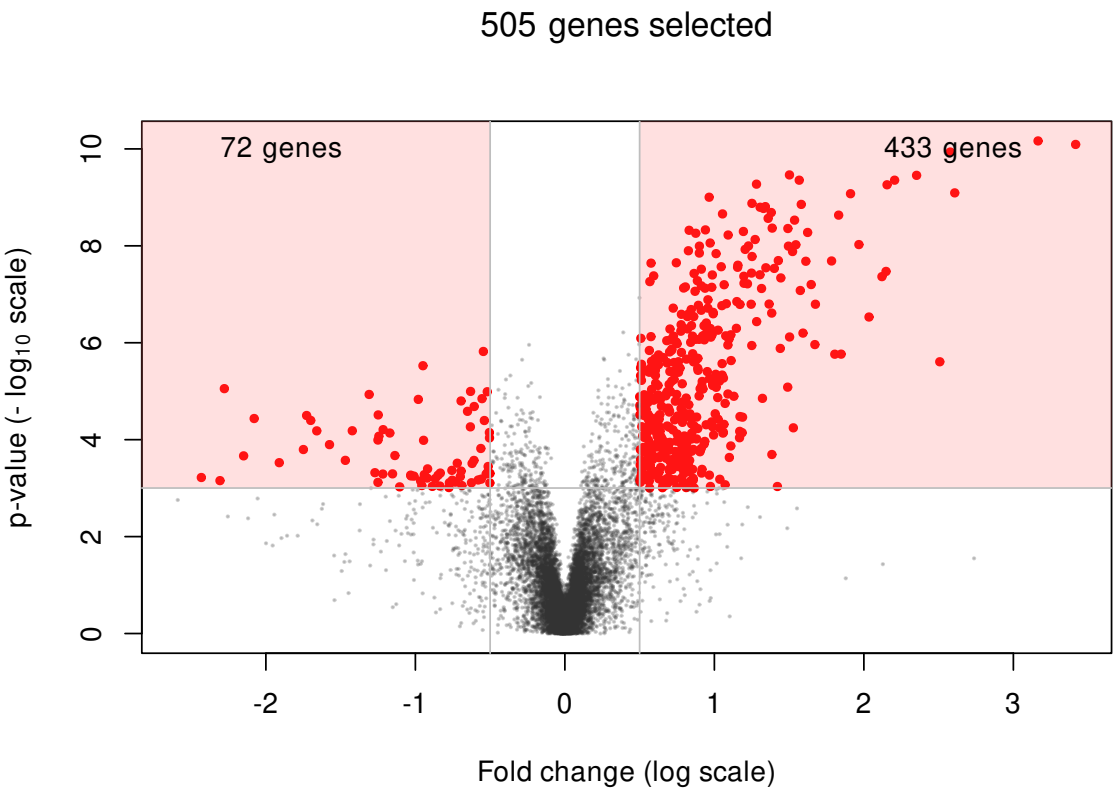

FIGURE 1.3 – Volcano plot représentant les résultats d'une analyse d'expression différentielle obtenue sur des données décrites en Section 1.1.2.1. Pour chaque gène, l'axe des abscisses représente le fold change (en échelle logarithmique), et l'axe des ordonnées, les p-valeurs associées au test de Wilcoxon en échelle $-\log_{10}$. Les points situés en haut du plot (zone rose et points rouges) avec une forte amplitude de fold change et une faible p-valeur sont considérés comme les gènes les plus pertinents pour l'étude. La zone est caractérisée par des seuils tels que la valeur absolue du log fold change est supérieure à 0.5 et les p-valeurs sont inférieures à 0.001. Cette sélection fournit 505 gènes.

La Figure 1.3 montre un exemple de volcano plot utilisé dans une étude d'expression différentielle sur des données transcriptomiques décrites en Section 1.1.2.1. Ainsi, les gènes considérés comme différentiellement exprimés se retrouvent dans les extrémités hautes du graphique. Les biologistes sélectionnent généralement un ensemble de gènes en appliquant un seuil non seulement sur les p-valeurs, mais aussi sur la taille d'effet associé (zone en rose sur la Figure 1.3). Les méthodes de contrôle du FDR offrent des garanties sur la sélection faite uniquement en seuillant les p-valeurs, mais ne fournissent pas de garanties pour le sous-ensemble sélectionné par seuillage sur la taille d'effet. Ebrahimpoor and Goeman (2021) ont montré dans une étude de simulation approfondie que ce type de stratégie de double filtrage produit des taux de faux positifs trop élevés.

Obstacle II : le contrôle du FDR ne correspond pas au contrôle du FDP. L'affirmation $\mathrm{FDR}(R) \leq q$ est souvent interprétée à tort comme signifiant que "la proportion de fausses découvertes (FDP) dans R est inférieure à q". En réalité, la proportion de fausses découvertes est une quantité *aléatoire*, et le FDR correspond à son espérance : $\mathrm{FDR} = \mathbb{E}(\mathrm{FDP})$. Informellement, $\mathrm{FDR}(R) \leq q$ implique seulement que la *moyenne des FDP sur des réplica-*

tions hypothétiques de la même expérience génomique et de la même procédure de seuillage sur les p-valeurs est majorée par q. Cette distinction n'aurait pas beaucoup d'importance si les expressions géniques étaient statistiquement indépendantes : en effet, lorsque le nombre m de tests tend vers l'infini, le FDP se concentre autour du FDR correspondant, avec un taux de convergence paramétrique typique : $m^{-1/2}$ (Neuvial, 2008). Toutefois, à mesure que la dépendance augmente, la distribution du FDP devient fortement asymétrique et à queue lourde, comme indiqué dans Korn et al. (2004), et illustré plus en détail dans Neuvial (2020, Fig. 2.1).

1.2.4 Inférence post hoc : contrôle de la Proportion de Faux Positifs

La notion d'inférence post hoc a été introduite par Goeman and Solari (2011) pour répondre à ces limitations. S'appuyant sur les travaux antérieurs de Genovese and Wasserman (2006), Goeman and Solari (2011) ont obtenu des bornes de confiance pour le FDP dans *des sous-ensembles d'hypothèses arbitraires, multiples et éventuellement guidés par des données* en utilisant la théorie du *closed testing* (Marcus et al., 1976). En pratique, *les bornes post hoc de Simes* sont recommandées par Goeman et al. (2019), car elles sont valides sous l'hypothèse PRDS et peuvent être calculées efficacement. Les bornes post hoc de Simes ont récemment été popularisées en génomique par Ebrahimpoor and Goeman (2021), mais aussi dans les études de neuro-imagerie par Rosenblatt et al. (2018), où cette approche a été appelée "All-resolutions inference" (ARI).

Malgré leurs propriétés théoriques très intéressantes, les méthodes post hoc ne sont pas encore largement connues et utilisées pour traiter les situations de tests multiples en génomique, où le contrôle de la FDR via la procédure BH reste la norme. Deux raisons possibles expliquent cette situation : contrairement à la procédure BH pour le contrôle de la FDR, la limite post hoc de Simes pour l'inférence post hoc est (i) typiquement conservatrice dans les applications génomiques et (ii) sa construction basée sur les méthodes de *closed testing* peut être difficile à comprendre pour les praticiens.

1.2.5 Méthodes de Simes adaptatives à la dépendance

Une autre construction de bornes post hoc a été proposée dans Blanchard et al. (2020) et explorée dans Durand et al. (2020) et Blanchard et al. (2021). Cette stratégie permet d'obtenir des bornes moins conservatrices, en s'adaptant à la dépendance statistique entre les tests à l'aide de permutations, et à la rareté du signal à l'aide d'un principe de réduction progressive. Cette section vise à détailler cette méthode.

1.2.5.1 Inférence post hoc basée sur l'interpolation

Objectifs : obtenir des bornes post hoc. Pour un sous-ensemble donné S de gènes déclarés DE, avec $s = |S|$, notre objectif est de trouver une fonction $\overline{\mathrm{FP}}_\alpha$ telle que, avec grande probabilité, $\overline{\mathrm{FP}}_\alpha(S)$ soit supérieure au nombre de faux positifs $\overline{\mathrm{FP}}_\alpha(S)$ dans S :

$$\mathbb{P}\left(\forall S, \mathrm{FP}(S) \leq \overline{\mathrm{FP}}_\alpha(S)\right) \geq 1 - \alpha. \tag{1.7}$$

En suivant Goeman and Solari (2011), une fonction $\overline{\mathrm{FP}}_\alpha$ satisfaisant l'Équation (1.7) sera appelée une *borne (supérieure) post hoc du nombre de faux positifs au niveau* α. L'inférence post hoc peut être formulée de manière équivalente en termes de bornes supérieures sur le FDP : $\overline{\mathrm{FDP}}_\alpha(S) = \overline{\mathrm{FP}}_\alpha(S)/s$, ou en termes de bornes inférieures sur le nombre ou la proportion de vrais positifs : $\overline{\mathrm{TP}}_\alpha(S) = s - \overline{\mathrm{FP}}_\alpha(S)$, $\overline{\mathrm{TDP}}_\alpha(S) = \overline{\mathrm{TP}}_\alpha(S)/s$.

Stratégie : Contrôle du JER et interpolation. Les bornes étudiées reposent sur un risque de tests multiples appelé *Joint Error Rate* (JER). Étant donnée une famille croissante de seuils $\mathbf{t} = (t_k)_{k\in[|K|]}$,

$$\mathrm{JER}(\mathbf{t}) = \mathbb{P}\left(\exists k \in [|K|] \,:\, q_k < t_k\right), \tag{1.8}$$

où pour $k \in [|K|]$, q_k désigne la k-ième plus petite p-valeur parmi l'ensemble des gènes véritablement non-DE (vraies hypothèses nulles, $\mathscr{H}_0$). Une famille croissante de seuils $(t_k)_{k\in[|K|]}$ contrôle le JER au niveau α si, avec une probabilité supérieure à $1-\alpha$,

$$\mathbb{P}\left(\forall\, k \in [|K|] \,:\, q_k \geq t_k\right) \geq 1-\alpha\,. \tag{1.9}$$

Un résultat clé est que toute famille $\mathbf{t}$ telle que $\mathrm{JER}(\mathbf{t}) \leq \alpha$ fournit une borne post hoc associée au niveau α via l'argument d'interpolation suivant.

Proposition 1 (Borne post hoc basée sur l'interpolation (Blanchard et al., 2020, Proposition 2.3))**.** *Si $\mathbf{t} = (t_k)_{k\in[|K|]}$ contrôle le JER au niveau α, alors l'Équation (1.7) est satisfaite pour la borne*

$$\overline{\mathrm{FP}}_\alpha(S) = \min_{k\in[|K|]} \left\{ \sum_{j\in S} \mathbb{1}_{\{p_j \geq t_k\}} + k - 1 \right\}. \tag{1.10}$$

Pour l'exhaustivité et afin de souligner la simplicité de l'argument, une preuve de la Proposition 1 est donnée dans l'Annexe A.1.1.

Bornes post hoc de Simes Un exemple important est la famille Simes $\mathbf{t}^{\mathrm{S}}(\alpha)$, définie par $t_k^S(\alpha) = \alpha k/m$ pour tout $k \in [|m|]$ (ici la famille $\mathbf{t}^{\mathrm{S}}(\alpha)$ est de la taille $K = m$). L'inégalité de Simes (1986) assure que $\mathrm{JER}(\mathbf{t}^{\mathrm{S}}(\alpha)) \leq \alpha$ dès que la famille de p-valeurs est PRDS (Sarkar et al., 2008). Comme noté par Blanchard et al. (2020), la borne post hoc alors obtenue par la Proposition 1 coïncide avec la borne post hoc de Simes introduite dans Goeman and Solari (2011).

Bien que l'inégalité de Simes soit précise lorsque les p-valeurs sont indépendantes, elle devient de plus en plus conservatrice à mesure que la dépendance entre les tests se renforce (Blanchard et al., 2020, Table 1). Le contrôle du JER associé et la borne post hoc héritent naturellement de ce conservatisme (ce point sera illustré dans les expériences numériques des Sections 2.4 et 2.5). Pour aborder cette question de conservatisme, il est utile de noter que pour $\lambda > 0$, le JER de la famille Simes $\mathbf{t}^S(\lambda)$ peut être écrit comme

$$\mathrm{JER}(\mathbf{t}^S(\lambda)) = \mathbb{P}\left(\min_{k\in[|m|]} \frac{mq_k}{k} < \lambda \right). \tag{1.11}$$

À la lumière de l'Équation (1.11), une idée naturelle pour obtenir un contrôle précis du JER est de sélectionner le λ le plus grand tel que $\mathrm{JER}(\mathbf{t}^S(\lambda)) \leq \alpha$. Cette idée est la base de la méthode de calibrage décrite en Section 1.2.5.2.

1.2.5.2 Calibrage du JER par permutation

Le JER défini dans l'Équation (1.8) dépend uniquement de la distribution conjointe des p-valeurs des vraies hypothèses nulles. Bien que cette distribution soit inconnue en pratique, dans les études DE à deux groupes, elle peut être approchée en permutant les étiquettes des groupes. En conséquence, la première étape de notre méthode de calibrage consiste à construire une matrice P de taille $B \times m$ contenant les p-valeurs des permutations : P_{bj} est la p-valeur du test du gène j associée à la b-ième permutation des étiquettes de groupe des

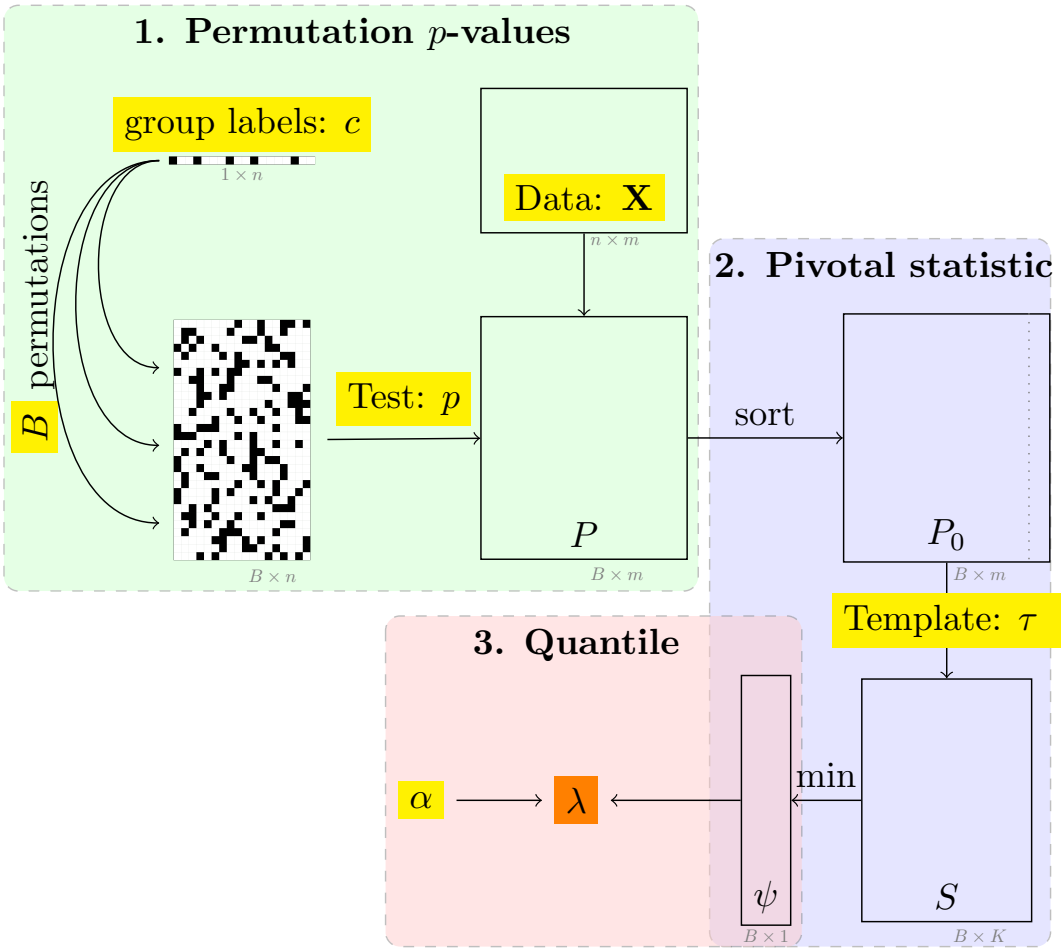

FIGURE 1.4 – Illustration des trois principales étapes du calibrage du JER basé sur les permutations. Le résultat est mis en évidence en orange, et les entrées sont mises en évidence en jaune. Les données d'entrée sont sous forme d'une matrice d'expression génique $n \times m$ $\mathbf{X}$ et d'un vecteur binaire c d'étiquettes de groupe (de taille n), spécifiant à quels groupes appartiennent les observations à comparer. Les paramètres sont le niveau de JER cible α et le nombre B de permutations, la fonction de p-valeur pour effectuer le test et le template τ (équivalent à la famille de seuils $\mathbf{t}$ tronquée ou complétée à la taille s).

échantillons. Ceci est illustré dans la première partie (carré vert, "1. Permutation p-values") de la Figure 1.4.

Les étapes suivantes du calibrage se comprennent plus facilement dans le cas particulier de la famille Simes. En effet, d'après l'Équation (1.11), $\mathrm{JER}(\mathbf{t}^S(\lambda))$ est la valeur de la fonction de répartition cumulative de $\psi = \min_{k\in[|m|]} mq_k/k$ en λ. En conséquence, la méthode de calibrage procède par le calcul de B échantillons de la "statistique pivotale" ψ, et le résultat est le quantile d'ordre α de ces statistiques. P_0 est la matrice de permutation où les p-valeurs sont triées par permutation. La méthode décrite dans la Figure 1.4 couvre non seulement le cas de la famille de Simes, mais aussi toute famille $\tau(\lambda) = (\tau_k(\lambda))_{k\in[|K|]}$ où les τ_k sont des fonctions inversibles.

Validité. Le Théorème 1 dans Blanchard et al. (2021) garantit que cette méthode de calibrage fournit bien λ tel que $\mathrm{JER}(\lambda) \leq \alpha$, pour les tests dont la p-valeur pour un gène donné dépend uniquement des valeurs d'expression de ce gène. Plus généralement, la théorie

développée dans Blanchard et al. (2020) est valide dès que la distribution conjointe des statistiques de test satisfait une hypothèse de randomisation (Romano and Wolf, 2005; Hemerik and Goeman, 2018). Dans le cas ci-dessus des tests à deux échantillons, cela est obtenu par permutation des étiquettes de groupe. Comme noté dans Blanchard et al. (2020), cette hypothèse est également valide pour les tests à un échantillon, où les permutations à l'Étape 1 sont remplacées par des inversions de signe.

1.2.6 Contributions sur les méthodes post hoc

Les contributions de cette thèse dans le domaine de l'inférence post hoc sont présentées dans la partie I. Le travail se concentre sur l'application de la méthode développée par Blanchard et al. (2020). Une nouvelle version de l'algorithme de calcul des bornes post hoc permet d'obtenir une complexité linéaire en temps de calcul. Cette amélioration permet de gagner en efficacité, notamment dans le cadre de l'application aux données transcriptomiques de comparaison de deux groupes. Afin de comprendre les enjeux de cette méthode pour ces données, une modélisation et un exemple d'application de la méthode sont présentés sur des données bulk RNAseq et des puces à ADN. Les performances statistiques de cette méthode sont évaluées sur des données transcriptomiques et comparées aux autres méthodes d'inférence post hoc. Ces contributions ont fait l'objet d'un article (Enjalbert-Courrech and Neuvial, 2022), retranscrit dans le Chapitre 2. La deuxième contribution pour l'inférence post hoc décrite dans ce manuscrit est le développement d'une interface interactive de la méthode précédemment décrite. Cette interface, nommée IIDEA (*Interactive Inference for Differential Expression Analyses*), permet de rendre accessible la méthode pour l'analyse d'expression différentielle sur des données bulk RNAseq et provenant de puces à ADN. En particulier, l'application permet d'obtenir des garanties sur les bornes post hoc (décrites dans la Section 1.2.5.1) pour des ensembles S sélectionnés sur un volcano plot (voir Section 1.2.3 pour une description de ce type de graphique) simultanément. Cette application est accessible via une interface web ainsi qu'un package R. L'ensemble de cette contribution est détaillé dans le Chapitre 3. Le Chapitre 4 discute des perspectives sur l'application des méthodes post hoc. Le Chapitre A contient les annexes dédiées à cette partie.

1.3 Inférence post-clustering pour la détection de gène marqueurs

Dans le cadre des données single-cell RNAseq, considérons un seul individu biologique dont l'expression génique est retranscrite pour un ensemble de cellules contenues dans un tissu. Soit $\mathbf{X}$ la matrice d'expression de taille $n \times m$, où n représente le nombre de cellules séquencées et m le nombre de gènes. La question posée dans ce cas est d'identifier les types cellulaires présents dans le tissu étudié. Cela revient à (i) regrouper les cellules de manière homogène et (ii) effectuer un test entre deux groupes obtenus à l'étape (i).

1.3.1 Procédures de clustering

L'étape (i) peut être résolue par une méthode de clustering (ou classification non supervisée). L'objectif d'une telle méthode est de fournir des groupes (ou classes) d'individus homogènes à partir de caractéristiques observées. Soit $\mathcal{C}$ une méthode de clustering et $\mathcal{C}(\mathbf{X}) := \{C_1, ..., C_K\}$ la partition en K classes obtenues où chaque individu est dans une seule classe, soit $C_k \cap \mathcal{C}_{k'} = \emptyset,\ \forall\ k, k' \in [|K|], k \neq k'$. Il existe de nombreuses méthodes de clustering, nous en décrivons trois types par la suite.

La première catégorie de méthodes regroupe celles basées sur la minimisation d'un critère de distance entre individus (ou entre classes). Dans cette catégorie, la méthode des K-means (Steinhaus et al., 1956; Macqueen, 1967) et celle de la Classification Ascendante Hiérarchique (HAC, Ward Jr, 1963) font partie des plus connues et utilisées.

La deuxième catégorie de méthodes regroupe celles basées sur des modélisations probabilistes. La méthode basée sur les modèles de mélange (McLachlan, 2000) illustre bien cette catégorie. Le principe est d'estimer la densité inconnue des données par une distribution de mélanges, qui tient compte d'une structure sous-jacente en classe, puis d'en déduire un clustering des données. Ce type de méthode apporte ainsi des garanties statistiques sur la partition obtenue. Les modèles de mélange gaussiens sont détaillés dans la Section 7.1.

Une troisième catégorie de méthodes est basée sur les graphes. Les individus représentent les noeuds d'un graphe et les relations entre individus par des arêtes. L'objectif est de découvrir des structures ou des communautés de nœuds fortement interconnectés, correspondant à des classes d'individus ayant des caractéristiques similaires. L'algorithme de Louvain (Blondel et al., 2008) est un exemple de ce type de méthode de clustering, et est utilisé par le package R `Seurat` pour l'analyse des données scRNAseq.

Estimer le nombre de classes K est un sujet de recherche à part entière dans le domaine du clustering (Thorndike, 1953; Rousseeuw, 1987). Nous n'abordons pas ce problème dans cette thèse, le nombre de classes est supposé connu.

1.3.2 Test statistique pour l'identification des gènes marqueurs

L'étape (ii) de test statistique reprend ceux présentés dans la Section 1.2.1. À nouveau, nous nous concentrerons ici sur la comparaison des moyennes de deux groupes G_k et $G_{k'}$ pour chaque variable $j \in [|m|]$. Notons que, dans le cadre de l'identification des gènes marqueurs, la comparaison peut être le groupe G_k contre les autres, mais nous n'aborderons pas ce cas-là dans ce manuscrit. Comme dans la Section 1.2.1, l'hypothèse nulle est $\mathcal{H}_0^{[j]} : \mu_{G_k}^{[j]} = \mu_{G_{k'}}^{[j]}$, où $\mu_{G_k}^{[j]}$ et $\mu_{G_{k'}}^{[j]}$ sont respectivement les moyennes marginales des groupes G_k et $G_{k'}$ pour la variable j.

Dans notre cas d'étude, les groupes sont estimés sur $\mathbf{X}$ par une méthode de clustering, tels que $G_k := C_k(\mathbf{X})$ et $G_{k'} := C_{k'}(\mathbf{X})$. Pour répondre à cette hypothèse, en pratique $\mathbf{X}$ est utilisé pour calculer la statistique de test et obtenir une p-valeur. Pour l'identification des gènes marqueurs, le package `Seurat` propose, entre autres, d'utiliser le test non paramétrique de Wilcoxon (voir Section 1.2.2) dont les p-valeurs obtenues sont corrigées par la méthode de Bonferroni (voir Section 1.2.3), permettant ainsi de définir la liste de gènes marqueurs comme étant ceux ayant une p-valeur ajustée inférieure à α. Malgré cette correction, les p-valeurs obtenues semblent encore plus faibles que ce qui serait attendu (Ioannidis, 2005; Zhang et al., 2019). Pour contourner ce problème, `Seurat` propose de filtrer la liste des gènes marqueurs en appliquant des seuils sur des caractéristiques provenant des données, comme le log fold change, la proportion de cellules qui s'exprime dans l'une des deux classes, ou encore l'aire sous la courbe ROC en entraînant un classifieur supervisé sur le gène testé pour estimer la séparabilité des deux classes (utilisées comme label). Une grande quantité de p-valeurs faibles, provenant du test de comparaison de deux classes, peut être le signe d'une procédure de test non valide. Nous nous concentrons donc dans cette partie sur l'étude des procédures de test comprenant l'étape (i) de clustering et (ii) d'inférence basée sur les résultats de l'étape (i), dites d'*inférence post-clustering*.

1.3.3 La problématique du *double dipping*

Formellement, la procédure de test décrite en Section 1.3.2 revient à définir l'hypothèse nulle

$$\mathcal{H}_0^{[j]} : \mu^{[j]}_{C_k(\mathbf{X})} = \mu^{[j]}_{C_{k'}(\mathbf{X})}. \tag{1.12}$$

Cette hypothèse, basée sur $C_k(\mathbf{X})$ et $C_{k'}(\mathbf{X})$, est considérée comme fixée, dans le test d'identification des gènes marqueurs, alors que les quantités $C_k(\mathbf{X})$ et $C_{k'}(\mathbf{X})$ sont aléatoires, dépendant de $\mathbf{X}$. Théoriquement, l'hypothèse nulle définie dans la Section 1.2.1 doit être fixée a priori, c'est-à-dire avant d'avoir observé les données. Ainsi, considérer les classes fixées (à tort) permet de se replacer dans ce cadre théorique et d'obtenir une p-valeur afin de conclure sur l'hypothèse nulle. L'exemple simulé suivant montre les effets d'une telle procédure.

Soit un échantillon gaussien centré et réduit de $n = 500$ individus indépendants et identiquement distribués (voir la Figure 1.5-A). Un clustering hiérarchique est appliqué aux données en utilisant la distance euclidienne et le lien d'agrégation de Ward pour obtenir un clustering en $K = 2$ classes (alors qu'en réalité, il n'y a qu'une seule classe dans les données). L'application d'un clustering sur des données sans signal crée des classes artificiellement séparés (voir la Figure 1.5-B). Un test de Student est alors réalisé afin de comparer la moyenne des deux classes obtenues. 2000 expériences sont effectuées pour évaluer les performances statistiques du test. La Figure 1.5-C représente la fonction de répartition empirique des p-valeurs obtenues avec le test de Student. Ici, il est largement visible que les p-valeurs sont stochastiquement plus petites que la distribution uniforme. Le test de Student ne contrôle pas le risque de première espèce lorsque les hypothèses de test sont basées à tort sur le résultat du clustering fixé. Gao et al. (2024) et Hivert et al. (2024a) montrent également cet effet numériquement.

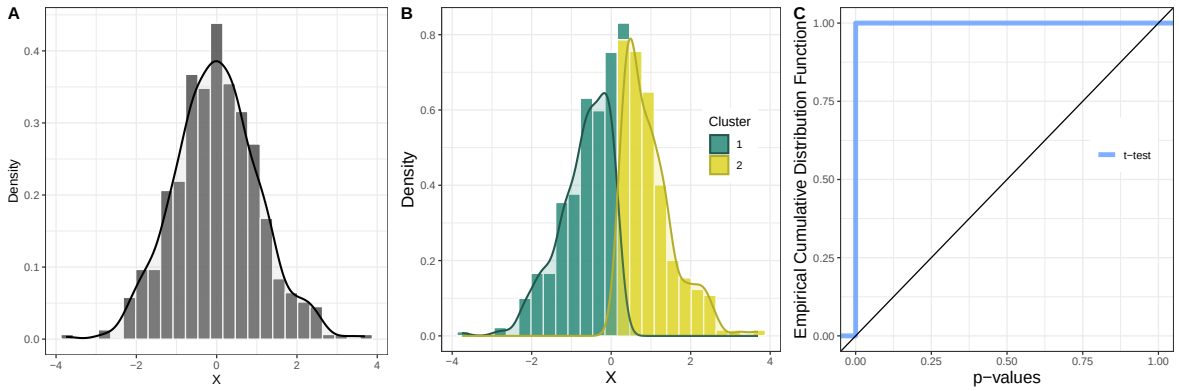

FIGURE 1.5 – Le clustering sur des données gaussiennes peut créer artificiellement une différence entre deux classes. **A** (gauche) : Distribution des données générées selon 500 individus d'une distribution gaussienne avec une moyenne de 0 et une variance de 1. **B** (centre) : Distribution des données en fonction de deux classes obtenues par classification hiérarchique avec lien de Ward. **C** (droite) : Distribution des p-valeurs issues du test de Student naïf (2000 simulations), comparées à une distribution uniforme théorique pour le contrôle du taux d'erreur de type I.

Considérer à tort que l'hypothèse nulle de l'équation (1.12) est fixée ne permet donc pas d'avoir une procédure de test d'inférence post-clustering valide. Ainsi, il semble que les procédures de test pour l'identification des gènes marqueurs ne contrôlent pas le nombre de faux positifs.

Kriegeskorte et al. (2009) aborde cette problématique d'analyse circulaire (ou la problématique de *double dipping*) où les données sont utilisées à la fois pour définir l'hypothèse

nulle et conclure dessus. Notre problématique se place dans un sous-cas de l'analyse circulaire où l'hypothèse nulle est définie après avoir effectué une étape de clustering sur les données. Formellement, l'hypothèse nulle testée $\mathcal{H}_0(\mathcal{C}(\mathbf{X}))$ dépend du clustering $\mathcal{C}(\mathbf{X})$, dont l'Équation (1.12) en est un exemple pour la comparaison des moyennes de deux classes. La statistique de test répondant à cette hypothèse nulle est $\mathcal{T}(\mathcal{C}(\mathbf{X}), \mathbf{X})$, qui dépend de $\mathbf{X}$ au travers du clustering et du test choisi. La p-valeur associée est une fonction du quantile

$$\mathbb{P}_{\mathcal{H}_0(\mathcal{C}(\mathbf{X}))}\left(\mathcal{T}(\mathcal{C}(\mathbf{X}), \mathbf{X}) \leq \mathcal{T}(\mathcal{C}(\mathbf{X}), \mathbf{x})\right).$$

En pratique, cette p-valeur n'est pas calculable en raison du caractère aléatoire du clustering $\mathcal{C}(\mathbf{X})$.

1.3.4 Les solutions d'inférence post-clustering

Cette question de *double dipping*, dans le cadre de l'inférence post clustering, a fait l'objet de plusieurs publications au cours des **cinq** dernières années. Deux types de méthodes se démarquent dans cet état de l'art.

Méthodes basées sur la partition d'information. Ce type de méthodes cherche à obtenir deux matrices indépendantes $\mathbf{X}^{(1)}$ et $\mathbf{X}^{(2)}$ à partir de $\mathbf{X}$ afin de calculer le test statistique indépendemment de l'estimation du clustering. Une première solution envisagée par Zhang et al. (2019) est d'utiliser la méthode de *data splitting* (Cox, 1975), généralement utilisée en apprentissage supervisé. Cette procédure vise à découper $\mathbf{X}$ en deux jeux de données indépendants, $\mathbf{X}^{(1)}$ et $\mathbf{X}^{(2)}$, respectivement de n_1 et n_2 observations, tels que $n_1 + n_2 = n$. Le modèle supervisé est appris sur $\mathbf{X}^{(1)}$ et testé sur $\mathbf{X}^{(2)}$. Les deux jeux de données étant indépendants, le modèle ne sur-apprend pas et permet de généraliser la règle de décision inférée.

Dans le cadre de l'inférence post-clustering, le clustering est estimé sur $\mathbf{X}^{(1)}$ et l'inférence est faite sur $\mathbf{X}^{(2)}$, comme décrit dans l'Algorithme 1. Pour faire le test, les individus de

Algorithm 1 Algorithme de *data splitting* utilisé dans le cadre de l'inférence post-clustering

1: Partager $\mathbf{X}$ pour obtenir $\mathbf{X}^{(1)} \in \mathbb{R}^{n_1 \times m}$ et $\mathbf{X}^{(2)} \in \mathbb{R}^{n_2 \times m}$
2: Estimer le clustering sur $\mathbf{X}^{(1)}$: $\mathcal{C}(\mathbf{X}^{(1)})$
3: Entraîner un modèle de classification supervisée où les labels sont les classes obtenus par le clustering $\mathcal{C}(\mathbf{X}^{(1)})$: $\hat{f}_{\mathbf{X}^{(1)}}$
4: Prédire les labels des observation de $\mathbf{X}^{(2)}$ avec $\hat{\mathcal{C}}_{\mathbf{X}^{(1)}}(\mathbf{X}^{(2)}) := \hat{f}_{\mathbf{X}^{(1)}}(\mathbf{X}^{(2)})$
5: Faire le test en utilisant la statistique de test $\mathcal{T}\left(\hat{\mathcal{C}}_{\mathbf{X}^{(1)}}(\mathbf{X}^{(2)}), \mathbf{X}^{(2)}\right)$

$\mathbf{X}^{(2)}$ sont labellisés d'après le clustering fait sur $\mathbf{X}^{(1)}$. Le clustering prend des informations provenant de $\mathbf{X}^{(1)}$ et les applique sur les informations $\mathbf{X}^{(2)}$ qui servent à faire le test, ce qui donne à nouveau un problème de *double dipping*. Cet effet a été montré par Gao et al. (2024) et Neufeld et al. (2024b). Un exemple unidimensionnel est repris dans la Figure 1.6. Pour cette expérience, un échantillon de $n = 500$ individus indépendants a été généré à partir d'une distribution gaussienne centrée et réduite. Le splitting est effectué avec $n_1 = n_2 = 250$. Le clustering HAC est appliqué sur $\mathbf{X}^{(1)}$ avec la méthode de Ward. La méthode des 5 plus proches voisins est utilisée pour apprendre les labels et partitionner $\mathbf{X}^{(2)}$. Le test de Student permet ensuite de faire l'inférence. La Figure 1.6-B illustre qu'une procédure de test basée sur le data splitting ne contrôle pas le risque de première espèce. La problématique soulevée par cette méthode est que les deux jeux de données $\mathbf{X}^{(1)}$ et $\mathbf{X}^{(2)}$ ne comportent pas les mêmes individus. La statistique de test $\mathcal{T}\left(\hat{\mathcal{C}}_{\mathbf{X}^{(1)}}(\mathbf{X}^{(2)}), \mathbf{X}^{(2)}\right)$ utilise $\mathbf{X}^{(2)}$ deux fois (problématique du *double dipping*), considérant $\hat{\mathcal{C}}_{\mathbf{X}^{(1)}}(\cdot)$ comme une nouvelle procédure de clustering.

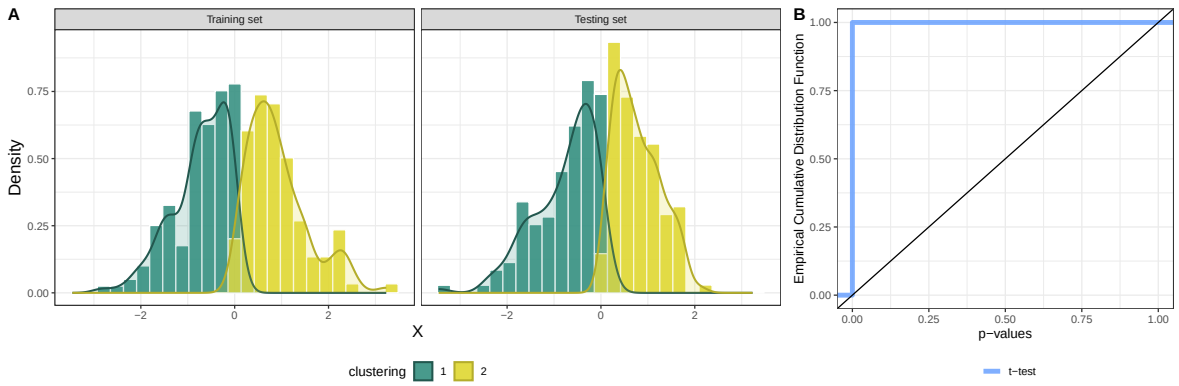

FIGURE 1.6 – Le data splitting n'est pas une procédure d'inférence post-clustering valide. **A** (gauche) : Distribution des données splittées (en *training set* et *testing set*). Le clustering est estimé sur le jeu d'entraînement et reporté sur le jeux de test par apprentissage supervisé (méthode des k plus proches voisins). **B** (droite) : Distribution des p-valeurs issues du test de Student naïf (2000 simulations), comparées à une distribution uniforme théorique pour le contrôle du taux d'erreur de type I.

Neufeld et al. (2024a), Dharamshi et al. (2024) et Leiner et al. (2023) proposent de nouvelles méthodes de partage de l'information permettant d'obtenir deux jeux de données $\mathbf{X}^{(1)}$ et $\mathbf{X}^{(2)}$ de même taille $n \times m$. Cela permet ainsi de reporter facilement le clustering $\mathcal{C}(\mathbf{X}^{(1)})$ sur $\mathbf{X}^{(2)}$ et de n'utiliser $\mathbf{X}^{(2)}$ uniquement pour l'étape d'inférence. Ces méthodes sont détaillées dans la Section 5.2.1. Neufeld et al. (2023) et Neufeld et al. (2024b) ont appliqué ces méthodes aux données scRNAseq dans le cadre de plusieurs problématiques non supervisées telles que celles du clustering.

Approches conditionnelles. Ce type de méthodes incorpore explicitement l'action de clustering dans le calcul de la p-valeur. Ces approches sont inspirées de la littérature récente de l'inférence post-sélection (Fithian et al., 2014), ayant déjà de nombreuses applications comme les statistiques en grande dimension (Lee et al., 2016), le machine learning et les réseaux de neurones profonds (Duy et al., 2022; Chen and Andrews, 2023), ou bien l'inférence causale en grande dimension (Belloni et al., 2017).

Dans les analyses classiques de statistique, le modèle statistique M est préalablement fixé avant d'avoir observé les données. L'hypothèse nulle ici ne dépend pas des données $\mathbf{X}$. Le test au niveau α associé au modèle M et à l'hypothèse nulle $\mathcal{H}_0$ contrôle le risque de première espèce

$$\mathbb{P}_{M,\mathcal{H}_0}(\text{reject } \mathcal{H}_0) \leq \alpha.$$

Dans le contexte de l'inférence post-sélection, le modèle testé $M(\mathbf{X})$ est défini en regardant les données, sur lesquelles l'hypothèse nulle $\mathcal{H}_0(M(\mathbf{X}))$ est construite. Fithian et al. (2014) montrent que, conditionnellement à la sélection du modèle M, le test contrôle l'erreur de première espèce

$$\mathbb{P}_{M(\mathbf{X}),\mathcal{H}_0}(\text{reject } \mathcal{H}_0 \mid (M(\mathbf{X}), \mathcal{H}_0) \text{ selected}) \leq \alpha.$$

Dans le cadre de l'inférence post-clustering, la sélection du modèle revient à faire un clustering sur les données, tel que $M(\mathbf{X}) := \mathcal{C}(\mathbf{X})$. Gao et al. (2024) basent leur développement sur ces outils en fournissant une p-valeur conditionnelle dans le cadre de données gaussiennes à covariance sphérique commune (les variables ont la même variance et sont indépendantes) pour

la comparaison multivariée de deux classes. En particulier, ils décrivent le calcul explicite de la p-valeur pour un clustering obtenu par une procédure HAC. Des extensions ont été faites, notamment pour l'obtention d'une p-valeur explicite pour une procédure K-means (Chen and Witten, 2023), pour des données gaussiennes avec une matrice de covariance sphérique inconnue (Yun and Foygel Barber, 2023), ou bien une matrice de covariance non sphérique (González-Delgado et al., 2023). Chen and Gao (2023) et Hivert et al. (2024a) ont adapté le test proposé par Gao et al. (2024) pour répondre aux questions de comparaison de moyennes marginales et se rapprocher de la question des gènes marqueurs. Bachoc et al. (2023) proposent une méthode conditionnelle pour le clustering convexe (Pelckmans et al., 2005), en se basant sur la méthode d'inférence post-sélection pour les modèles de régression lasso (Tibshirani, 1996) développée par Lee et al. (2016).

1.3.5 Contributions sur l'inférence post-clustering

Les contributions de cette thèse sur le sujet de l'inférence post-clustering sont détaillées dans la Partie II de ce manuscrit.

La littérature scientifique très récente (moins de cinq ans) de solution d'inférence post-clustering aborde deux questions distinctes. La première problématique cherche à comparer la moyenne multivariée de deux classes. La deuxième question est de comparer marginalement les moyennes de deux classes. Dans les deux cas, les méthodes développées dans la littérature ont été comparées à la solution naïve (utilisée dans les simulations de la Figure 1.5) mais pas aux autres méthodes du domaine. La première contribution est une revue de l'ensemble de ces méthodes en comparant leur cadre théorique pour pointer leurs avantages et inconvénients théoriques. Une analyse numérique comparative des méthodes permet d'affiner et d'illustrer les performances statistiques de ces méthodes, ainsi que d'identifier et d'étaudier leurs limitations. Les comparaisons numériques sont faites à partir de données gaussiennes simulées afin d'être en adéquation avec les cadres théoriques utilisés dans la plupart des méthodes. Les Chapitres 5 et 6 contiennent ces contributions pour respectivement la comparaison multivariée (question 1) et la comparaison marginale (question 2).

Le Chapitre 7 est consacré à l'exploration d'extensions de la méthode de Gao et al. (2024) en utilisant le clustering provenant d'un modèle de mélange gaussien. Ce type de clustering fournit des informations supplémentaires (estimation des paramètres du mélange, probabilité a posteriori d'appartenance des individus aux classes, ...) qui peuvent permettre le raffinement de la méthode. Le Chapitre 8 discute des perspectives d'études sur ce sujet concernant les extensions envisagées pour ces méthodes et leur analyse numérique.

Part I

Post hoc inference

Chapter 2

Powerful and interpretable error control for two-group differential expression studies

This chapter is adapted from the published article:
Enjalbert-Courrech, N. and Neuvial, P. (2022). Powerful and interpretable control of false discoveries in two- group differential expression studies. *Bioinformatics*, 38(23):5214–5221.

2.1 Introduction

Two-sample comparison problems are ubiquitous in genomics. The usual example is that of differential expression studies aimed at identifying genes whose mean expression level differs significantly between two populations. A common strategy is to test, for each gene, the null hypothesis that its mean expression is identical in both populations. *Differentially expressed* (DE) genes are those that pass a threshold on p-values after correction for multiple testing (name adjusted p-values).

The state-of-the-art method for large-scale multiple testing is false discovery rate (FDR) control, introduced by Benjamini and Hochberg (1995). The FDR is the expected proportion of false positives among the selected genes. The most widely used method to control FDR is the Benjamini-Hochberg (BH) procedure, which has been shown to control FDR when the hypotheses corresponding to the non-differentially expressed genes are independent or satisfy a specific type of positive dependence called PRDS (Positive Regression Dependency Structure) (Benjamini and Yekutieli, 2001). PRDS is widely accepted as a reasonable assumption in differential gene expression studies and in genomic studies in general (see, e.g. Goeman and Solari, 2014).

However, two major limitations remain:

1. The FDR of a subset of genes is not controlled. The list of genes identified as differentially expressed can be modified manually, leading to an inflation of the false discovery rate, as noted by Goeman and Solari (2011). A typical example is the case of volcano plots (Cui and Churchill, 2003), where genes are selected according to a threshold on p-values and a threshold on the fold change (difference of average gene expression on the log scale), see Figure 2.2. Ebrahimpoor and Goeman (2021) have recently shown in an extensive simulation study that this type of double filtering strategy yields inflated false discovery rates.
2. FDR control does not mean FDP control (proportion of false positives), which is a *random quantity*. The fact is that FDR is the *average FDP over hypothetical replications*. If gene expressions are independent, then the FDP concentrates on the corresponding FDR with a parametric convergence rate $m^{-1/2}$ (Neuvial, 2008), with m the number of tested variables. In the presence of dependencies between tests, the FDP can greatly

vary, despite a controlled FDR on average, as reported by Korn et al. (2004), and illustrated in Neuvial (2020, Fig. 2.1).

To overcome these limitations, the notion of post hoc inference was introduced by Goeman and Solari (2011), based on Genovese and Wasserman (2006). This method provides confidence bounds for FDP in *arbitrary, multiple and possibly data-driven subsets of hypothesis* using the theory of closed testing (Marcus et al., 1976). Goeman et al. (2019) recommend using *Simes post hoc bounds* in practice, as they are valid under PRDS assumption. Ebrahimpoor and Goeman (2021) have popularized these bounds in genomics studies and Rosenblatt et al. (2018) in neuroimaging studies, where this approach has been called *All-Resolutions Inference* (ARI). Despite their theoretical advantages, these methods are not yet widely adopted in genomics, where FDR control via the Benjamini-Hochberg procedure remains the norm.

An alternative construction of post hoc bounds has been proposed by Blanchard et al. (2020) and further explored by Blanchard et al. (2021) and Durand et al. (2020). This strategy can yield sharper bounds by adapting to the statistical dependency between tests using permutations and to the sparsity of the signal using a step-down principle. The main goal of the present chapter is to popularize the use of the post hoc bounds introduced by Blanchard et al. (2020) called *Adaptive Simes* in the context of two-group DE studies. The mathematical framework of this method is detailed in Section 2.2 with a short and self-contained introduction to interpolation-based post hoc inference (Section 2.2.1) and the use of the permutation-based calibration methods introduced by Blanchard et al. (2020) for DE studies (Section 2.2.2). Accordingly, the main contributions of this chapter can be summarized as follows:

1. Proving that generic interpolation-based post hoc bounds can be computed in linear time (Section 2.3);
2. Applying the resulting *Adaptive Simes* method to a specific RNAseq DE study to illustrate that it yields more interpretable results than those derived from FDR control and sharper bounds than Simes post hoc bounds (Section 2.4);
3. Assessing the statistical performance of the method (control of the target risk, and statistical power) for DE studies via comprehensive numerical experiments based on real genomic data, both for microarray and bulk RNAseq data sets (Section 2.5).

Altogether, the results presented in this chapter illustrate that substantial gains in power can be achieved with respect to state-of-the-art post hoc bounds as of 2022 in the case of two-group DE studies, without sacrificing computational efficiency.

These developments are implemented in the R package `sanssouci` [1] (Neuvial et al., 2024). The R code used for the numerical experiments is available from the associated GitHub repository [2]. The Supplementary data (texts, figures, algorithms) are provided in Appendix A.

2.2 Background: Adaptive Simes methods to dependence

This section revisits elements discussed in detail in Section 1.2.

2.2.1 Interpolation-based post hoc inference

We consider a *differential expression* (DE) study with m features. These features are called genes for simplicity, but the methods described below are also applicable more generally. Let $\mathbf{X}$ be the expression matrix of size $n \times m$, with n the number of samples divided into two

1. https://sanssouci-org.github.io/sanssouci/
2. https://github.com/sanssouci-org/IIDEA-method-paper

groups (corresponding to two biological conditions) of sizes n_1 and n_2 (with $n_1+n_2=n$). The null hypothesis for gene j tests the equality of the means of the two groups being compared. Formally, let $\mu_1^{[j]}$ and $\mu_2^{[j]}$ be the expression means of groups 1 and 2, respectively. The null hypothesis associated with gene j is written as $\mathcal{H}_0^{[j]} : \mu_1^{[j]} = \mu_2^{[j]}$, and the corresponding p-value is denoted by p_j. m statistical tests are computed, one per gene, and the corresponding vector of p-values is denoted by $(p_1, \ldots, p_m)$. Differentially expressed (DE) genes are usually defined as those whose p-value is below a significance threshold, which is obtained by a multiple testing procedure. For now, we only assume that a p-value is available to test each gene's differential expression. More specific assumptions on how these p-values are obtained are given in Section 2.2.2.

2.2.1.1 Objective: post hoc bounds

For a given subset S of genes declared DE, with $s = |S|$, we denote by $\mathrm{FP}(S)$ the number of false positives in S, that is, the number of genes in S that are not truly DE. Our goal is to find a function $\overline{\mathrm{FP}}_\alpha$ such that with high probability, $\overline{\mathrm{FP}}_\alpha(S)$ is larger than the number of false positives in S:

$$\mathbb{P}\left(\forall S, \mathrm{FP}(S) \leq \overline{\mathrm{FP}}_\alpha(S)\right) \geq 1-\alpha\,. \tag{2.1}$$

Following Goeman and Solari (2011), a function $\overline{\mathrm{FP}}_\alpha$ satisfying Equation (2.1) will be called an α-level *post hoc upper bound on the number of false positives*. Post hoc inference can be equivalently formulated in terms of upper bounds on the FDP: $\overline{\mathrm{FDP}}_\alpha(S) = \overline{\mathrm{FP}}_\alpha(S)/s$, or in terms of lower bounds on the number or proportion of true positives: $\overline{\mathrm{TP}}_\alpha(S) = s - \overline{\mathrm{FP}}_\alpha(S)$, $\overline{\mathrm{TDP}}_\alpha(S) = \overline{\mathrm{TP}}_\alpha(S)/s$.

2.2.1.2 Strategy: JER control and interpolation

The bounds studied rely on a multiple-testing risk called the Joint Error Rate (JER) and introduced by Blanchard et al. (2020). Given a non-decreasing family of thresholds $\mathbf{t} = (t_k)_{k\in[|K|]}$,

$$\mathrm{JER}(\mathbf{t}) = \mathbb{P}\left(\exists k \in [|K|] \,:\, q_k < t_k\right)\,, \tag{2.2}$$

where for $k \in [|K|]$, q_k denotes the k-th smallest p-value among the set of truly non-DE genes (true null hypotheses). A key result is that any family $\mathbf{t}$ such that $\mathrm{JER}(\mathbf{t}) \leq \alpha$ yields an associated α-level post hoc bound by the following interpolation argument.

Proposition 2 (Interpolation-based post hoc bound (Blanchard et al., 2020, Proposition 2.3))**.** *If $\mathbf{t} = (t_k)_{1\leq k\leq K}$ controls the JER at level α, then Equation (2.1) is satisfied for the bound*

$$\overline{\mathrm{FP}}_\alpha(S) = \min_{1\leq k\leq K}\left\{\sum_{j\in S}\mathbb{1}_{\{p_j\geq t_k\}} + k - 1\right\}\,. \tag{2.3}$$

For completeness and to emphasize the simplicity of the argument, a proof of Proposition 2 is given in Appendix A.1.1.

2.2.1.3 Simes post hoc bounds

An important example is the Simes family $\mathbf{t}^{\mathrm{S}}(\alpha)$, defined by $t_k^S(\alpha) = \alpha k/m$ for all k. The Simes (1986)'s inequality ensures that $\mathrm{JER}(\mathbf{t}^{\mathrm{S}}(\alpha)) \leq \alpha$ as soon as the p-value family is PRDS (Sarkar et al., 2008). As noted by Blanchard et al. (2020), the post hoc bound then derived

by Proposition 2 coincides with the Simes post hoc bound introduced in Goeman and Solari (2011).

Although the Simes inequality is sharp when the p-values are independent, it is increasingly conservative as the dependence between tests gets stronger (Blanchard et al., 2020, Table 1). The associated JER control and post hoc bound naturally inherit this conservativeness (as illustrated in the numerical experiments of Sections 2.4 and 2.5). In order to address this conservativeness issue, it is useful to note that for $\lambda > 0$, the JER of the Simes family $\mathbf{t}^S(\lambda)$ can be written as

$$\mathrm{JER}(\mathbf{t}^S(\lambda)) = \mathbb{P}\left(\min_{1\leq k\leq m} \frac{mq_k}{k} < \lambda\right). \tag{2.4}$$

In view of Equation (2.4), a natural idea in order to obtain a tight JER control is to select the largest λ such that $\mathrm{JER}(\mathbf{t}^S(\lambda)) \leq \alpha$. This idea is the basis of the calibration method described in Section 2.2.2.

2.2.2 JER calibration by permutation

The JER defined in Equation (2.2) only depends on the joint p-value distribution of true null hypotheses. Although this distribution is unknown in practice, in two-group DE studies, it can be approximated by permuting the group labels. Accordingly, the first step of our calibration method is to build a $B \times m$ matrix P of permutation p-values: P_{bj} is the p-value of the test of gene j associated to the b-th permutation of the group labels of samples. This is illustrated in the first panel of Figure 1.4.

The next steps of the calibration are best explained in the particular case of the Simes family. Indeed, by Equation (2.4), $\mathrm{JER}(\mathbf{t}^S(\lambda))$ is the value of the cumulative distribution function of $\psi = \min_{1\leq k\leq m} mq_k/k$ at λ. Accordingly, the calibration method proceeds by calculating B samples from the "pivotal statistic" ψ, and the output is the quantile of order α of these statistics. P_0 is the permutation matrix where p-values are sorted by permutation.

The method as described in Figure 1.4 covers not only the case of the Simes family, but any family $\tau(\lambda) = (\tau_k(\lambda))_{k\in[|K|]}$ where the τ_k are invertible functions. We have also implemented a "step-down" version: it is a slightly more powerful algorithm that is also adaptive to the unknown proportion of true null hypotheses (Blanchard et al., 2020, Proposition 4.5). Algrithm 3 transcribes the calibration illustrated in Figure 1.4.

Validity. Theorem 1 in Blanchard et al. (2021) ensures that this calibration method yields $\mathrm{JER}(\lambda) \leq \alpha$, for tests whose p-value for a given gene depends on the data only via its own expression values. In particular, this is the case for two-sample Student tests or Wilcoxon rank sum tests, which can be used for microarray and bulk RNA sequencing (RNAseq) DE studies, respectively. However, note that this permutation-based strategy is formally only valid for two-group comparisons with no adjustment factors. More generally, the theory developed in Blanchard et al. (2020) is valid as soon as the joint distribution of the test statistics satisfies a randomization assumption (Romano and Wolf, 2005; Hemerik and Goeman, 2018). In the above case of two-sample tests, this is obtained via permutation of group labels. As noted in Blanchard et al. (2020), this assumption also holds for one-sample tests, where permutations at Step 1 are replaced by sign-flips. It also holds when testing for the marginal independence between each gene's expression and a continuous outcome via a correlation test, as further explained and illustrated in Appendix A.6.

Let us recall that our methods rely on *group label permutations* in order to obtain statistically valid procedures that adapt to the dependency between genes observed in the data set at hand. While being theoretically valid regardless of the sample size n, such permutation techniques require a large enough sample size to learn gene dependencies appropriately. As a

rule of thumb, we advocate the use of this method in studies with more than 5 samples per group.

Complexity. Assuming a linear time complexity $O(n)$ to perform the test of one single null hypothesis, the overall time complexity of the calibration method is $O(mB(n+\log(m)))$. Indeed, the most costly step is the calculation of P_0, which involves mB tests followed by B sorting operations on a vector of size m. The overall space complexity is $O(m(B+n))$.

Figure 1.4 also illustrates the modularity of Algorithm 3, where the three main steps are highlighted in different colors. This modularity is important in practice. For example, it makes it possible to obtain the result for several values of α without re-computing the permutation matrix P_0. This modularity is also useful for the computational efficiency of the above-mentioned step-down version of the calibration algorithm. The calibration method has been available since 2017 within the R package `sanssouci`.

2.3 Linear time interpolation-based post hoc bound

Post hoc bounds can be used for multiple gene selections S without compromising the corresponding error control. For post hoc inference to be applicable in practice $\overline{\mathrm{FP}}_\alpha(S)$ must be computed efficiently.

A naive implementation of the bound $\overline{\mathrm{FP}}_\alpha(S)$ defined in Equation (2.3) would require s^2 operations by performing a loop on both $k \in [|s|]$ and $j \in S$ in order to calculate $v_k(S) = \sum_{j\in S} \mathbb{1}_{\{p_j \geq t_k\}} + k - 1$ for all k. This induces a quadratic worst-case time complexity $O(m^2)$, which is achieved when evaluating $\overline{\mathrm{FP}}_\alpha$ on the set of all genes. A quadratic time complexity for a single set is too slow for DE studies with $m \geq 10{,}000$. Moreover, a useful application of post hoc bounds is to build the false positive confidence curve associated with S, that is, all the bounds $\overline{\mathrm{FP}}_\alpha(S_j)$ for $j \in [|s|]$, where S_j is the index set of the j smallest p-values in S. Using the above naive algorithm, this would require $O(s^3)$ operations, implying a cubic worst-case time complexity $O(m^3)$ to build the false positive confidence curve associated with all hypotheses.

In contrast, Algorithm 2 computes $\overline{\mathrm{FP}}_\alpha(S)$ in linear time and space $O(s)$ for a given S. Inputs of this algorithm are the set sorted p-values $p = (p_{(1:S)}, \dots p_{(s:S)})$ where $p_{(j:S)}$ is the j-th smallest p-value in the set S and the threshold family $\mathbf{t}$. In fact, it even outputs the entire false positive confidence curve associated to S. For example, the largest set S such that $\overline{\mathrm{FDP}}_\alpha(S) \leq \gamma$ is then obtained in linear time and space for any user-defined γ. This complexity cannot be improved since the output vector size is s. The validity of Algorithm 2 relies on the following formulation for $\overline{\mathrm{FP}}_\alpha(S_j)$.

Proposition 3. *For $j \in [|s|]$, let S_j be the index set of the j smallest p-values in S, and $\kappa_j = \sum_{k\in[|s|]} \mathbb{1}_{\{p_j \geq t_{k\wedge K}\}}$. Then*

$$\overline{\mathrm{FP}}_\alpha(S_j) = \min\left(\kappa_j, \min_{1\leq k\leq \kappa_j} v_k(S) - (s-j)\right). \tag{2.5}$$

Proposition 3 is proved in Appendix A.1.3. The fact that $\overline{\mathrm{FP}}_\alpha(S_j)$ depends on j only via κ_j but not S_j in Equation (2.5) is crucial for obtaining a linear time complexity. The properties of Algorithm 2 can be summarized as follows:

Corollary 1 (Validity and complexity of Algorithm 2)**.** *Algorithm 2 returns the vector $(\overline{\mathrm{FP}}_\alpha(S_j))_{1\leq j\leq s}$ in $O(s)$ time and space complexity.*

Proof of Corollary 1. **Valididy.** The **for** loop at lines 1- 8 stores the thresholds $(t_{k\wedge K})$ for $k \in [|s|]$. The **while** loop at lines 11-19 outputs both $(\kappa_j)_{j\in S}$ and $(r_k)_{1\leq k\leq K}$, where $r_k =$

Algorithm 2 Linear algorithm for interpolation-based post hoc bounds.

Require: $p = (p_{(1:S)}, \dots p_{(s:S)}), \mathbf{t} = (t_1, t_2, \dots t_K)$

1: $\tau \leftarrow \texttt{rep}(0, s)$
2: **for** $k \leftarrow 1$ to s **do**
3: **if** $k \leq K$ **then**
4: $\tau[k] \leftarrow t[k]$
5: **else**
6: $\tau[k] \leftarrow t[K]$
7: **end if**
8: **end for**
9: $\kappa, r \leftarrow \texttt{rep}(s, s)$
10: $k, j \leftarrow 1$
11: **while** $(k \leq s)$ & $(j \leq s)$ **do**
12: **if** $(p[j] < \tau[k])$ **then**
13: $\kappa[j] \leftarrow k - 1$ ▷ $\kappa[j] = |\{k/p[j] \geq t[k]\}|$
14: $j \leftarrow j + 1$
15: **else**
16: $r[k] \leftarrow j - 1$ ▷ $r[k] = |\{j/p[j] < t[k]\}|$
17: $k \leftarrow k + 1$
18: **end if**
19: **end while**
20: $V, A, M \leftarrow \texttt{rep}(0, s)$
21: **for** $k \leftarrow 1$ to s **do**
22: $A[k] \leftarrow r[k] - (k - 1)$
23: **if** $k > 1$ **then**
24: $M[k] \leftarrow \max(M[k-1], A[k])$
25: **else**
26: $M[k] \leftarrow A[k]$
27: **end if**
28: **end for**
29: **for** $j \leftarrow 1$ to s **do**
30: **if** $\kappa[j] > 1$ **then**
31: $V[j] \leftarrow \min(\kappa[j], j - M[\kappa[j]]))$
32: **end if**
33: **end for**
34: **return** V

$|\sum_{j \in S} \mathbb{1}_{\{p_j < t_k\}}|$. Noting that $r_k = s - v_k(S) + (k-1)$, the **for** loop at lines 21-28 outputs $M_k = \max_{k' \leq k} s - v_{k'}(S)$, that is, $M_k = s - \min_{k' \leq k} v_{k'}(S)$. Thus, the **for** loop at lines 29-33 outputs $V_j = \overline{\mathrm{FP}}_\alpha(S_j)$ by Proposition 3.

Complexity. All the vectors stored within the algorithm are of size s, so the space complexity of Algorithm 2 is $O(s)$. For the time complexity, the $(\kappa_j)_j$ and $(r_k)_k$ are calculated within a single **while** loop of size s, in which exactly one of j or k is incremented at each step. The rest of the algorithm consists of two **for** loops of size s consisting of $O(1)$ operations. □

Algorithm 2 has been available since 2016 within `sanssouci`. The original implementation of Algorithm 2 had $O(K \vee s)$ time and space complexity, i.e. $O(m)$ worse case complexity. It has been improved to $O(s)$ complexity in 2021 by adding the lines 1-8.

2.4 Urothelial Bladder Carcinoma data set

In this section, we focus on an Urothelial Bladder Carcinoma (BLCA) bulk RNA sequencing data set from the TGCA et al. (2014, TCGA). This preprocessed data set is available from the R/Bioconductor package `curatedTCGAData`. Internally, this package itself relies on the R/Bioconductor package `RTCGAToolbox` to download TCGA data that have already been preprocessed by TCGA pipelines. For convenience, this data set has also been made available in the R package `sanssouci.data` (Neuvial and Enjalbert-Courrech, 2024). This data set consists of gene expression measurements for $n = 270$ patients, classified into two subgroups: stage II ($n_1 = 130$) and stage III ($n_2 = 140$). Bladder cancer stages range from 0 to IV, quantifying how much the cancer has spread. We have filtered out unexpressed genes, here defined as those for which raw expression counts were lower than 5 in at least 75% of the patients. This results in $m = 12,534$ genes. To identify DE genes between the stage II and stage III populations, we test for each gene the null hypothesis that the gene expression distribution is identical in the two populations. The calibration method described in Section 2.2.2 is performed using a Wilcoxon (1945) rank sum test (also known as Mann and Whitney (1947) test) with the Simes template, with $B = 1000$ permutations and target risk (JER) set to $\alpha = 10\%$. The resulting method is called the **Adaptive Simes** method.

2.4.1 Confidence curves

In the absence of prior information on genes, a natural idea is to rank them by decreasing statistical significance. Post hoc methods provide confidence curves on the number (or proportion) of true positives (truly DE genes) among the most significant genes. Such curves are displayed in Figure 2.1 for the BLCA data set. The black lines in Figure 2.1 are $1 - \alpha = 90\%$ confidence curves obtained by the Adaptive Simes method. Upper bounds on FDP and lower bounds on TP are displayed in the left and right panels, respectively. For reference, the corresponding curves obtained by ARI are displayed in gray; they are almost identical to the ones obtained from the original bound of Goeman and Solari (2011) (see Equation (2.3) and Appendix A.2.1).

Post hoc guarantees. Post hoc inference makes it possible to define DE genes as the largest set of genes for which the FDP bound is less than a user-given value q. The arbitrary choice $q = 0.1$ is illustrated in Figure 2.1, corresponding to the horizontal line in the left panel. The black lines in Figure 2.1 correspond to the set S of 1064 genes for which the adaptive Simes method ensures that $\mathrm{FDP}(S) \leq q$. This corresponds to at least $\overline{\mathrm{TP}}_\alpha(S) = 958$ true positives (since $1 - 958/1064 = 0.1$), as illustrated in the right panel.

Adaptation to dependence. The above example also illustrates the increase in power obtained by Adaptive Simes thanks to the calibration described in Section 2.2.2. Indeed, for an identical statistical guaranty ($FDP \leq 0.1$), ARI yields a substantially smaller subset of 703 DE genes on a selection of 781 genes. More generally, the comparison between the black and gray curves in Figure 2.1 illustrates the gain in power obtained by using permutation methods to adapt to the dependence between genes.

Comparison to FDR control. For this data set, the BH procedure calls a set S of 1787 genes DE for a target FDR level of 0.05. As stated in Section 2.1, *the BH procedure does not provide guarantees on the FDP of these genes*, but only on their FDR, that is, the average FDP over hypothetical replications of the same genomic experiment and p-value thresholding procedure. Note that this remark is not specific to the BH procedure: the same would be true

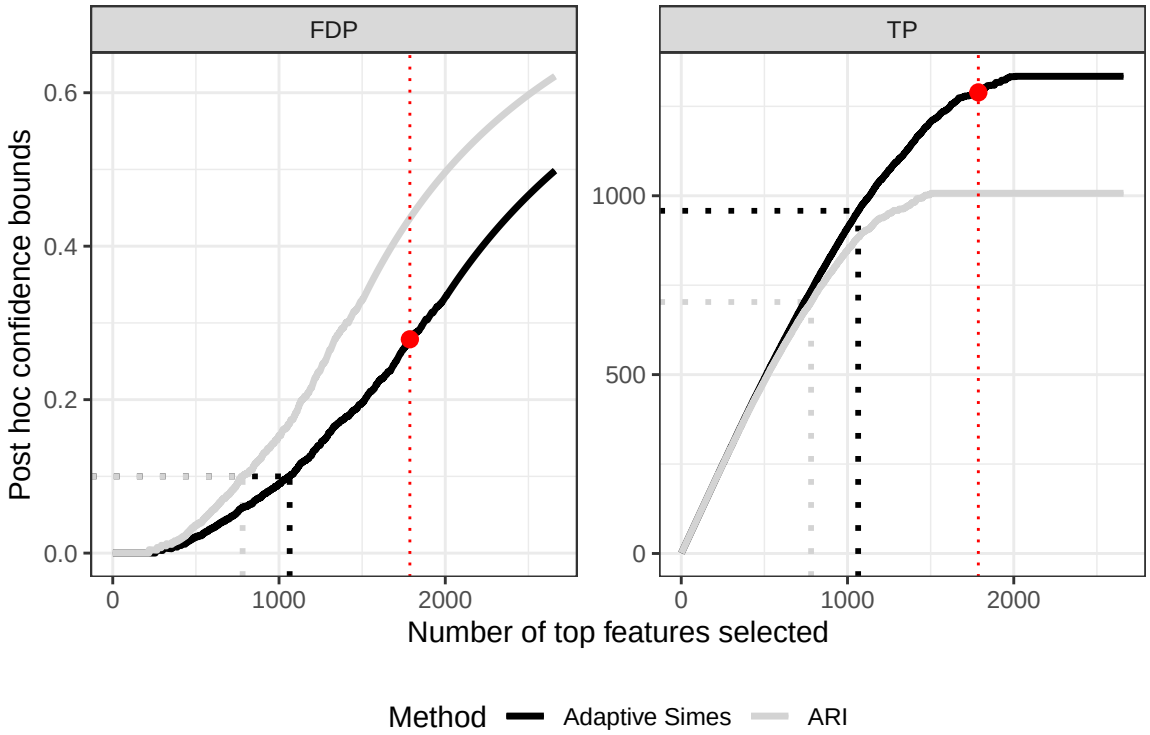

Figure 2.1 – 90% confidence curves on "top k" lists for the Urothelial bladder carcinoma data set. Left: upper bound on the False Discovery Proportion (FDP); right: lower bound on the number of true positives (TP). The adaptive Simes bound (black curves) outperforms ARI (gray curves). The horizontal dotted line on the left panel correspond to a common target upper bound on FDP of $q = 0.1$. Adaptive Simes selects 1064 genes at this confidence level, while ARI selects only 781. This corresponds to a lower bound on TP of 958 for the *adaptive Simes* method and 703 for the ARI method (dotted lines in the right panel). For reference, the set of 1787 genes called DE by the classical BH(0.05) procedure is represented by a red dot.

	Confidence curve (Fig. 2.1)		Volcano plot (Fig. 2.2)	
	ARI	Adaptive Simes	ARI	Adaptive Simes
Number of genes in S	781	1064	569	569
$\overline{\mathrm{TP}}_\alpha(S)$	703	958	456	492
$\overline{\mathrm{FDP}}_\alpha(S)$	0.1	0.1	0.199	0.135

Table 2.1 – Post hoc bounds on BLCA data set for ARI and Adaptive Simes, for gene selections S illustrated in Figures 2.1 (target FDP= 0.1) and 2.2 (genes filtered by p-value).

for any FDR controlling procedure. In contrast, the Adaptive Simes bound guarantees (with 90% confidence) that the number of true positives in S^{BH} is at least 1289, or, equivalently, that the corresponding FDP is less than 0.279.

2.4.2 Volcano plots

Volcano plots are a commonly used graphical representation of the results of a differential expression analysis (Cui and Churchill, 2003), illustrated in Figure 2.2. Each gene is

represented in two dimensions by estimates of its effect size (or "fold-change", x axis) and significance (y axis). The fold change of a gene is generally defined as the difference between the average or median (log-scaled) gene expressions of the two compared groups. Its significance is quantified by $-\log_{10}(p)$-values for the test of its differential expression, where the "$-\log_{10}$" transformation ensures that large values of y correspond to genes which are likely to be differentially expressed.

As noted by Ebrahimpoor and Goeman (2021), post hoc inference makes it possible to select genes of interest based on both fold change and significance without compromising the validity of the corresponding bounds. Moreover, even if Wilcoxon tests have been performed for the *calibration of the post hoc bounds*, our proposed post hoc bounds are still valid when relying on other statistics for the *selection* of genes of interest. Figure 2.2 illustrates this idea with a volcano plot based on the p-values and log-fold changes obtained from the limma-voom method of Law et al. (2014), which is implemented in the `limma` package of Smyth (2004). In this example, the function $\overline{\mathrm{FP}}_\alpha$ defined in Equation (2.3) depends on the Wilcoxon tests via the p-values $(p_j)_j$ and the thresholds $(t_k)_{k\in[|K|]}$ obtained at the calibration step, but it is statistically valid for arbitrary gene selections S. By construction, for a given selection size $|S|$, the tightest bound $\overline{\mathrm{FP}}_\alpha(S)$ corresponds to the set of the $|S|$ smallest Wilcoxon p-values. More generally, smaller bounds $\overline{\mathrm{FP}}_\alpha(S)$ will be obtained for selections S that consist of small Wilcoxon p-values. A quantitative comparison between the Wilcoxon and limma-voom p-value is provided in Figure A.2. It illustrates the coherence of the two methods for identifying DE genes in the settings considered.

An example selection of 569 genes is highlighted in red in Figure 2.2. It corresponds to genes whose p-value is less than 10^{-3} and fold change larger than 0.5 in absolute value.

The Adaptive Simes method ensures that with probability larger than 90%, the proportion of false discoveries (FDP) is less than 0.14. It also ensures that the FDP among the subset of 493 genes with positive fold change is less than 0.14, and that the FDP among the subset 76 of genes with negative fold change is less than 0.63. As already noted, the proposed bounds can be computed for multiple, arbitrary gene subsets (obtained e.g. by changing the p-value and fold change thresholds in Figure 2.2) without comprising their validity. Here again, the Adaptive Simes method yields tighter bounds than ARI, as illustrated in Table 2.1.

2.4.3 Influence of the number of permutations

The adaptive Simes method relies on random permutation of class labels. As such, running this method several times could lead to different results. Larger values of the number B of permutations are expected to give more stable results. However, this comes at a higher computational price, since the theoretical time complexity of the calibration step is linear in B, see Section 2.2.2. To study the impact of B, we use the BLCA data set where the number of genes is $m = 12{,}534$ and the number of patients is $n = 270$, divided in two subgroups of size $n_1 = 130$ and $n_2 = 140$. As in Section 2.4, the calibration is performed with the Wilcoxon rank sum test, the Simes template and a target confidence level $1 - \alpha = 90\%$. We focus on three different gene selections:

BH_05: the genes selected by the Benjamini and Hochberg (1995) procedure at level $q = 5\%$

first_1000: the 1000 genes with smallest p-value

H: all genes in the data set.

The parameter of the experiment is $B \in \{100, 200, 500, 1000, 2000, 5000\}$. Figure 2.3 summarizes, for each value of B, the empirical distribution of the post hoc bounds obtained across 1000 draws of B permutations. As expected, the dispersion of the bounds decreases as the

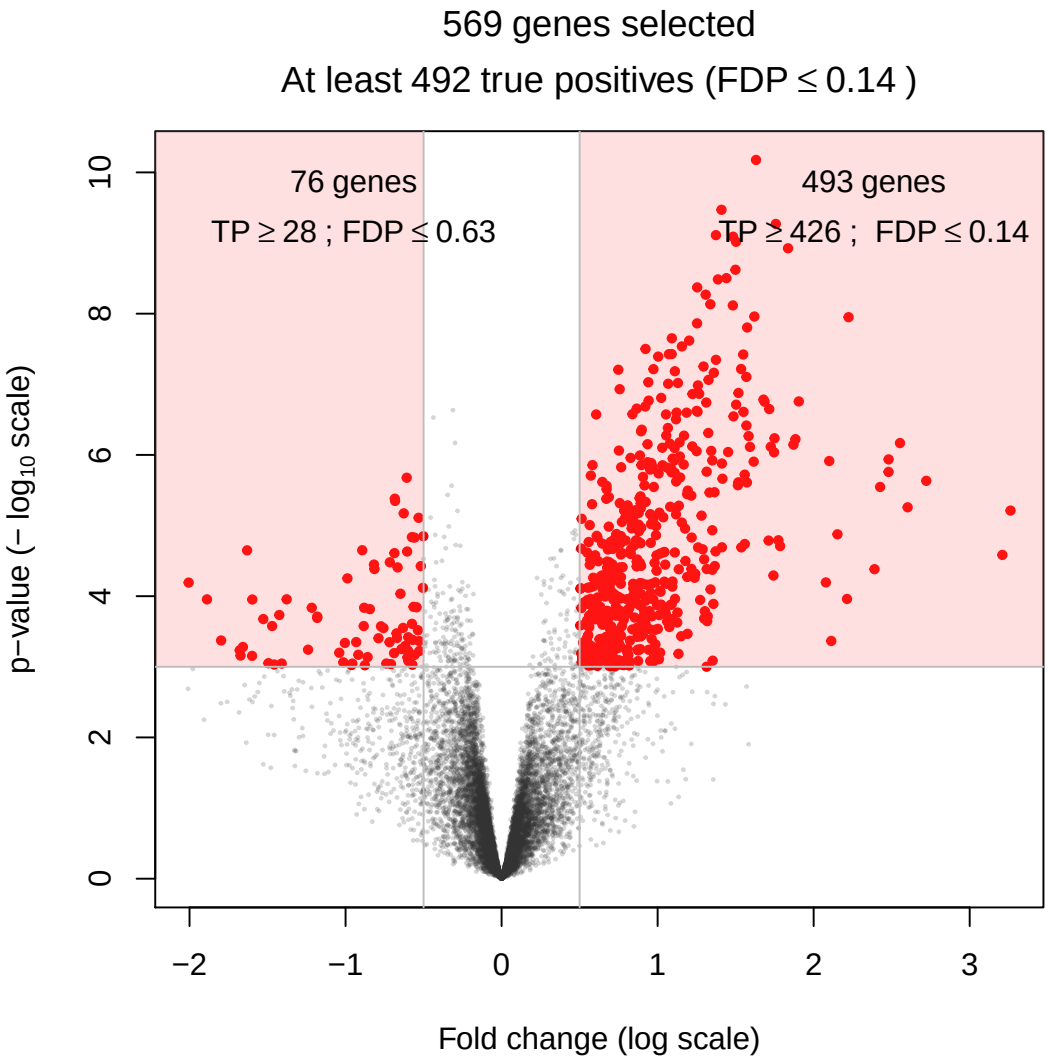

Figure 2.2 – Volcano plot for the Urothelial Bladder Carcinoma data set. Each dot corresponds to a gene, represented by its fold change (x axis) and p-value (y axis) on the log scale. Fold changes and p-values were obtained by the limma-voom method (Smyth, 2004). The 569 genes with p-value less than 10^{-3} and fold change larger than 0.5 are highlighted. The Adaptive Simes method ensures that at least 492 of these genes are true positives.

number of permutations increases. Figure 2.4 summarizes the same experiments by plotting the variability of the FDP bound (across 1000 draws of B permutations) against the corresponding average computation time. The variability is quantified by the inter-centile range (ICR), that is, the difference between the 99% quantile and the 1% quantile of the empirical distribution. Based on these results, $B = 1000$ appears to be a reasonable default choice to balance computation time and precision. For example, for the 1787 genes selected by the BH procedure at level $q = 0.05$, the FDP bound is between 0.23 and 0.37 for 99% of the 1000 replications of the calibration procedure. $B = 1000$ is the default value used in the `sanssouci` package. Users may of course change this parameter manually.

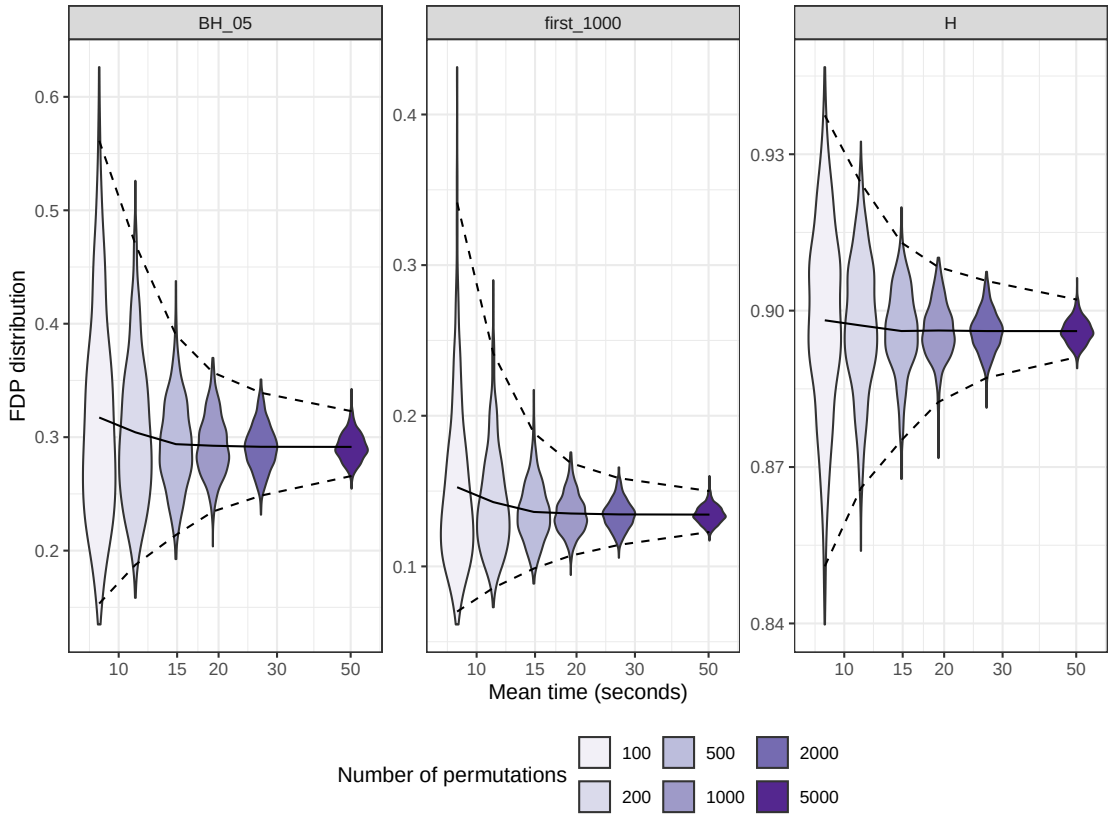

Figure 2.3 – Empirical distribution of the post hoc bounds on FDP across 1000 draws of B permutations, for $B \in \{100, 200, 500, 1000, 2000, 5000\}$. Each panel corresponds to a specific gene selection. Solid lines indicate the average FDP, and dashed lines indicate the 1% and 99% quantiles.

2.5 Statistical performance for DE studies

2.5.1 Existing post hoc inference methods

The first post hoc inference methods introduced were not adaptive to the dependence between tests, since they were obtained from probabilistic inequalities:

— The **Simes** bound was first proposed in Goeman and Solari (2011) together with a quadratic algorithm ($O(m^2)$). It has been implemented in the R package `cherry`.

— A slightly sharper version of the Simes bound has been introduced by Goeman et al. (2019), together with an algorithm of linearithmic complexity. This method is known as **ARI** for "All-Resolution Inference" and implemented in the R package `hommel`.

The idea of using randomization to obtain sharp risk control is not new in the multiple-testing literature. In particular, resampling or permutations have been used to control the Family-Wise Error Rate (FWER, (Ge et al., 2003; Westfall and Young, 1993)) and the k-FWER (Romano and Wolf, 2007). For post hoc inference:

— The **Adaptive Simes** method described in this chapter exploits sign-flipping and permutation-based approaches introduced by Blanchard et al. (2020, 2021) in order to build post hoc bounds. It has been implemented since 2017 in the R package `sanssouci`.

— A closely related approach called **pARI** has recently been proposed by Andreella et al. (2023) for the analysis of neuroimaging data. It is implemented in the R package `pARI`.

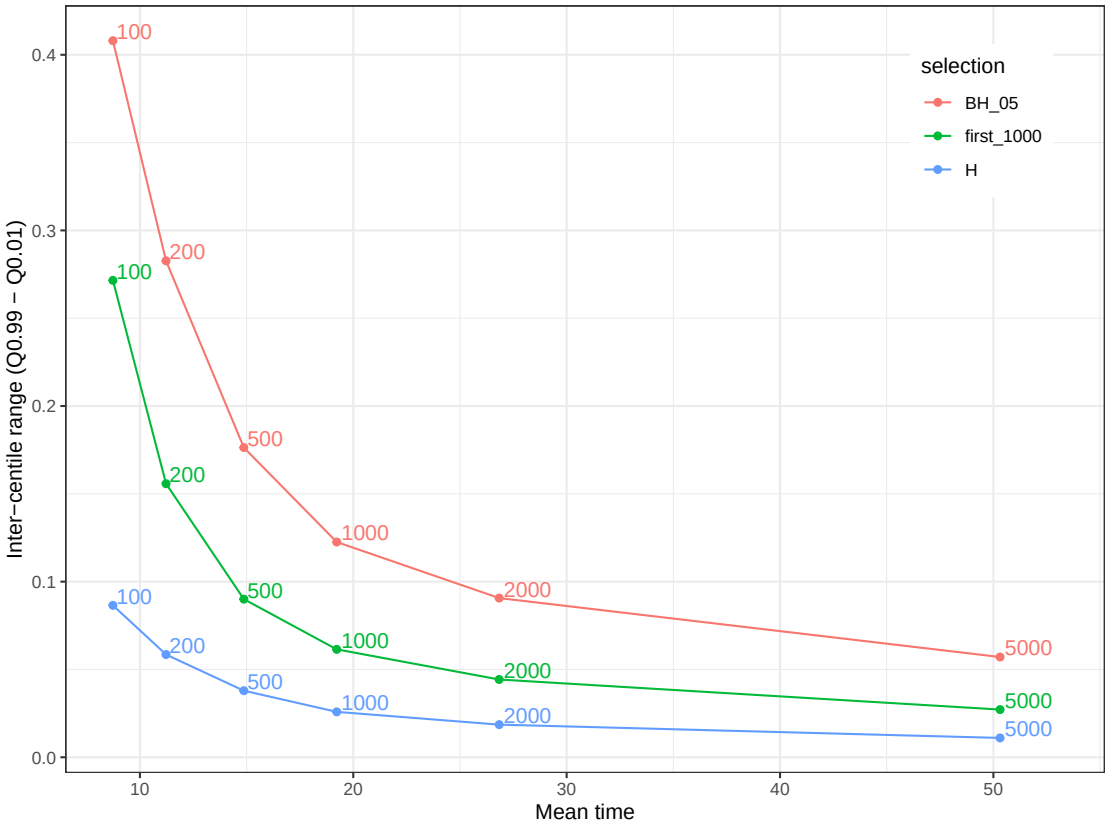

Figure 2.4 – Variability of the post hoc bounds (measured by the inter-centile range, ICR) as a function of the computation time, for $B \in \{100, 200, 500, 1000, 2000, 5000\}$. The ICR is estimated from 1000 draws of B permutations, .

Both pARI and Adaptive Simes, rely on the calibration method described in Section 2.2.2, combined with the interpolation bound in Equation (2.3). An important difference between Adaptive Simes (R package `sanssouci`) and pARI is that `sanssouci` implements the linearithmic time complexity algorithm described in Section 2.3. In contrast, the algorithm used in `pARI` to calculate the post hoc bound after calibration is the "naive" interpolation algorithm described at the beginning of Section 2.3, which has a quadratic complexity for a single set.

While `pARI` initially only implemented a single-step version, it has been updated after the initial submission of Enjalbert-Courrech and Neuvial (2022), so that both `sanssouci` and `pARI` now implement a step-down principle in order to adapt to the unknown quantity of signal (or, equivalently, to the proportion π_0 of true null hypotheses). These two step-down methods have the same goal, but they are based on different principles. The experiments reported below and in Appendix A.2.1 show that our proposed step-down typically only provides marginal performance improvements on real data due to the sparsity of the signals. Similar observations have been made for the step-down version of `pARI` (see Andreella et al., 2023), so we have not included this method in our comparisons.

The main features of existing post hoc bounds are summarized in Table 2.2.

2.5.2 Evaluation framework

The mathematical validity of the post hoc bounds considered in this chapter has been proved in Blanchard et al. (2020), where their numerical performance has also been illustrated by experiments on synthetic data. This section aims to complement these results with

Method	R Package	Adaptivity to: dependence	Adaptivity to: π_0	Time complexity
Simes	`cherry`	NO	NO	quadratic
ARI	`ARI`	NO	YES	linear
permutation ARI	`pARI`	YES	YES	quadratic
Adaptive Simes	`sanssouci`	YES	YES	linear

Table 2.2 – Main features of existing post hoc inference methods and software.

numerical experiments based on gene expression data, which are more realistic for DE studies.

Data set generation. Without a gold-standard data set from which one would know which genes are truly DE or not, we created the following data sets. Starting from a $n \times m$ gene expression data set X, where each column corresponds to a gene and each row to an experiment or statistical observation, we have

1. partitioned the observations into two groups of size n_1 and n_2, such that $n_1 + n_2 = n$;
2. partitioned the genes into m_0 null genes and m_1 non-null genes, with $m_0 + m_1 = m$
3. modified the expression of the non-null genes in Group 2 by shifting or scaling the corresponding submatrix of X of size $n_2 \times m_1$.

This process results in a perturbed gene expression data set Y where the null and non-null genes are known. Following Blanchard et al. (2020), for a set of such experiments, we have quantified estimates of the risk (JER) and the power of each method considered for each value of the target risk α. The JER results are presented in Section 2.5.3. The power results, which are highly consistent with the JER results, are postponed to Appendix A.3.

The empirical risk of a given method is estimated by the proportion of experiments for which the corresponding confidence curve on the false positives is not always below the actual number of false positives. In other words the empirical risk is given by $\widehat{\mathrm{JER}}(\alpha) = \frac{1}{Q}\sum_{i=1}^{Q} \mathbb{1}_{\{\exists\ j\in[|m|], \overline{\mathrm{FP}}_\alpha(S_j) < \mathrm{FP}(S_j)\}}$, with S_j the set of the j smallest p-values in the set of the m p-values and Q the number of experiments. This quantity is the empirical counterpart of the JER defined in Equation (2.2), and can be compared to the target risk α: JER is empirically controlled if the empirical JER is lower than α, and the closer it is to α, the tighter JER control.

The parameters of such a numerical experiment are the proportion $\pi_0 = m_0/m$ of null genes, and a measure of distance (or signal to noise ratio) between null and non-null genes. The numerical results obtained for RNA sequencing data are reported Section 2.5.3. We have also performed the same type of experiments with microarray data. The results are similar, and they are reported in Appendix A.4.

A core feature of our proposed method is to use *group label permutations* in order to obtain statistically valid procedures that adapt to the dependency between genes observed in the data set at hand. However, the number of distinct permutations is limited for lower sample sizes: for example, 252 distinct permutations are available for a comparison between two groups of size 5. We have evaluated the impact of the sample size on the performance of the methods in Appendix A.5. This can be done by down-sampling the BLCA data set and retaining only a smaller number of observations. Our experiments demonstrate that even for sample sizes less than 10, the Adaptive Simes method yields sharper bounds than its competitors. Although the Adaptive Simes method is formally valid regardless the sample size, we recommend using it in studies with more than 5 samples per group.

2.5.3 Results for bulk RNA sequencing data

Our starting point is the data set used in Section 2.4. We have only selected stage III samples for this experiment and performed the same filtering as in Section 2.4 on these samples only. We obtained a "null" data set (with no signal), consisting of 130 patients and $m = 12{,}418$ genes, after applying the same process [3] described in Section 2.4 to filter out unexpressed genes. The parameters of the experiments are set as follows. The proportion of null genes is set to $\pi_0 \in \{0.8, 0.95, 1\}$. We have considered a multiplicative signal for differential expression: for each gene j among the m_1 non-null genes, the original expression values of j are multiplied by a constant ς_j for n_1 of the n observations, where ς_j is drawn uniformly between 1 and a signal to noise (SNR) parameter. The SNR value is set to 1 (no signal), 2 or 3 (weak to strong signal). We have used a two-sided Wilcoxon rank sum test to compare the two groups.

The results are summarized by Figure 2.5, where the average empirical risk (JER) achieved across 1000 experiments is plotted (together with 95% confidence curves) against the target risk α for the methods described in Section 2.5.1. In particular, the single-step version of pARI is represented by the "Adaptive Simes (single step)" method.

Each panel corresponds to a combination of the parameters $\pi_0 \in \{0.8, 0.95, 1\}$ (in columns) and SNR $\in \{1, 2, 3\}$ (in rows). The JER is controlled for all methods and all parameter combinations, since all curves are below the diagonal. The risk for the Adaptive Simes methods is substantially closer to the target risk than for the parametric Simes methods (Simes and ARI). This illustrates the systematic gain in tightness provided by the calibration method described in Section 2.2.2. We also note that the gain obtained from the adaptation to π_0 is very small, and even negligible for $\alpha \leq 0.2$. Indeed, the Simes and ARI methods are essentially indistinguishable from each other, and the same holds for the single-step and step-down Adaptive Simes methods. These results are also confirmed by those of the power assessment (see Appendix A.3).

2.6 Discussion

This chapter advocates for the use of post hoc inference in two-group DE studies, which provide more interpretable statistical guarantees than classical inference based on the False Discovery Rate. The methods proposed in this chapter make it possible to obtain post hoc bounds that are both fast to compute, and powerful (in the sense of the proportion of true signal recovered). The resulting improvement over the state-of-the-art is illustrated by realistic numerical experiments based on RNAseq and microarray data. These methods are implemented in the open-source R package `sanssouci`. The code used for the numerical experiments of this chapter and to generate the figures is also provided [4]. Based on these numerical experiments, we recommend using $B = 1000$ permutations as an acceptable compromise between computation time and power. Moreover, even though our proposed methods are theoretically valid regardless of the sample size, we recommend using them in studies with at least 5 samples per condition so that the number of distinct group-label permutations is large enough to provide adaptation to the dependency between genes observed in the data set at hand.

The methods proposed in this chapter and their implementation in the R package `sanssouci` are generic in the sense that they can be used with reference families (or *templates*) of arbitrary shape. The most natural choice is the Simes family (which corresponds to a linear

3. The number of genes differs because the procedure is applied only to the $n_1 = 130$ patients in group 1 instead of the $n = 270$ patients in the study.

4. https://github.com/sanssouci-org/IIDEA-method-paper

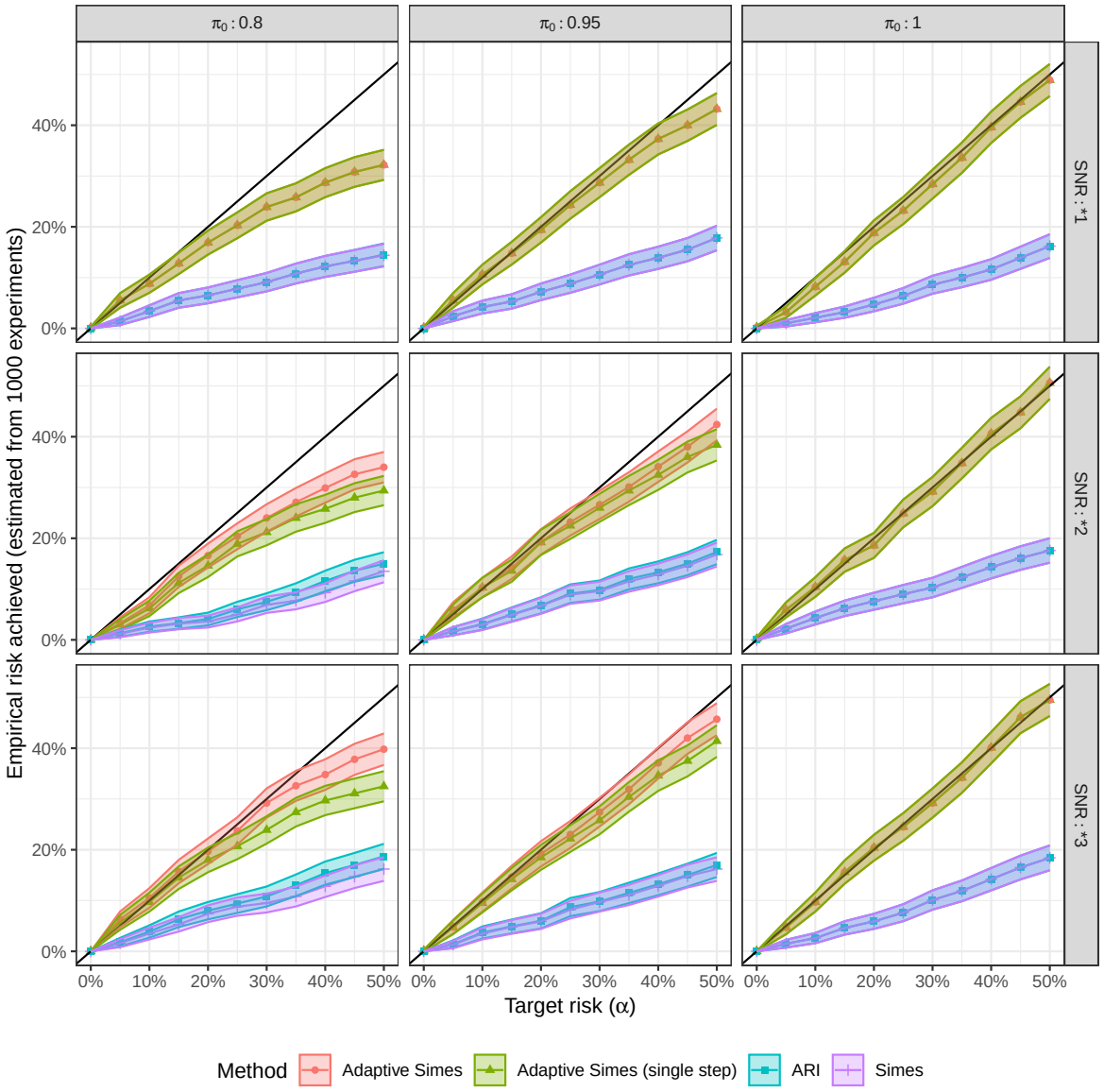

Figure 2.5 – Validity and compared tightness of the post hoc bounds on RNA-seq based numerical experiments. The average empirical JER achieved across 1000 experiments is plotted (together with 95% confidence curves) against the target risk α for all considered methods. Each panel corresponds to a combination of the parameters π_0 and SNR.

template), as it is closely related both to FDR control and to the first post hoc bounds introduced by Goeman and Solari (2011). The resulting method, which is called the Adaptive Simes method, is used in the numerical experiments reported in this chapter. An interesting perspective of this work is to compare the performance of other templates. Our experience in DE studies indicates that improving on the Simes family by changing the template is challenging; similar conclusions have been reported in Andreella et al. (2023) for the analysis of fMRI data. Recent works in this field have shown the superiority of a fully non-parametric approach, whereby the entire family of templates (instead of a single parameter λ as in the present work) is learned from external data (Blain et al., 2022). Applying this method to genomic data is another exciting perspective for the present work (see Chapter 4).

While this chapter focuses on DE studies, these methods and our implementation are applicable to any practical situation involving multiple two-sample tests, or one-sample tests, or tests of association with a continuous outcome. Such situations are frequent in genomics (differential expression, differential splicing, differential methylation) but also in neuroimaging, which is another field where post hoc inference methods have been introduced (Rosenblatt et al., 2018). However, for studies that have more complex designs, such as multi-sample comparisons or studies including covariates, the calibration-based approach proposed here cannot be applied directly. This is a limitation of the Adaptive Simes method when compared to the state-of-the-art ARI method, which is applicable in any multiple testing framework where the Simes inequality holds. We view this limitation at the (current) price to pay in order to obtain the substantial power gains that are illustrated numerically in this chapter. Extensions of the present work to the problem of testing parameters of a general linear model are developed by Davenport et al. (2022) and described in Chapter 4.

Chapter 3

IIDEA: Interactive Inference for Differential Expression Analyses

3.1 Introduction

A commonly used method for selecting genes of interest in genomic studies is to use a volcano plot (Cui and Churchill, 2003), which displays an effect size measure (the log fold change) and statistical significance (the p-values) simultaneously (see an example in Figure 2.2). Biologists are used to select genes of interest by selecting genes with a small p-value and a large log fold change. As argued in Section 1.2.3, FDR control is not adapted to such data-driven gene selections. This issue has been studied numerically in Ebrahimpoor and Goeman (2021), which demonstrate the inflation of the false discovery rate based on simulation experiments and the analysis of bulk RNAseq dataset. The authors advocate the use of post hoc inference methods in this context.

The strength of post hoc inference lies in its ability to provide statistical guarantees for any selection made from prior knowledge or after observing the data. Moreover, this approach maintains these guarantees even after successive selections, which is crucial for the reproducibility of biological results. With post hoc inference, users can refine their selection to identify sets of genes of interest by manually re-running scripts and trying different threshold values. However, this requires programming skills and lacks interactivity. To address these limitations, we have developed an interactive application with the R package `shiny` (Chang et al., 2021) named IIDEA (Interactive Inference for Differential Expression Analyses). IIDEA makes it possible to select genes of interest directly from the volcano plot, for which the post hoc bounds from the *Adaptive Simes* method (introduced in Section 2.2) are calculated instantly.

This solution offers several major advantages for biologists. Firstly, it allows manipulation of the method without requiring coding skills, thereby facilitating the selection of rejection hypotheses. Secondly, it minimizes the risk of errors when applying the method. Finally, it eliminates the need to install R packages using the deployed application on an accessible platform. IIDEA can be used for DE studies comparing two groups for data from microarray and bulk RNA seq data as described in Chapter 2. Section 3.2 provides an overview of IIDEA. The description of the interactive selection is provided in Section 3.3. Section 3.4 describes how the post hoc bounds are computed. IIDEA's gene set enrichment analysis is provided in Section 3.5. The deployment strategy used is described in Section 3.6. The application is available in the R package `IIDEA` (Enjalbert-Courrech, 2024) and the deployed version is available online [1].

3.2 Overview

The primary objective of IIDEA is to facilitate the selection of genes of interest while providing statistical guarantees for such selections. Figure 3.1 presents an overview of the

1. https://shiny-iidea-sanssouci.apps.math.cnrs.fr/

application. Users start by choosing a public dataset or uploading their own data, and selecting parameters for the post hoc method (blue dashed box in Figure 3.1). After pressing the 'Run' button, a volcano plot is generated on the right, enabling direct gene selection in the green dashed box in Figure 3.1. Simultaneously, the post hoc bounds associated to these gene selections is displayed in a table on the left side (red dash-dotted box in Figure 3.1).

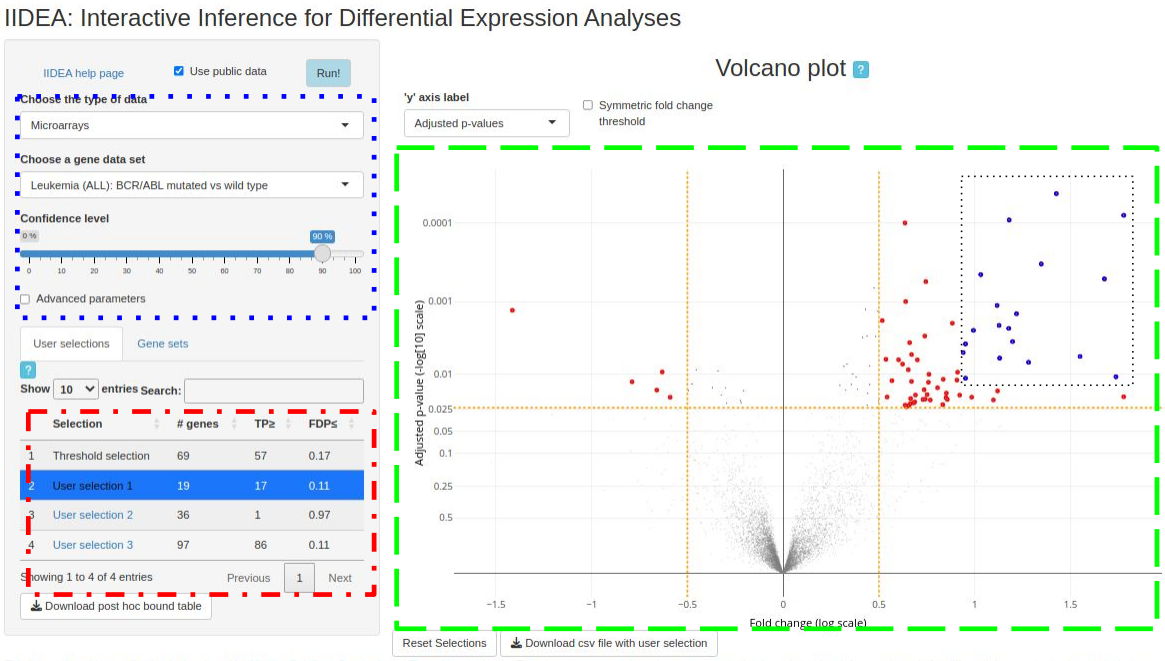

Figure 3.1 – Overview of IIDEA. Elements in the blue dotted box correspond to inputs of the methods. The selection on the volcano plot can be made in the green dashed box. The corresponding table of post hoc bounds is contained in the red dashed-dotted box.

3.3 Post hoc bounds for interactive gene selections

An interactive graphical user is intended to be fast to use in order to provide user interactivity. Making a selection on the volcano plot instantly updates the post hoc bounds table (see the red dashed-dotted box in Figure 3.1). This display is fast thanks to Algorithm 2 described in Section 2.3, which operates in linear time in the size of the selection. However, the volcano plot used in the IIDEA application includes as many points as there are genes ($m \approx 10{,}000$ to $20{,}000$). With a naive implementation, any modification to the volcano plot requires redrawing the entire plot. For example, when selection thresholds are adjusted, the points must be re-displayed to highlight the selected ones (in red). This can reduce the application's interactivity for the user.

To address this issue, the volcano plot is created with the R package `plotly` (Sievert, 2020), an open-source graphing library that offers various features for creating interactive and dynamic plots. The package `plotly` is implemented in several programming languages, including Python, R, JavaScript, and Julia. It offers sharing and embedding options for integration into web applications. IIDEA relies on the R version of this package to display the volcano plot (green dashed box of Figure 3.1).

The `plotly` package displays a multi-layered plot to avoid loading the entire plot with each modification made. Figure 3.2 illustrates the layer composition of the volcano plot

in IIDEA. The backbone of the plot (the grey points representing all genes, first layer) is constantly displayed, and only the layers displaying gene selections (red points for current threshold selections and blue points for previous selections, second and third layers) or orange thresholds (fourth layer) are modified. This modification of only a small part of the plot saves display time, which otherwise can be lengthy due to the large number of points.

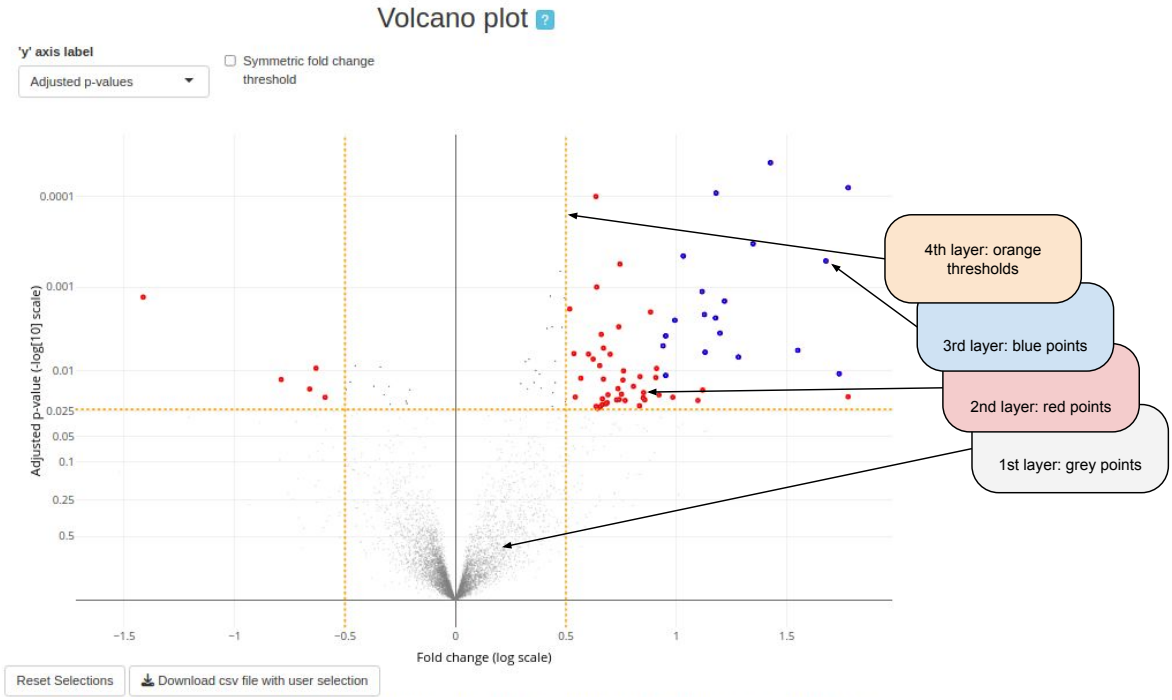

Figure 3.2 – Utilizing multiple layers of the `plotly` package for constructing the volcano plot in IIDEA. Each layer contains an object that can be independently modified. The graph is based on the first layer consisting of a grey scatter plot (representing all studied genes). This layer is almost fixed, since the user can change the y-axis labels (see Figure 3.4 for the y-axis displayed in IIDEA). The second layer consists of a red scatter plot (representing points selected by thresholds). The changes to this layer are due to the drag of the orange thresholds on the volcano plot. The third layer comprises a blue scatter plot (representing points re-displayed from a previous selection). This layer is modified by clicking on the post hoc bound table (see table in the red dash-dotted box in Figure 3.1). The final layer corresponds to the orange thresholds. The user can move these to make the selection.

The `plotly` package provides different ways of selecting points of interest by orange thresholds but also by "box selection" (rectangular area) or "lasso selection" (area closed by free drawing), allowing users to select areas other than those induced by the thresholds. The thresholds can be directly adjusted on the plot, allowing for a visual adjustment of the selected points. This feature is a major advantage for users, who benefit from easy threshold selection. While the "box" and the "lasso" selections are by default integrated into plotly's plots, we implemented the orange thresholds specifically for the purpose of IIDEA. An example of thresholds, box and lasso selection is given in Figure A.10 (in Appendix A.7).

IIDEA also provides the following features:

— All successive selections are stored in the table of post hoc bounds (red dashed-dotted box in Figure 3.1). Clicking on a line of this table displays the corresponding selection in the volcano plot, as illustrated in Figure 3.1 with the "box selection" (black dotted box) corresponding to the second row in the table of post hoc bounds.

— Clicking on the name of a gene selection in the table displays a graph between the corresponding genes based on the *string-db* collection [2]. This information may help to interpret the interaction results of the volcano plot in terms of interactions between genes (see Figure 3.3).
— All gene selections and the table of post hoc bounds can be downloaded as a *.csv* file, in the form of a binary matrix assigning genes to selections (see download buttons in Figure 3.1 respectively below the green and the red boxes).

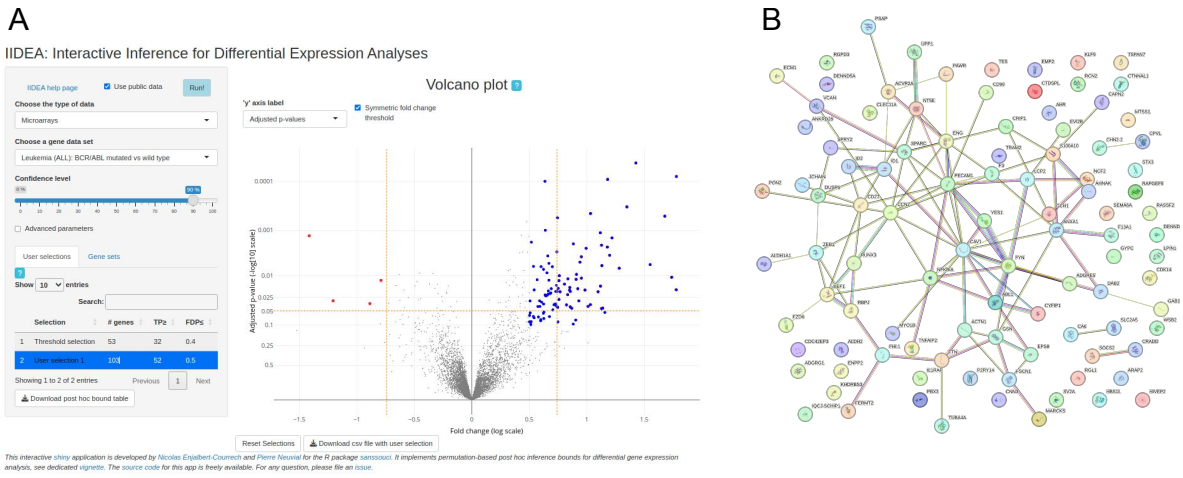

Figure 3.3 – Example of an interactive graph between selected genes from the *string-db* collection. Users can access this graph from the table of post hoc bounds. In this example, a user has made a selection on the volcano plot (see blue points and the corresponding row in the post hoc bound table in Panel A). The corresponding interactive graph in displayed in Panel B.

Alternative labeling for y axis. The y axis of the volcano plot displays the values of $-\log_{10}$ p-values, so that larger values correspond to genes more significantly differentially expressed. It is possible to change the labels of the y-axis without changing the points displayed, as illustrated by the three panels of Figure 3.4. The first alternative option for the y label corresponds to p-values adjusted by the BH method on a $-\log_{10}$ scale (see Figure 3.4-B). With this scale, points above the horizontal orange line correspond to genes selected by the BH method at the corresponding threshold value. The second alternative option for the y label directly involves post hoc bounds (see Figure 3.4-C). The values displayed on the axis are the post hoc upper bounds on the false positives among the set of all genes above the corresponding ordinate. Note that both of these alternative representations make sense because the ordering of the p-values, the adjusted p-values, and the post hoc bounds are identical. This display feature is implemented efficiently thanks to the modularity of the `plotly` graphs, allowing modifications to a specific part of the graph (here, the y label) without affecting the rest.

2. https://string-db.org/

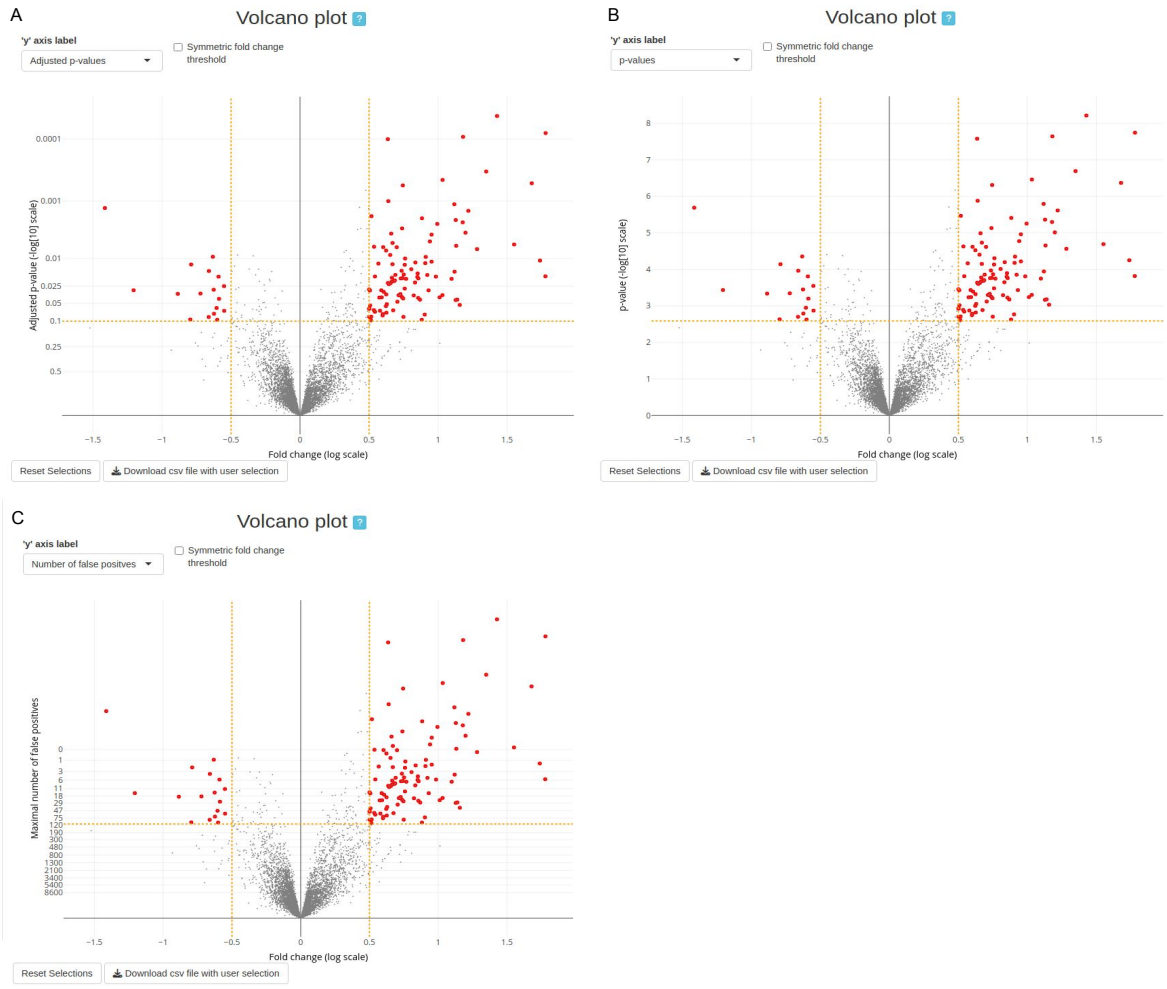

Figure 3.4 – y-axes display in IIDEA. Users can change the y-axis of the volcano plot to refine the selection. Possible choices are: 'p-values', 'Adjusted p-values' and 'Number of false positives'. In this example, the threshold on p-value is set such that the set of genes above the threshold controls the FDR at level 0.1 (see Panel A). That corresponds to a threshold on p-values of 0.0026 (see Panel B). The third axis indicates that in the genes above the threshold, there are not more than 120 false positives (see Panel C).

3.4 Computation of post hoc bounds: richer inputs yield richer outputs

To display the volcano plot, the user needs to provide at least a matrix containing, for each gene, the associated p-value of the test as well as the log fold change measurement. With this information, IIDEA can display the volcano plot to allow the user to make their selection. With this information, the bounds based on the Simes inequality can be computed, since these bounds only depend on the p-values: this corresponds to Equation (1.10) with $t_k = \alpha k/m$. As argued in Chapter 2, these bounds are typically conservative and adaptive bounds should be preferred. To fully harness the power of the *Adaptive Simes* method (Blanchard et al., 2020; Enjalbert-Courrech and Neuvial, 2022), the user must provide the original expression matrix $\mathbf{X}$ containing gene expression profiles for each biological sample. With this matrix, the method can use the calibration algorithm introduced in Section 2.2.1.2 by permuting group

labels to learn the joint p-value distribution of the true null hypotheses, and obtain sharper bounds (see the numerical analysis proposed in Section 2.5).

However, computing the permuted p-values can be time-consuming, which can make user interaction cumbersome in the context of IIDEA. To decrease computation time, a matrix-based version of the tests has been implemented in the `sanssouci` package. This computational approach relies on vectorization techniques and is based on matrix products between a permutation matrix of the labels and the expression matrix, which are illustrated in Figure 1.4. This method substantially reduces execution time by generating a $B \times m$ matrix of p-values without looping over permutations and genes. A similar idea is implemented in the R package `matrixTests` (Koncevičius, 2023). However, the approach in Koncevičius (2023) only implements one layer of vectorization (with respect to the genes), and not the second layer (with respect to the permutations). Numerical simulations demonstrating the computational gains are presented in Figure A.3. The blue and green curves represent matrix-based and loop-based calculations, respectively. The matrix-based calculation offers a 100x speedup compared to the loop-based calculation, for any tested value of B. Additionally, the execution time appears constant in m for the tested values of B in the matrix-based version, in contrast to the increasing time for the loop-based version. This implementation improves execution time, allowing the application to be more responsive.

To optimize the reactivity of IIDEA, we have exploited the modularity of the calibration algorithm illustrated in Figure 1.4. In case of a parameter change in the method, all calculations do not need to be re-executed. For example, the method has already been executed for a certain number of permutations B, a template, and a confidence level α. If only the confidence level changes, then the permuted p-values and the pivotal statistic can remain the same. Applying the new confidence level to find the quantile λ that controls the JER is sufficient. Another example is changing the template: the permuted p-values remain the same, and only the pivotal statistic and the quantile are updated. This modularity saves computation time depending on which parameters are changed. However, changing the number of permutations currently requires recalculating everything.

By default, the p-values and log fold changes displayed by IIDEA are computed based on the same test statistics as those used at the calibration step to obtain the post hoc bounds, that is, Welch two-sample tests for microarray data, and Wilcoxon rank tests for bulk RNAseq data. In practice however, users may be interested in performing gene selections based on alternative test statistics, as discussed in Section 2.4.2, where the volcano plot is generated using p-values obtained from the `limma-voom` method, while the calibration step computes B permuted p-values from the Wilcoxon test to approximate the joint p-value distribution under the true null hypotheses. This feature has been implemented in IIDEA by decoupling the statistics used for display from those used for post hoc inference. In order to use this feature, users can provide a matrix containing p-values and log fold changes corresponding to a statistic of their choice (in addition to $\mathbf{X}$ which is then only used for the calibration of post hoc bounds).

3.5 Gene set enrichment analysis

Enrichment analysis aims to identify sets of genes associated with known biological functions or metabolic pathways as described in the introduction (Section 1.1.2.1). For this purpose, several strategies exist. One of them is to simultaneously consider groups of genes known beforehand and estimate a post hoc bound on each group. Post hoc methods provide simultaneous guarantees on multiple selected gene sets and can, therefore, be applied to gene set enrichment analysis as shown by Ebrahimpoor et al. (2020), introducing Simultaneous Enrichment Analysis (SEA). They particularly show that self-contained and competitive ap-

proaches (Goeman and Bühlmann, 2007) can be addressed with post hoc inference. These two approaches are distinguished as follows:

self-contained tests the null hypothesis "No gene in the set is active" (see, e.g., the Global Test of Goeman et al. (2004)). This null hypothesis is common in statistics (as in ANOVA). This method is considered very powerful since only one DE gene in the set is sufficient to consider the set of genes active. As noted by Ebrahimpoor et al. (2020), gene sets significant for this test are those for which at least one true positive is detected (TP ≥ 1).

competitive tests the null hypothesis "The genes in the set are at least as active as those in the complementary set," with the complementary set of genes being all those not in the tested set (see, e.g., GSEA (Subramanian et al., 2005) or GSA (Efron et al., 2007)). Ebrahimpoor et al. (2020) note that significant gene sets for this test are those where the proportion of DE genes in the tested set is at most as large as that in the complementary set.

This type of analysis is implemented in the IIDEA application (see Figure 3.5). By default, gene sets from the Gene Ontology (2015, GO) are available in IIDEA. Calibration is performed, and post hoc bounds on the tested gene sets are displayed in the corresponding table. A filter on the gene sets is possible by meeting the criteria of self-contained or competitive. With this approach, the volcano plot is not used for selecting genes. However, the genes from a selected gene set can be displayed on the volcano plot (as blue points) to refine the interpretations. For example, in Figure 3.5, genes contained in the Gene Ontology "GO:0007186" are printed with blue points in the volcano plot. To refine the biological interpretation, clicking on the name of a gene set links to the dedicated Gene Ontology web page [3]. Users can also upload their annotation matrix. If the matrix provides GO identification codes, then the user will have access to the same additional information. If not, they will not have access to these additional annotations but will still obtain the post hoc bounds corresponding to their annotations.

3.6 Application deployment

The IIDEA application is available through its corresponding R package [4] (Enjalbert-Courrech, 2024). While users familiar with the R environment may easily use the package, it can present challenges for those less experienced. Deploying the application on a web-based platform can address this issue. One practical solution is to use RStudio's shinyapps.io, which provides a secure and reliable deployment option. However, the free version is limited to a maximum of 5 applications, 25 active hours per month, and 1 GB of memory per application. Paid versions offer additional resources, but with an important cost. To overcome these limitations, we have deployed the application on an institutional server, making IIDEA accessible to a broader audience [5]. We chose to use CNRS servers, which can be accessed through PLM, and deployed the application using the Kubernetes (Luksa, 2017) framework. This approach leverages the institutional infrastructure, aligning with our objective of avoiding reliance on private solutions. Utilizing PLM tools provides access to additional resources at no extra cost, which is beneficial for maintaining the performance and availability of the IIDEA application.

However, when we deployed IIDEA on the CNRS servers, there were few applications shiny in place, implying limited documentation at the time. This lack of documentation posed problems in terms of web programming skills for `shiny` interface developers. Indeed,

3. For example, for the selection made in Figure 3.5, the dedicated web page is https://www.ebi.ac.uk/QuickGO/term/GO:0007186
4. https://github.com/sanssouci-org/IIDEA
5. https://shiny-iidea-sanssouci.apps.math.cnrs.fr/

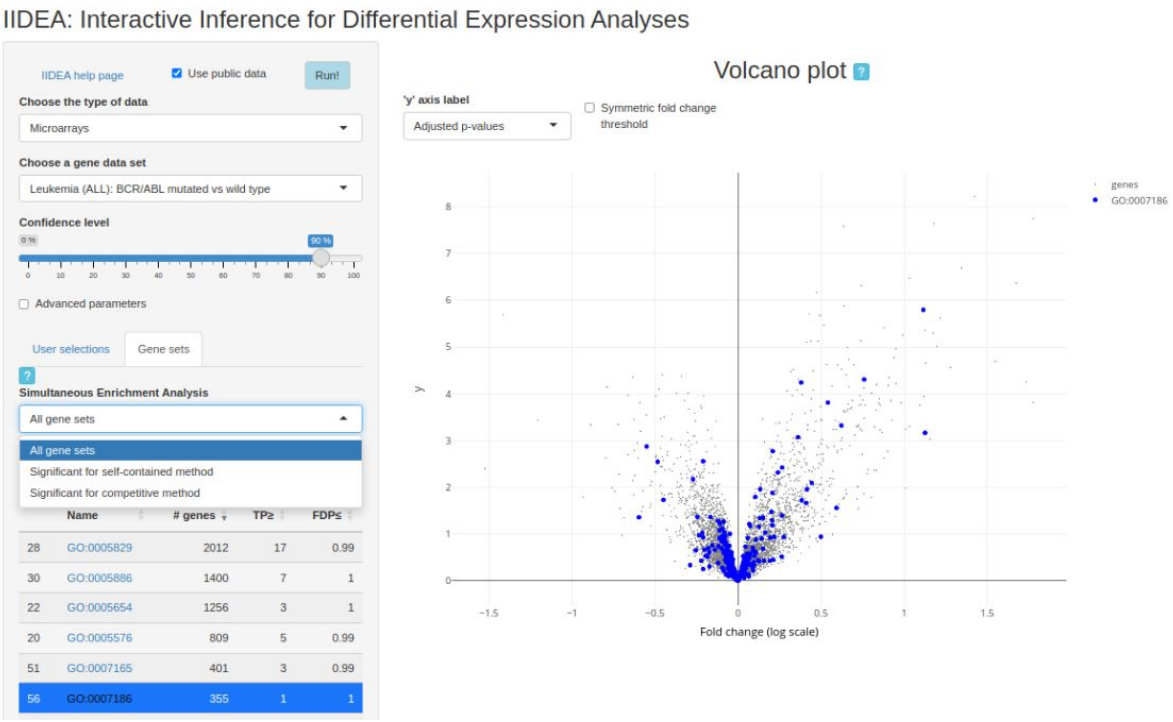

Figure 3.5 – Overview of IIDEA for the enrichment analysis. Interfaces from user selection and gene set analysis display the same elements. As gene sets are previously known, the volcano plot is not used for gene selection. Gene sets can be filtered according to self-contained and competitive assumptions.

environment variables (e.g., names, resources) and deployment parameters (e.g., strategies, number of instances), needed to be configured to deploy the application. In addition, manual configurations adapted to the needs of the applications were necessary. For example, packages hosted on `GitHub` or `Bioconductor` have to be installed manually in the container created for the application, rather than automatically like those from CRAN. The help provided by the tool's support team was invaluable in ensuring that the application was deployed correctly.

Recently, the SK8 team at INRAE developed a hosting solution for `shiny` applications (Maigné et al., 2023), also based on Kubernetes. This solution offers flexibility, is user-friendly for `shiny` developers, ensures stability in package installation, and simplifies long-term deployment management. A similar tool, accessible via PLM for CNRS agents, could provide an effective solution for deploying `shiny` applications and making research tools readily available to all scientists.

3.7 Conclusion

Post hoc methods provide statistical guarantees on flexible and multiple selections of statistical hypotheses of interest by the user. To facilitate interactive selections, the IIDEA application was developed. Designed for differential expression analysis of microarray and bulk RNAseq data, this application allows users to choose their genes of interest directly on a volcano plot, by adjusting the volcano plot thresholds or manually selecting genes using `plotly`. IIDEA provides post hoc guarantees on the selections made using the *Adaptive Simes* method. As such, IIDEA gives new visibility to this method and, generally, to the post hoc methods in DE studies for biologists and bioinformaticians.

At the start of the conception and development of IIDEA, no other interactive application existed that integrated the post hoc method into DE studies. Concurrently, Ebrahimpoor and Goeman (2021) developed an interactive R Shiny application that offers similar statistical guarantees for gene sets selected by threshold on a volcano plot. However, threshold selection does not occur directly on the volcano plot, making it more challenging to define these thresholds. Additionally, this application takes as input a matrix of p-values and fold changes, limiting the adaptability of post hoc methods using only the Simes and ARI method proposed by Goeman and Solari (2011) and Goeman et al. (2019). Therefore, IIDEA offers a more intuitive interface using the *Adaptive Simes* method while providing less conservative post hoc bounds. This approach facilitates data manipulation while ensuring statistical robustness, positioning IIDEA as an advanced interactive differential gene expression analysis solution.

Chapter 4

Perspectives on the application of post hoc methods

In differential expression analyses, procedures controlling the FDR do not give guarantees on genes sets selected on the volcano plot, due to the double filtering on p-values and fold change, as discussed in Section 1.2.3. Part I discusses and popularizes the application of post hoc inference methods introduced by Blanchard et al. (2020) to transcriptomic data. In Chapter 2, extensive numerical experiments based on transcriptomic data illustrate the performance of the *Adaptive Simes* method introduced by Blanchard et al. (2020). A generic linear time complexity to compute post hoc bound is also described in this chapter. These developments have enabled the creation of an interactive application called IIDEA, which is described in Chapter 3. This application aims to make it easier for biologists and bioinformaticians to apply the *Adaptive Simes* method in a controlled environment to transcriptomic data (including data from microarrays and bulk RNAseq data). The application allows users to select a set of genes of interest directly on an interactive volcano plot. Post hoc bounds corresponding to the selection are displayed instantly.

Section 4.1 explores existing developments (published after Enjalbert-Courrech and Neuvial (2022)) and potential improvements for the IIDEA application. Section 4.2 discusses a perspective of creation of an interactive application for neuroimaging data.

4.1 Improvement of IIDEA

The IIDEA application is a differential expression analysis software coded in R with the package `shiny`, based on functions from the R package `sanssouci`, which implements the *Adaptive Simes* method. Currently, this application is limited to comparing two groups of individuals. After discussions with bioinformaticians from the GenPhySE laboratory at INRAE, they are also interested in paired tests with a single group. Blanchard et al. (2020) describe the post hoc procedure for addressing the paired one-group test. Therefore, developing a new feature in IIDEA to include this test would be feasible and useful.

Multivariate linear models. Biologists may need to compare more than two groups or include multiple covariates simultaneously. Davenport et al. (2022) have generalized the calibration Blanchard et al. (2020) to multivariate linear models in which the expression of each gene is explained by a fixed set of covariates (possibly including groups of observations). For each gene, one or more tests are performed based on contrasts of interest in the corresponding linear model. As the permutation-based framework used by Blanchard et al. (2020) to calibrate post hoc bounds is not valid in this setting, Davenport et al. (2022) have introduced an alternative calibration method based on bootstrapping the residuals of the linear model to approximate the p-value distribution of under the null hypothesis. This method is implemented in Python. Its applicability is illustrated on both neuroimaging data and transcriptomic (microarray) data Bahr et al. (2013). Incorporating this method into IIDEA is an interesting perspective which requires additional developments. Since the linear model formulation encompasses the one-group and two-group tests, the multiple linear method of Davenport et al.

(2022) could in theory replace the permutation-based calibration currently implemented in IIDEA. However, we expect this method to be slower than the entirely permutation-based approach currently implemented: a detailed comparison of statistical performance and computation time is thus required. From a user's perspective, the implementation should be flexible so that they can select contrasts to be tested in the case of multiple covariates settings.

Learned template: sharper post hoc bounds. Post hoc methods use a family of thresholds $\mathbf{t} = (t_k)_{k\in[|K|]}$ that control the JER($\mathbf{t}$) at level α (see Equation (2.2)). To circumvent the conservativeness due to the dependence between the p-values, the Simes template is calibrated to find the largest value of λ such that JER($\mathbf{t}^S(\lambda)$) $\leq \alpha$ (see Section 2.2 for more details). Here, only λ is optimized for a given template shape. Blain et al. (2022) have proposed to learn a template from real data as follows. Let $\mathbf{X}^{train}$ be an independent dataset of the same type as $\mathbf{X}$, the studied dataset. Perform B^{train} permutations of the labels and calculate the p-values (see Step 1 of the calibration algorithm described in Algorithm 1.4 and illustrated in Figure 1.4). For each permutation $b \in [|B^{train}|]$, the b/B^{train} quantile associated to the sorted p-value forms the learned templates. The calibration is then performed on the inference dataset $\mathbf{X}$, by finding the largest b such that the associated JER is empirically controlled.

This method is already implemented (in Python) for both one-sample and two-sample tests. While its application is predominantly demonstrated on functional Magnetic Resonance Imaging (fMRI) data, Blain et al. (2022) discuss its adaptability for transcriptomic data, highlighting its broad utility across diverse datasets. Nevertheless, it is interesting to study the parameters and their impacts on the results for transcriptomic data as highlighted by Blain et al. (2022) for fMRI data. For neuroimaging data, they suggest to take $B^{train} =$ 10,000. Another important parameter is the size K of the template. Optimal choices for these parameters may be influenced by the larger voxel count in fMRI data (approximately 50,000) compared to transcriptomic data (around 10,000 genes tested).

A limitation of the method is that the associated statistical guarantees have been obtained when the train data set $\mathbf{X}^{train}$ is independent from $\mathbf{X}$. For the implementation of this method in IIDEA, a practical strategy would be to fix the training data set once and for all. Blain et al. (2022) have observed that the data used at the learning step affects the statistical performance. To select an appropriate training dataset for transcriptomic data analysis, a numerical evaluation of the method using a benchmark of transcriptomic datasets to identify those yielding less conservative post hoc bounds would be necessary. In order to avoid relying on an external data set, an interesting statistical perspective is to design alternative template learning methods for which X is used for both at the train and inference steps.

4.2 Interactive interface for fMRI data

Post hoc methods have applications beyond transcriptomics, including the analysis of fMRI data, which is a common use case for post hoc methods (Rosenblatt et al., 2018; Andreella et al., 2023; Blain et al., 2022; Davenport et al., 2022; Goeman et al., 2023; Vesely et al., 2023; Blain et al., 2024). fMRI data analysis aims to select connected regions of the brain that are activated by a specific task, and post hoc methods can provide statistical guarantees for such selections. An application similar in spirit to IIDEA could be created to facilitate the interactivity of fMRI data analysis.

The interactive interface could offer a brain mapping feature, allowing users to select a set of voxels representing a region of interest in the brain. These regions could be chosen based on the p-values of each voxel and other data-derived information (such as the mean difference in expression between the two compared groups, similar to the log fold change used

in transcriptomics). It would also be interesting to incorporate regions of interest (ROIs) defined by brain atlases. The application would then provide post hoc bounds for the selected regions.

The analysis of fMRI data typically uses the Python programming language, as the dedicated packages like `nilearn` (Abraham et al., 2014). Moreover, the method proposed by Blanchard et al. (2020) is implemented in the Python package `sanssouci.python`, allowing easy integration of methodological developments around the learned template (Blain et al., 2022) and multiple contrast settings (Davenport et al., 2022). Therefore, the development of an interactive interface would naturally occur in Python.

Similar to the R package `shiny` (Chang et al., 2021), its Python counterpart exists under the same name. The most widely used Python package for creating web interfaces is `dash`, maintained by Plotly, which facilitates easy integration of interactive graphics. One of these packages can be used to develop this application. However, given the large dimensions of fMRI data, development and deployment robustness must be considered.

Appendix A

Appendix of post hoc inference part

A.1 Technical results

A.1.1 Proof of Proposition 1 (interpolation-based post hoc bound)

In this proof we simply reproduce the argument used in the original proof of (Blanchard et al., 2020, Proposition 2.3) in the particular case studied in Chapter 1

We denote by $\mathscr{H}_0 \subset \{1, \dots, m\}$ the (unknown) subset of true null hypotheses. Then for a given subset S of genes, the number of false positives in S can be written as $\mathrm{FP}(S) = |S \cap \mathscr{H}_0|$.

Proof of Proposition 2. The proof relies on the following simple observation: for any subsets S and R of $\{1, \dots, m\}$, we have

$$\begin{aligned} |S \cap \mathscr{H}_0| &= |S \cap R^c \cap \mathscr{H}_0| + |S \cap R \cap \mathscr{H}_0| \\ &\leq |S \cap R^c| + |R \cap \mathscr{H}_0|. \end{aligned}$$

Let $R_k = \{j \in \{1, \dots, m\}, p_j < t_k\}$ be the set of genes whose p-value is less than t_k for $1 \leq k \leq K$. Then, we note that

$$(i)\ |S \cap R_k^c| = \sum_{j \in S} \mathbb{1}_{\{p_j \geq t_k\}}$$

$$(ii)\ |R_k \cap \mathscr{H}_0| \leq k - 1 \iff q_k \geq t_k$$

By Equation (2.2), it holds with probability greater than $1 - \alpha$ that for all $1 \leq k \leq |\mathscr{H}_0| \wedge K$, $q_k \geq t_k$. Therefore, there exists an event of probability greater than $1 - \alpha$ such that for any $S \subset \{1, \dots, m\}$,

$$\forall 1 \leq k \leq |\mathscr{H}_0| \wedge K, \quad |S \cap \mathscr{H}_0| \leq \sum_{j \in S} \mathbb{1}_{\{p_j \geq t_k\}} + k - 1$$

which concludes the proof. □

A.1.2 Calibration algorithm

The calibration algorithm described in Section 2.2.2 and illustrated in Figure 1.4 is formalized in Algorithm 3. Inputs of the algorithm are the expression matrix $\mathbf{X} \in \mathbb{R}^{n \times m}$, the vector of contrast $c \in \{0, 1\}^n$, the number of permutation $B \in \mathbb{N}$, the function p computing the p-values, a vector of the inverse of a reference family $(\tau_k^{-1})_{k=1 \dots K}$ and an error risk α.

A.1.3 Validity of Algorithm 3 (linear time interpolation bound)

A first step is to rewrite the bound $\overline{\mathrm{FP}}_\alpha$ on arbitrary subsets of S as a minimum over $|S|$ items:

Lemma 1. *Let $S \subset \{1, \dots, m\}$. Then for any $R \subset S$,*

$$\overline{\mathrm{FP}}_\alpha(R) = |S| \wedge \min_{1 \leq k \leq |S|} v_k(R)\,, \tag{A.1}$$

where $v_k(R) = \sum_{j \in R} \mathbb{1}_{\{p_j \geq t_{k \wedge K}\}} + k - 1$.

Algorithm 3 Calibration

Require: $X, c, B, p, (\tau_k^{-1})_{k=1\dots K}, \alpha$

for $b \leftarrow 1$ to B **do** ▷ 1. Permutation p-values:
  $c^b \leftarrow \text{sample}(c)$ ▷ $O(mB(n+\log(m)))$
  for $j \leftarrow 1$ to m **do**
    $P[b,j] \leftarrow p(X[,j], c^b)$
  end for
  $P_0[b,] \leftarrow \text{sort}(P[b,])$
end for
for $b \leftarrow 1$ to B **do** ▷ 2. Pivotal statistics:
  for $k \leftarrow 1$ to K **do** ▷ $O(BK)$
    $S[b,k] \leftarrow \tau_k^{-1}(P_0[b,k])$
  end for
  $\psi[b] \leftarrow \min(S[b,])$
end for
$\lambda \leftarrow \text{quantile}(\psi, \alpha)$ ▷ 3. Quantile: $O(B)$
return λ

Proof of Lemma 1. Let $R \subset S \subset \{1,\dots,m\}$. Let $v_k^0(R) = \sum_{j\in R} \mathbb{1}_{\{p_j \geq t_k\}} + k - 1$ for $k = 1,\dots,K$. With this notation, we have

$$\begin{aligned}
\overline{\text{FP}}_\alpha(R) &= \min_{1\leq k\leq K} v_k^0(R) \\
&= |S| \wedge \min_{1\leq k\leq K} v_k^0(R) && \text{(A.2)} \\
&= |S| \wedge \min_{1\leq k\leq |S|\wedge K} v_k^0(R) && \text{(A.3)} \\
&= |S| \wedge \min_{1\leq k\leq |S|\wedge K} v_k(R) && \text{(A.4)} \\
&= |S| \wedge \min_{1\leq k\leq |S|} v_k(R)\,. && \text{(A.5)}
\end{aligned}$$

Above, Equation (A.2) holds since $v_1^0(R) = \sum_{j\in R} \mathbb{1}_{\{p_j\geq t_1\}} \leq |R| \leq |S|$. Equation (A.3) is obvious if $|S| \geq K$; if $|S| < K$ then for $k > |S| \wedge K$, $v_k^0(R) \geq k - 1 \geq |S|$. Equation (A.4) holds since $v_k(R) = v_k^0(R)$ for $k \leq |S| \wedge K$. Finally, Equation (A.5) is obvious if $|S| \leq K$; if $|S| > K$ then for $k \geq |S| \wedge K (= K)$, we have $t_{k\wedge K} = t_K$ which implies that $v_k(R) = (k-K) + v_k(R) \leq v_k(R)$. □

We are now ready to prove Proposition 3.

Proof of Proposition 3. We consider the nested subsets S_j for $j = 1,\dots,s$. By Lemma 1, we have

$$\overline{\text{FP}}_\alpha(S_j) = s \wedge \min_{1\leq k\leq s} v_k(S_j)\,. \tag{A.6}$$

If $p_j \geq t_{s\wedge K}$, then $\kappa_j = s$, which implies that

$$\overline{\text{FP}}_\alpha(S_j) = \kappa_j \wedge \min_{1\leq k\leq \kappa_j} v_k(S_j)\,. \tag{A.7}$$

If $p_j \geq t_{s\wedge K}$, then $\kappa_j < s$. This implies that $v_k(S_j) \geq \kappa_j$ for all $k \in \{\kappa_j + 1,\dots,s\}$, and $v_{\kappa_j+1}(S_j) = \kappa_j$. Therefore, we have

$$\min_{\kappa_j < k\leq s} v_k(S_j) = \kappa_j\,,$$

so that Equation (A.7) holds as well. We conclude by noting that for $k \leq \kappa_j$,

$$\begin{aligned} v_k(S) - v_k(S_j) &= \sum_{j' \in S \setminus S_j} \mathbb{1}_{\{p_{j'} \geq t_{k \wedge K}\}} \\ &= |S \setminus S_j| = s - j \,, \end{aligned}$$

since for $j' \in S \setminus S_j, p_{j'} \geq p_j > t_{\kappa_j \wedge K} \geq t_{k \wedge K}$. □

A.2 Numerical results for the BLCA data set

A.2.1 Comparison between existing post hoc bounds

This section complements the results presented in Section 2.4, by providing the results of the (non-adaptive) Simes method and the single-step adaptive Simes method (which corresponds to the single-step pARI method). All methods are described in Section 2.5.1. This figure illustrates the fact that for DE studies, the adaptation to dependency provided by the calibration method described in Section 2.2.2 (black vs gray curves) yields a more substantial improvement than the adaptation to the proportion of true null hypotheses (dashed vs solid curves).

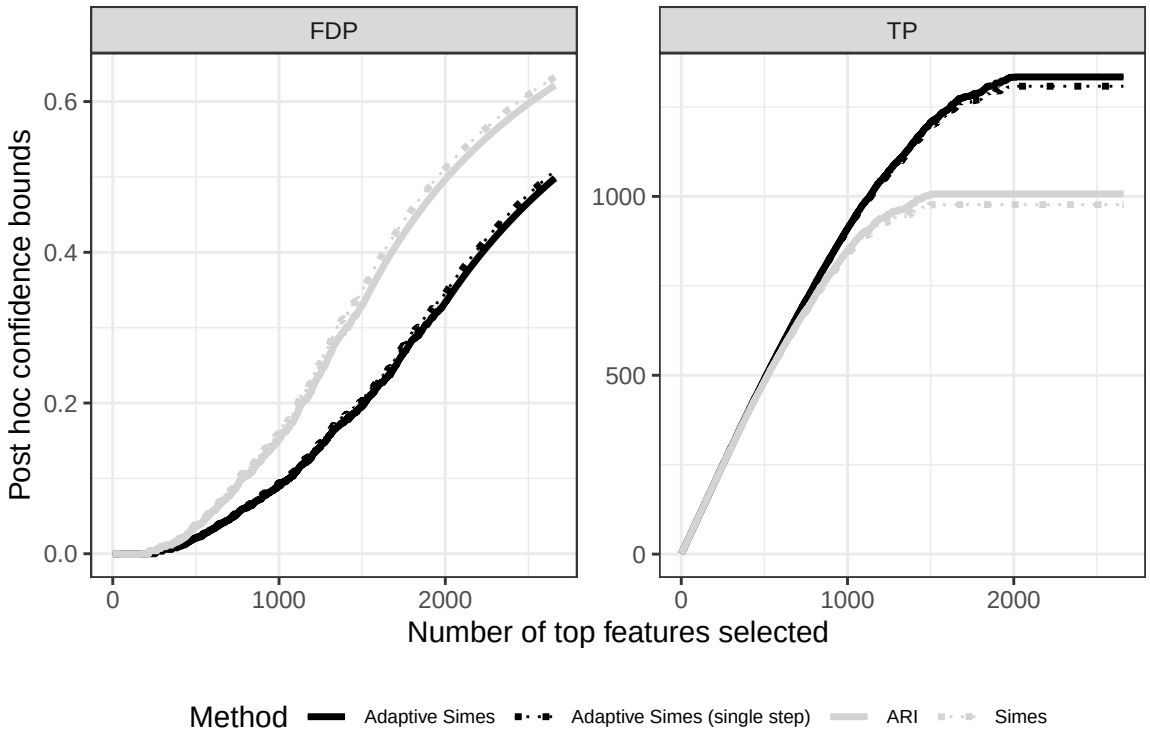

Figure A.1 – 90% confidence curves on "top k" lists for the Urothelial bladder carcinoma data set. Left: upper bound on the False Discovery Proportion (FDP); right: lower bound on the number of true positives (TP). Adaptive methods (black curves) outperforms non-adaptive ones (gray curves).

	Confidence curve (Fig. 2.1)			Volcano plot (Fig 2.2)		
	$\lvert S\rvert$	$\overline{\mathrm{TP}}_\alpha(S)$	$\overline{\mathrm{FDP}}_\alpha(S)$	$\lvert S\rvert$	$\overline{\mathrm{TP}}_\alpha(S)$	$\overline{\mathrm{FDP}}_\alpha(S)$
Adaptive Simes	1042	958	0.1	569	492	0.135
Adaptive Simes (single step)	1042	938	0.1	569	490	0.139
ARI	781	703	0.1	569	456	0.199
Simes	757	682	0.1	569	452	0.206

Table A.1 – Post hoc bounds on BLCA data set for the four compared methods. Left panel: Gene selections from Figures 2.1 and A.1, with target FDP set to 0.1. Right panel: gene selection for the volcano plot (Figure 2.2).

A.2.2 Comparison between limma-voom and Wilcoxon p-values

As explained in Section 2.2.2, the theoretical results underlying our calibration method require the test statistic (or p-value) for each gene to depend on the data only via the expression measurements associated with this gene. However, the most commonly used statistical tests in DE studies with RNAseq data (`DEseq2` (Love et al., 2014), `edgeR` (Chen et al., 2014) and limma-voom (Smyth, 2004)) do not formally meet this assumption. In particular, these methods are using moderated variance estimators that borrow information from all genes, which is crucial when dealing with very low sample sizes. Let us emphasize that, by the very nature of post hoc bounds and is illustrated in Figure 2.2, our methods can still be used to evaluate the number of false positives in any gene selection, *including selections obtained from the above-mentioned methods*. Figure A.2 provides a graphical comparison between the p-values obtained by the limma-voom and the Wilcoxon rank sum test for the BLCA data set. It illustrates the consistency between these p-values.

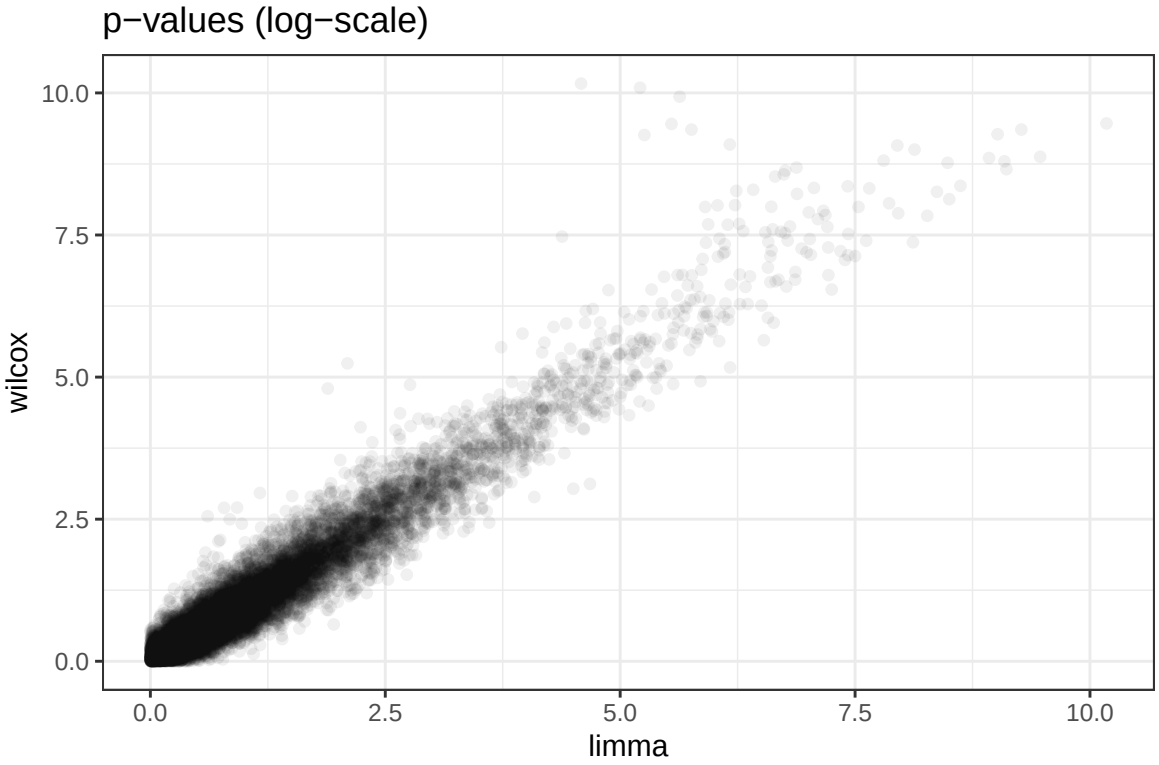

Figure A.2 – Comparison between limma-voom and Wilcoxon p-values on the BLCA data set.

A.2.3 Comparison of the execution time of the limma-voom method and the Wilxocon test

A current reason for using the Wilcoxon test in IIDEA is its computational efficiency compared to standard methods developed for omics data (`limma-voom`, `DEseq2`, and `edgeR`) for comparing two groups of individuals. These standard methods implement specific computational procedures for omics data, which can potentially increase their computation time. For the limma-voom methods, each contrast tested follows a linear model, allowing simultaneous calculation of p-values. Conversely, the Wilcoxon test is applied directly to the data for each gene tested. Thus, a naive implementation involves looping over the genes to obtain the desired m p-values. A solution to facilitate this calculation is to perform matrix operations between a vector coding for the observation groups and the expression matrix. This solution is implemented in the `rowWilcoxonTest` function of the `sanssouci` package.

We constructed a simulation setting to compare the execution times of the methods. Using the BLCA dataset, we selected the first m genes from the dataset with $m \in \{10, 20, 50, 100, 200, 500\}$. m tests were performed for a single contrast. The compared methods were the standard use of limma-voom, the Wilcoxon test with a naive loop-based implementation, and the matrix-based version of Wilcoxon. Figure A.3 shows the time as a function of m for $B = 1$. The matrix-based calculation is the fastest for the tested values of m. The limma-voom method is slower but is caught up by the iterative Wilcoxon test for $m \geq 100$, resulting in the same execution time.

Furthermore, the post-hoc bounds involve computing B permuted p-values to estimate the

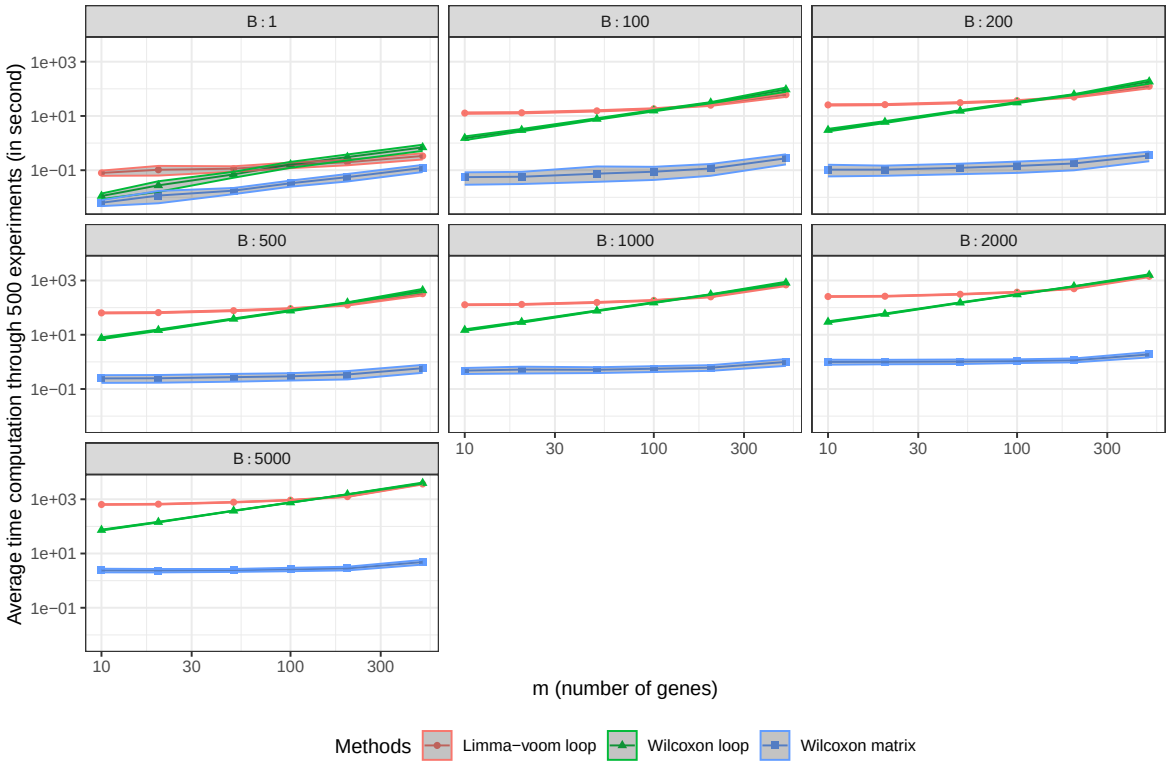

Figure A.3 – Computation time of limma-voom and Wilcoxon tests in function of the number of realized on 500 simulations. The limma-voom loop method iterates on the B permutations. The Wilcoxon loop algorithm interacts on the B permutations and the m genes. The Wilcoxon matrix algorithm only use matrix operations and none loop.

data distribution of the joint p-value under the null hypothesis, adding complexity to execution time. For the limma-voom method, the current naive implementation iteratively loops over the permutations of contrasts to obtain the desired $B \times m$ p-values. The naive version of the Wilcoxon test involves an additional loop to compute all permutations. The matrix-based implementation allows for providing a permutation matrix, enabling matrix computation of the test without iterative loops.

The same simulation setting is used, adding $B \in \{100, 200, 500, 1000, 2000, 5000\}$ permutations performed. Figure A.3 shows the results. The computation time increases with B. The same interpretations regarding the methods' order based on their execution time hold. Additionally, the execution time appears to become linear for the Wilcoxon test after a certain number of permutations and remains significantly faster than the other two tests. Therefore, the current use of the Wilcoxon test provides a 100x speed up, ensuring a responsive interface to user requests.

A.3 Power assessment for RNAseq data

For a given subset S of genes, the power of a method providing the post hoc bound $\overline{\mathrm{TP}}_\alpha$ is defined in Blanchard et al. (2020) as

$$\mathbb{E}\left(\frac{\overline{\mathrm{TP}}_\alpha(S)}{\mathrm{TP}(S)}\middle|\, \mathrm{TP}(S) > 0\right). \tag{A.8}$$

This quantity corresponds to the expected proportion of signal in S actually recovered by the method considered. Again, it is estimated by its empirical counterpart, that is, the average proportion of signal in S recovered over those experiments for which some signal was actually present in S. We considered four different gene selections S for estimating the power as defined in Equation (A.8):

BH_05: the set genes selected by the Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995) at level 5%

first_100: the 100 genes with lower p-value

p_05: the genes whose p-value is less than 5%

H: all genes in the data set.

Figure A.4 displays the empircal power for the following choice of parameters: $\pi_0 = 0.8$ and $SNR \in \{1.5, 3\}$. The power of the Adaptive Simes methods is higher than the one of parametric Simes methods. This is coherent with the increased tightness of the Adaptive Simes bounds relative to their parametric counterpart already observed in Figure 2.5. We also note than the adaptivity to π_0 does not make a difference in terms of power: the Simes and ARI methods are essentially indistinguishable from each other, and the same holds for the single-step and step-down Adaptive Simes methods.

A.4 Performance evaluation for microarray data

We consider the GSE19188 data set (Hou et al., 2010) available from GEO (Barrett et al., 2012) and the R package `GSEABenchmarker` (Geistlinger et al., 2020). This data set consists of 91 non-small cell lung cancer tissue samples and 62 normal samples. Each of these observations corresponds to a vector of $m = 21{,}407$ gene expression values.

The parameters of the experiments have been set as follows. The proportion of null genes is set to $\pi_0 \in \{0.5, 0.8, 1\}$. We have considered an additive signal for differential expression:

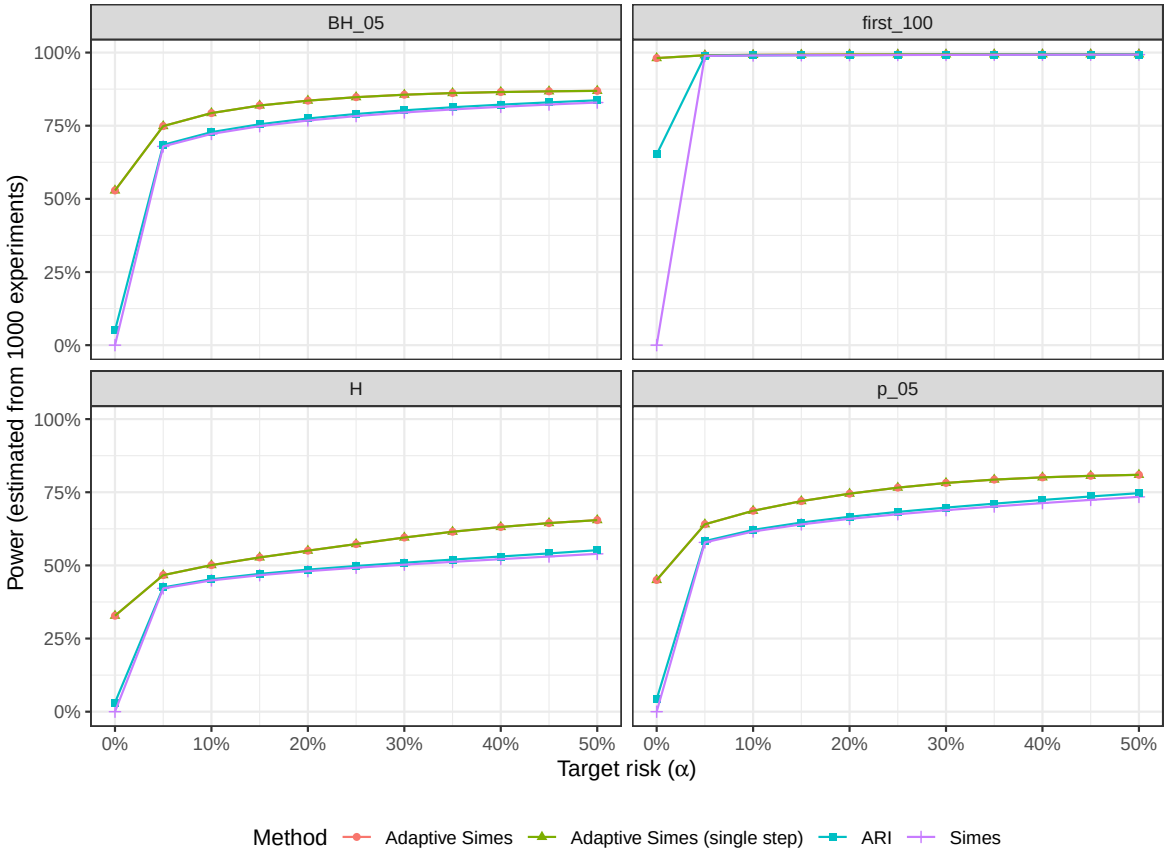

Figure A.4 – Empirical power of four methods, as a function of the target risk (JER) level. Simes-based methods are outperformed by their adaptive couterpart. Each panel corresponds to a different gene selection (as described in main text).

the expression level for the m_1 non-null genes are then shifted by a constant value for n_1 of the n observations. We have used a two-sided Welch test for comparing the two groups.

The results are summarized by Figure A.5. The conclusions are identical to those obtained for RNA-seq-based experiments in Section 2.5.3: JER is controlled for all methods and all parameter combinations, and the risk for the Adaptive Simes methods is substantially closer to the target risk than for the parametric Simes methods (Simes and ARI), illustrating a substantial gain provided by the calibration method described in Section 2.2.2. The gain obtained from the adaptation to π_0 is negligible, except for the case with strongest and most dense signal ($\pi_0 = 0.5$ and SNR$= 5$), which is arguably the less realistic scenario for DE studies.

A.5 Influence of sample size

In this section, we assess the performance of our calibration method when a limited number of samples is available on both RNA-seq and microarray data.

A.5.1 RNA seq data

We report the results of a subsampling study based on the BLCA data set, with the number of subsampled observations is set to $n \in \{10, 50, 90\}$ with $n_0 = n_1 = n/2$. The settings

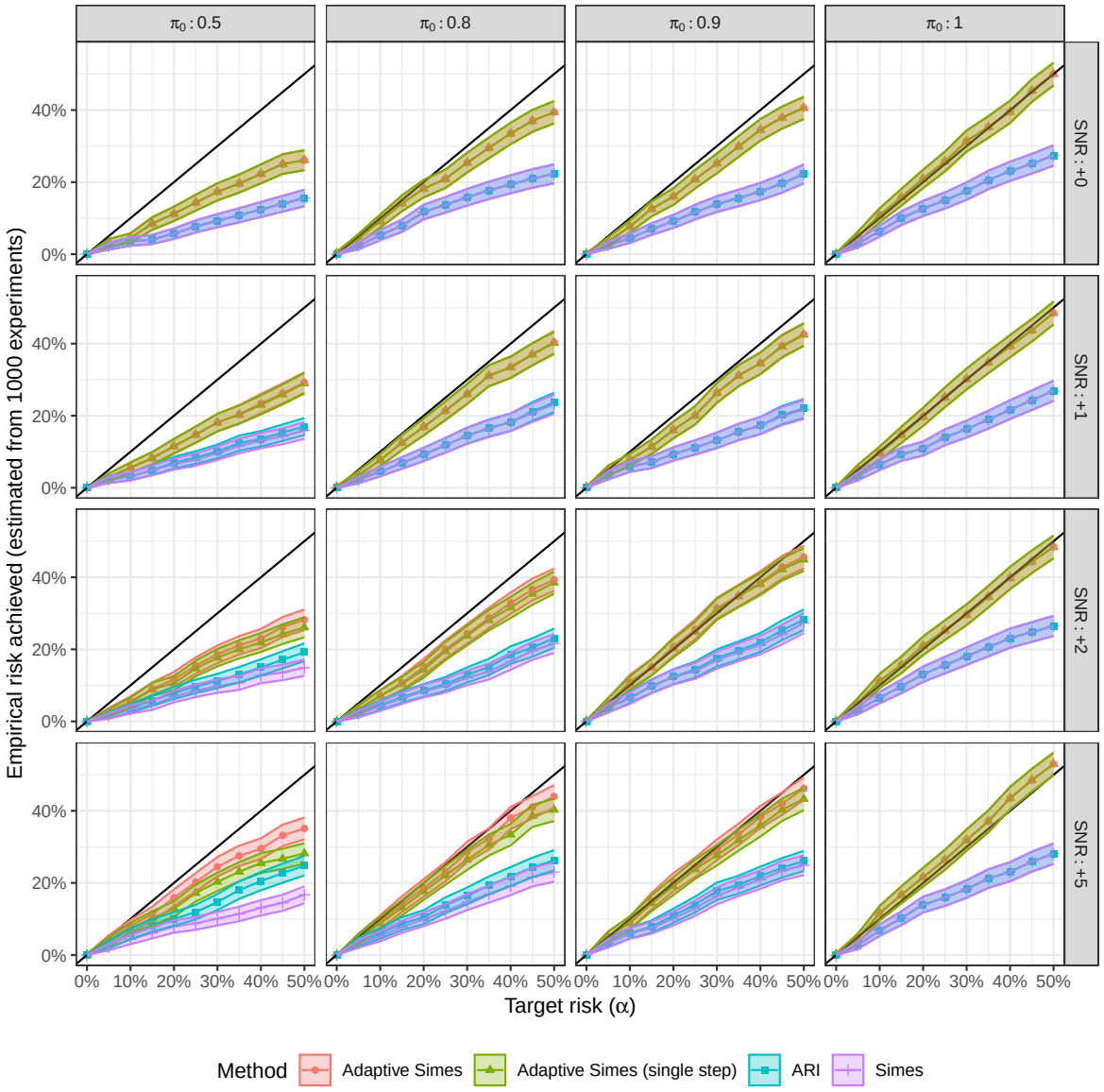

Figure A.5 – Validity and compared tightness of the post hoc bounds on microarray-based numerical experiments. The average empirical JER achieved across 1000 experiments is plotted (together with 95% confidence curves) against the target risk α for all considered methods. Each panel corresponds to a combination of the parameters π_0 and SNR.

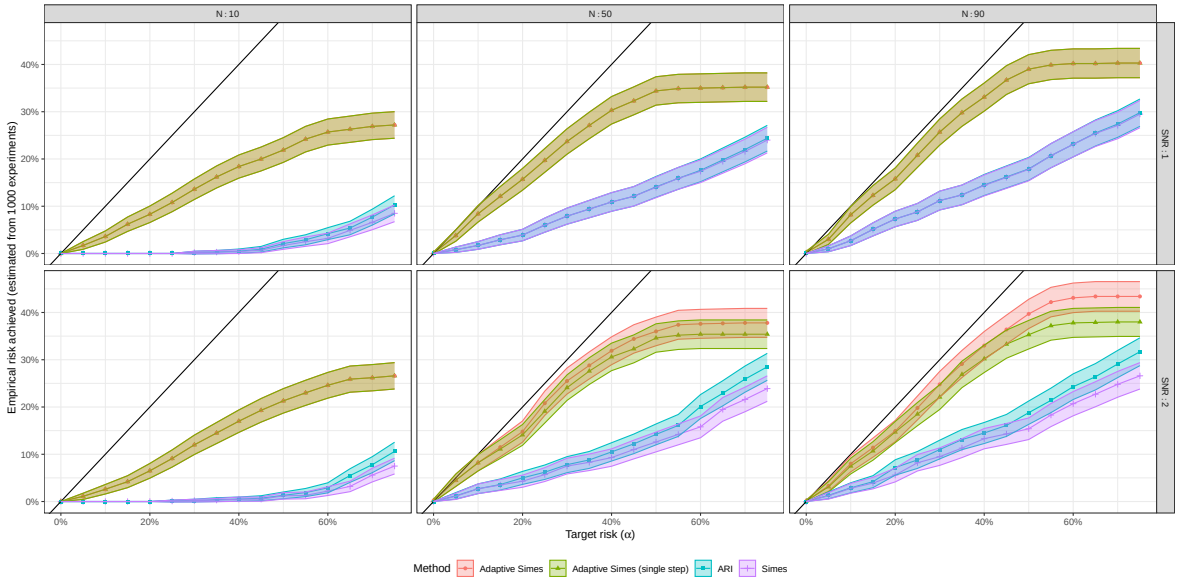

Figure A.6 – Control of the empirical risk for several sample sizes (RNA-seq data). The BLCA dataset is sampled by n observations and a multiplicative signal is added for 20% of genes, so that $\pi_0 = 0.8$. The Wilcoxon rank sum test is used for comparing the two groups.

are those described in Section 6 of the paper, with $\pi_0 = 0.8$. The results are displayed in Figure A.6 for JER control and Figure A.7 for power. The JER is controlled for all methods and parameter combinations, and the risk for the Adaptive Simes methods is substantially closer to the target risk than for the parametric Simes methods (Simes and ARI), illustrating a substantial gain provided by the calibration method described in Section 2.2.2. This is confirmed by the power plot. Note that for $n = 5$ the BH procedure did not select any gene, which is why the corresponding panel is empty.

A.5.2 Microarray data

We report the result of a subsampling study based on the lung cancer data described in Section A.4. Here, the number of subsampled observations is set to $n \in \{10, 30, 50\}$ with $n_0 = n_1 = n/2$ and $\pi_0 = 0.8$. The corresponding JER control results are displayed in Figure A.8. These results indicate that the proposed methods provide consistent improvement with respect to other methods.

A.6 Tests of association with a continous covariate

This section stems from an interesting suggestion from an anonymous reviewer. A natural extension to the multiple two-sample (and one-sample) testing setting considered in Chapter 2 is the case of tests of marginal association between the expression of each gene $\mathbf{X}_{.j}$ with a continuous covariate Y. If $(\mathbf{X}_{.j}, Y)$ are bivariate Gaussian, then tests of the null hypothesis $H_{0,i} : cor(\mathbf{X}_{.j}, Y) = 0$ based the Pearson correlation coefficient do satisfy the randomization assumption that is used in Blanchard et al. (2020) to prove the validity of the calibration procedure. If $(\mathbf{X}_{.j}, Y)$ are not bivariate Gaussian, then the calibration procedure is also valid with the same test statistic, for the test of the (stronger) null hypothesis that $\mathbf{X}_{.j}$ and Y are independent.

In order to illustrate this point, we have made numerical experiments based a synthetic

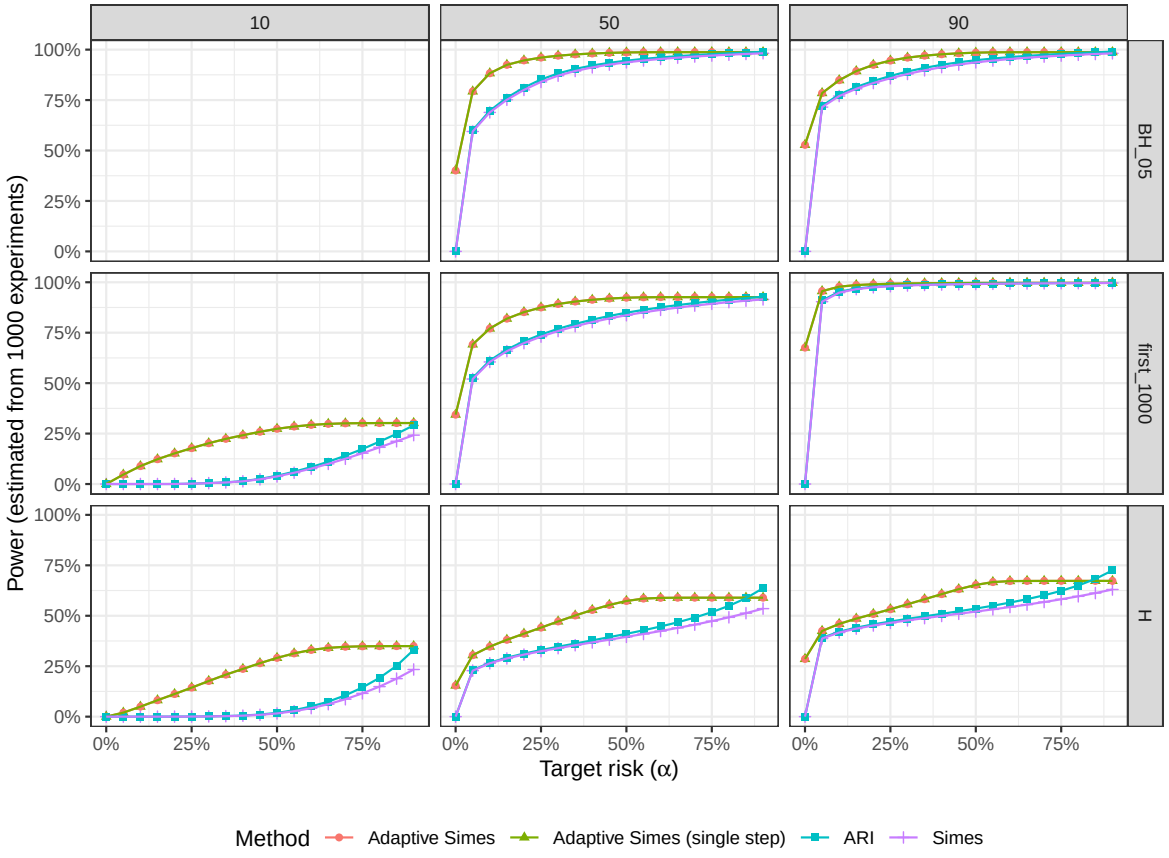

Figure A.7 – Empirical power for several sample sizes (RNA-seq data). The BLCA dataset is sampled by n observations and a multiplicative signal is added to 20% of the genes in one of the groups, leading to $\pi_0 = 0.8$. The Wilcoxon rank sum test is used for comparing the two groups.

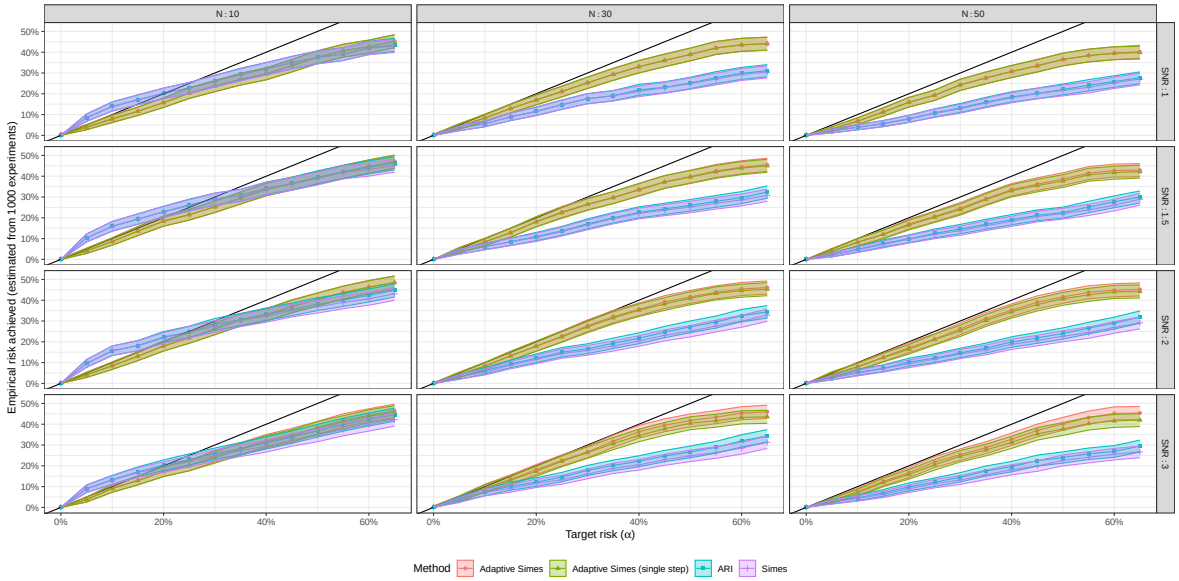

Figure A.8 – Control of the empirical risk for several sample sizes (microarray data). The lung cancer dataset is sampled by n observations and an additive signal is added for 20% of genes in one of the groups, leading to $\pi_0 = 0.8$. The Welch test is used to compare the two groups.

data set of $n = 60$ samples for $m = 10{,}000$ gaussian features (genes) generated as follows. We simulate a multivariate normal distribution $X \in \mathbb{R}^{n \times m}$ with mean $\mu = 0$ and a block-diagonal covariance matrix Σ with 100 blocks. The block sizes are drawn uniformly. The pairwise correlation between any two features from the same block is set to $\rho \geq 0$. The continuous outcome is generated by a linear model of the form $y_i = X_{ij}\delta_j + \epsilon_i$, where $\delta_i = 1$ if gene j is truly associated with the outcomes and $\delta_i = 0$ otherwise. The vector δ is constant within blocks, and designed in such a way that the proportion of non associated genes (that is, the proportion of null items in δ) is controlled (in expectation) by the parameter $\pi_0 \in [0, 1]$. For each gene j, the test statistic is based on the Pearson correlation coefficient between $\mathbf{X}_{.j}$ and Y. 1000 simulations have been performed for each value of $\rho \in \{0, 0.2, 0.4, 0.6\}$ and $\pi_0 \in \{0.8, 0.95, 0.99\}$.

Figure A.9 reports the results for $\rho > 0$ since the results for $\rho = 0$ (independence) are close to those shown for $\rho = 0.2$. We obtain similar conclusions as for the other numerical experiments in the manuscript: the JER is controlled for all methods and all parameter combinations, and the risk for the Adaptive Simes methods is closer to the target risk than for the parametric Simes methods (Simes and ARI), illustrating a substantial gain provided by the calibration method described in Section 2.2.2. The gain obtained from the adaptation to π_0 is negligible.

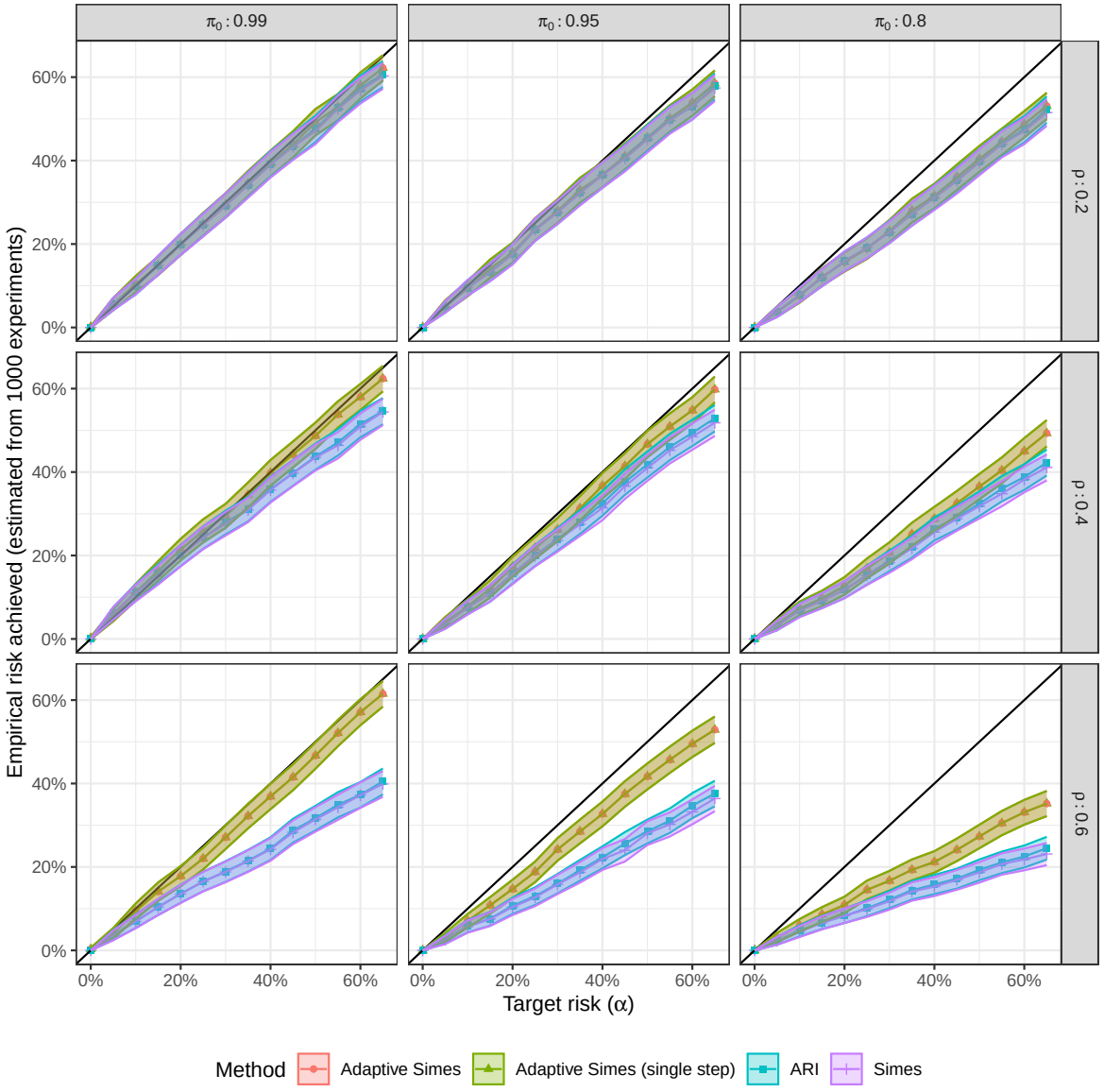

Figure A.9 – Validity and compared tightness of the post hoc bounds on synthetic data for testing the marginal association with a continuous covariate. The average empirical JER achieved across 1000 experiments is plotted (together with 95% confidence curves) against the target risk α for all considered methods. Each panel corresponds to a combination of the parameters π_0 and ρ.

A.7 Additional plots for IIDEA

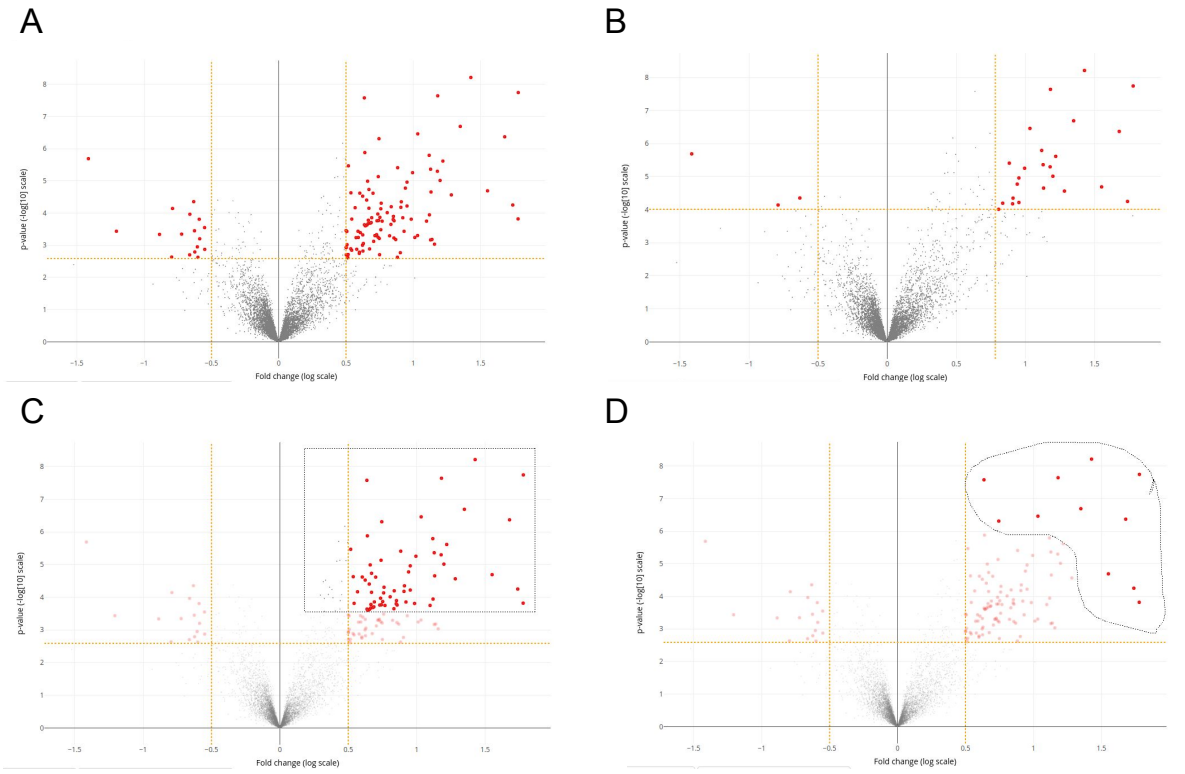

Figure A.10 – Example of interactive selection on IIDEA. **A** (upper left): A volcano plot with by default selection. **B** (upper right): Selection by dragging the orange line to select genes on both upper corners. **C** (lower left): Selection with a box proposed by `plotly`. **D** (lower right): Selection with a lasso proposed by `plotly`.

Part II

Post-clustering inference

Chapter 5

Multivariate methods: review and numerical comparison

5.1 Introduction

The problem of post-clustering inference arises from using the same dataset to (i) estimate a clustering partition of individuals, and (ii) test a hypothesis stemming from the clustering step. As explained in Section 1.3.3, ignoring the clustering step when performing the test leads to violations of the type I error rate control.

Let $\mathbf{X} = (X_i)_{i\in[|n|]}$ be a $n \times m$ matrix of samples. We assume that, for each $i \in [|n|]$, $X_i \sim \mathcal{N}_m(\mu_i, \Sigma)$ with $\mu_i \in \mathbb{R}^m$ the vector of means and the covariance matrix $\Sigma \in \mathbb{R}^{m\times m}$, and X_i are independent. The mean vectors μ_i are grouped together in the matrix $\boldsymbol{\mu} = (\mu_i)_{i\in[|n|]} \in \mathbb{R}^{n\times m}$. The test compares the behavior of $\boldsymbol{\mu}$ between some groups, using tools of statistical testing described in Section 1.2.1. In this chapter, we focus on the comparison between two multivariate clusters.

Firstly, we define the test for two groups G_k and $G_{k'}$ defined before seeing data. The null hypothesis is

$$\mathcal{H}_0 : \eta(G_k, G_{k'})^\top \boldsymbol{\mu} = 0_m, \tag{5.1}$$

for a contrast vector $\eta(G_k, G_{k'})$ that only depends on the groups G_k and $G_{k'}$. Let $\mathcal{T}(\eta, \mathbf{X})$ be a test statistic that assesses the null hypothesis in Equation (5.1) and depends on the contrast vector η and the dataset $\mathbf{X}$. The associated p-value for the observed data $\mathbf{x}$ is

$$p(\mathbf{x}) = \mathbb{P}_{\mathcal{H}_0}\left(\mathcal{T}(\eta(G_k, G_{k'}), \mathbf{X}) \geq \mathcal{T}(\eta(G_k, G_{k'}), \mathbf{x})\right) \tag{5.2}$$

and the test controls the type I error rate, as described in Section 1.2.1. Example 1 defines the comparison of mean between two clusters using the Wald's test statistic.

Example 1 (Comparison of mean between two groups)**.** *Let G_k and $G_{k'}$ be two groups. For the mean comparison of two groups, the contrast vector is defined as*

$$\eta_i(G_k, G_{k'}) = \frac{\mathbb{1}_{\{i\in G_k\}}}{|G_k|} - \frac{\mathbb{1}_{\{i\in G_{k'}\}}}{|G_{k'}|}, \ \forall\ i \in [|n|]. \tag{5.3}$$

Thus, in this example, the null hypothesis in Equation (5.1) can be rewritten as

$$\mathcal{H}_0 : \bar{\boldsymbol{\mu}}_{G_k} = \bar{\boldsymbol{\mu}}_{G_{k'}} \tag{5.4}$$

with $\bar{\boldsymbol{\mu}}_{G_k} = \frac{1}{|G_k|}\sum_{i\in G_k} X_i$. Several test statistics exist to assess this null hypothesis, such as Hotelling (1931)s t-squared statistic. For this example, we will consider the Wald's test statistic, defined as $\mathcal{T}(\eta, \mathbf{X}) = \left\|\eta^\top \mathbf{X}\right\|_2$ which follows a scale Chi distribution under the null hypothesis. The expression of the p-value is

$$p(\mathbf{x}) = \mathbb{P}_{\mathcal{H}_0}\left(\left\|\eta(G_k, G_{k'})^\top \mathbf{X}\right\|_2 \geq \left\|\eta(G_k, G_{k'})^\top \mathbf{x}\right\|_2\right). \tag{5.5}$$

The distribution of the p-value in Equation (5.5) is known under the null hypothesis for a fixed contrast vector, i.e. the compared groups G_k and $G_{k'}$ are known before seeing the data. But in some applications, the contrast vector is *data-driven* since the groups are built from data. Let $\mathcal{C}$ be a clustering method and $\mathcal{C}(\mathbf{X}) = \{C_1(\mathbf{X}), ..., C_K(\mathbf{X})\}$ be the associated partition of $\mathbf{X}$ into K clusters. Let define the vector contrast $\eta(C_k(\mathbf{X}), C_{k'}(\mathbf{X}))$ on clusters $C_k(\mathbf{X})$ and $C_{k'}(\mathbf{X})$. Then the null hypothesis defined in Equation (5.1) is now

$$\mathcal{H}_0 : \eta(C_k(\mathbf{X}), C_{k'}(\mathbf{X}))^\top \boldsymbol{\mu} = 0_m. \tag{5.6}$$

Then, the naive p-value in Equation (5.2) becomes

$$p(\mathbf{x}) = \mathbb{P}_{\mathcal{H}_0} \left(\mathcal{T}(\eta(C_k(\mathbf{X}), C_{k'}(\mathbf{X})), \mathbf{X}) \geq \mathcal{T}(\eta(C_k(\mathbf{X}), C_{k'}(\mathbf{X})), \mathbf{x}) \right). \tag{5.7}$$

As discussed in Section 1.3.3, the null hypothesis in Equation (5.6) is random, and then the distribution of the statistic $\mathcal{T}(\eta(\mathcal{C}_k(\mathbf{X}), C_{k'}(\mathbf{X}))$ is unknown under the null hypothesis. As shown numerically by Gao et al. (2024) and Hivert et al. (2024a), and reported in Figure 1.5, these naive p-values are stochastically smaller than the uniform distribution illustrating that the 'Naive' test does not control the type I error rate.

To address this problem, several solutions have recently been published, which can be divided into two categories: 1) methods based on an information partitioning and 2) conditional approaches. For the first category, Leiner et al. (2023), Neufeld et al. (2024a) and Dharamshi et al. (2024) have introduced information partition solutions that split the information contained in $\mathbf{X}$ into two independent or conditionally independent datasets: the clustering is computed on the first dataset and the statistical test is performed on the second dataset. The methods in the the second category condition by the clustering event $\{C_k(\mathbf{x}), C_{k'}(\mathbf{x}) \in \mathcal{C}(\mathbf{X})\}$ to control the type I error rate. Among these methods, Gao et al. (2024); Chen and Witten (2023); Yun and Foygel Barber (2023) and González-Delgado et al. (2023) have addressed the question of mean comparison using the Wald's test described in Example 1 with the random contrast $\eta(C_k(\mathbf{X}), C_{k'}(\mathbf{X}))$. All of these publications have appeared in the last **three** years. Even though each has compared itself to the naive method, demonstrating its improvement, none have compared themselves to the others. Accordingly, our contribution consists of conducting a systematic review of state-of-the-art methods and then performing a numerical comparison of these methods to understand their limitations and possible extensions.

This chapter is organized as follows. Section 5.2 is devoted to a review of post-clustering inference methods, divided into the above two categories. A comparison through numerical experiments between methods under a known spherical covariance matrix assumption ($\Sigma = \sigma^2 I_m$) is addressed in Section 5.3. In Section 5.4, the numerical comparison is extended to the spherical case with unknown σ and to a specific non-spherical (auto-regressive) case.

5.2 Review of methods

5.2.1 Information partitioning

The first category of methods to address the post-clustering inference problems is based on the information partitioning. As mentioned in the introduction (Section 1.3.4), splitting individuals is not an adequate solution to the problem of post-clustering inference. In this section, we will examine how splitting the information contained in $\mathbf{X}$ can address this problem.

A perfect procedure would involve having two independent copies of each observation, one for the clustering procedure the second for the test. However, this scenario is rare in practice. To circumvent this issue, Leiner et al. (2023) and Neufeld et al. (2024a) propose methods for partitioning information to create two datasets $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$, each of them containing all

observations from a dataset $\mathbf{X}$. These methods enable the use of post-clustering inference processes (see Algorithm 4).

Algorithm 4 Information partitioning methods used for post-clustering inference

1: Split $\mathbf{X}$ to obtain $\mathbf{X}^{(1)} \in \mathbb{R}^{n\times m}$ and $\mathbf{X}^{(2)} \in \mathbb{R}^{n\times m}$
2: Compute the clustering on $\mathbf{X}^{(1)}$: $\mathcal{C}\left(\mathbf{X}^{(1)}\right)$
3: Perform a statistical test with contrast vector $\eta(\mathcal{C}(\mathbf{X}^{(1)}))$ using $\mathbf{X}^{(2)}$.

Leiner et al. (2023) propose the so-called *data fission* process which aims to create two data sets $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$ such as the distributions of $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}|\mathbf{X}^{(1)}$ are known and $\mathbf{X} = h(\mathbf{X}^{(1)}, \mathbf{X}^{(2)})$ for a given function $h()$. Note that, for a test that depends on the clustering on $\mathbf{X}^{(1)}$, the independence between the two data sets $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$ is not required if the distribution of $\mathbf{X}^{(2)}|\mathbf{X}^{(1)}$ is known. In the multivariate Gaussian context, let $X_i \sim \mathcal{N}_m(\mu_i, \Sigma)$ for each $i \in [|n|]$ and $Z_i \sim \mathcal{N}_m(0, \Sigma)$, with X_i and Z_i are independent. Then, $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$ are built such that

$$\begin{cases} X_i^{(1)} = X_i + \tau Z_i \sim \mathcal{N}_m(\mu_i, (1+\tau^2)\Sigma) \\ X_i^{(2)} = X_i - \tau^{-1} Z_i \sim \mathcal{N}_m(\mu_i, (1+\tau^{-2})\Sigma), \end{cases}$$

where $\tau > 0$ is called the fission parameter. This decomposition ensures that $X_i^{(1)}$ and $X_i^{(2)}$ are independent.

Neufeld et al. (2024a) have developed a related process called *data thinning*, valid for convolution-closed distributions (see Jørgensen and Song (1998) for more details) which include Gaussian, Poisson, Negative Binomial distributions among others (see Neufeld et al. (2024a) and Jørgensen and Song (1998) for a complete list). These convolution-closed distributions allow the decomposition of $\mathbf{X}$ into two data sets $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$ such that (i) $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$ are independent, (ii) $\mathbf{X} = \mathbf{X}^{(1)} + \mathbf{X}^{(2)}$ and (iii) $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$ follow the same type of distribution as $\mathbf{X}$. Dharamshi et al. (2024) generalize the procedure for other distributions by relaxing properties of distribution (iii) where $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$ follow the same distribution which can be different from the distribution of $\mathbf{X}$ and reconstruction (ii) with a deterministic and known function $h(.)$ such that $h(\mathbf{X}^{(1)}, \mathbf{X}^{(2)}) = \mathbf{X}$. For a multivariate Gaussian distribution where $X_i \sim \mathcal{N}_m(\mu_i, \Sigma)$, the *data thinning* process generates

$$\begin{cases} X_i^{(1)} | X_i = x_i \sim \mathcal{N}_m(\varepsilon x_i, \varepsilon(1-\varepsilon)\Sigma) \\ X_i^{(2)} = X_i - X_i^{(1)}, \end{cases}$$

where $\varepsilon \in (0,1)$ is the thinning parameter. This process provides two independent data sets with known distributions: $X_i^{(1)} \sim \mathcal{N}_m(\varepsilon\mu, \varepsilon\Sigma)$ and $X_i^{(2)} \sim \mathcal{N}_m((1-\varepsilon)\mu, (1-\varepsilon)\Sigma)$. For post-clustering inference, the clustering can be obtained from $\mathbf{X}^{(1)}$, $\mathcal{C}(\mathbf{X}^{(1)})$, and the contrast $\eta(C_k, C_{k'})$, with $C_k, C_{k'} \in \mathcal{C}(\mathbf{X}^{(1)})$ is considered as fixed for the test procedure on $\mathbf{X}^{(2)}$ since $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$ are independent.

Both data thinning and data fission can be used to address the global mean comparison. Under the null hypothesis, the p-value associated to a test statistic $\mathcal{T}$ is

$$p(\mathbf{x}) = \mathbb{P}_{\mathcal{H}_0}\left(\mathcal{T}\left(\eta(C_k(\mathbf{x}^{(1)}), C_{k'}(\mathbf{x}^{(1)}))^\top \mathbf{X}^{(2)}\right) \geq \mathcal{T}\left(\eta(C_k(\mathbf{x}^{(1)}), C_{k'}(\mathbf{x}^{(1)}))^\top \mathbf{x}^{(2)}\right)\right). \tag{5.8}$$

For the mean comparison (see Example 1), Wald's test depends on the distribution of $\mathbf{X}^{(2)}$ such that for a contrast vector η, $\eta^\top \mathbf{X}^{(2)} \sim \mathcal{N}_m((1-\varepsilon)\eta^\top\mu, (1-\varepsilon)\|\eta\|_2^2\,\Sigma)$. Then, under the null hypothesis, the distribution of the test statistic is known, and the p-value can be computed.

The sharing of information between $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$ is controlled explicitly by a parameter for both approaches (ε and τ for *data thinning* and *data fission* respectively). In practice, this parameter has to be calibrated. It seems natural to try to put enough information into $\mathbf{X}^{(1)}$ to obtain a good clustering partition. However, if the remaining part of information for $\mathbf{X}^{(2)}$ is insufficient, then the test keeps control of the type I error rate but could lose statistical power. On the contrary, if there is not enough information in $\mathbf{X}^{(1)}$, then the clustering obtained will not recover the true underlying partition. In the *data thinning* procedure, the parameter $\varepsilon \in (0, 1)$, corresponds to the proportion of information puts into $\mathbf{X}^{(1)}$ and will be numerically studied in Section 5.3.3. For the parameter of *data fission*, τ can take any positive value: the lower the value of τ, the more information is contained in $\mathbf{X}^{(1)}$. The interpretation of the part of information is not straightforward.

Note that, the *data thinning* procedure allows splitting the information into more than two data sets, while the data fission procedure does not allow it. Neufeld et al. (2024a) and Neufeld et al. (2023) use it for choosing the number of principal components in PCA or the number of clusters using the idea of cross-validation.

5.2.2 Conditional approaches

A second family of methods for post-clustering inference involves conditioning the p-value by the clustering select event. This approach is inspired by recent literature on post-selection inference (Fithian et al., 2014), which is introduced in Section 1.3.4. In post-clustering inference, the conditioning event $\{C_k(\mathbf{x}), C_{k'}(\mathbf{x}) \in \mathcal{C}(\mathbf{X})\}$ could fix the tested clusters. The naive p-value in Equation (5.5) is replaced by the following conditional p-value:

$$\begin{aligned}
&\mathbb{P}_{\mathcal{H}_0}\Big(\mathcal{T}(\eta(C_k(\mathbf{X}), C_{k'}(\mathbf{X})), \mathbf{X}) \geq \mathcal{T}(\eta(C_k(\mathbf{X}), C_{k'}(\mathbf{X})), \mathbf{x}) \;\Big|\; C_k(\mathbf{x}), C_{k'}(\mathbf{x}) \in \mathcal{C}(\mathbf{X})\Big) \\
&= \mathbb{P}_{\mathcal{H}_0}\Big(\mathcal{T}(\eta(C_k(\mathbf{x}), C_{k'}(\mathbf{x})), \mathbf{X}) \geq \mathcal{T}(\eta(C_k(\mathbf{x}), C_{k'}(\mathbf{x})), \mathbf{x}) \;\Big|\; C_k(\mathbf{x}), C_{k'}(\mathbf{x}) \in \mathcal{C}(\mathbf{X})\Big). \qquad (5.9)
\end{aligned}$$

The conditioning event in Equation (5.9) allows to consider the contrast vector as fixed, since it only depends on the clustering outcome. However, to compute the p-value, the conditional null distribution must be known. A common practice is to over-condition the p-value to obtain a tractable distribution.

From now on, we consider a Gaussian case with known spherical covariance matrix $\Sigma = \sigma^2 I_m$. We will discuss the relaxation of this assumption on the covariance matrix in Section 5.2.2.3.

5.2.2.1 p-value through over-conditioning

To obtain an analytically tractable p-value, Gao et al. (2024) over-condition the p-value (5.9) for the specific case of mean cluster comparison with the Wald's test statistic $\mathcal{T}(\eta, \mathbf{X}) := \left\|\eta^\top \mathbf{X}\right\|_2$ with η defined in Equation (5.3). In this section, we rewrite this work for a general contrast vector η only depending on Clusters C_k and $C_{k'}$. To understand the considered over-conditioning, let consider the decomposition of $\mathbf{X}$ into its projection on η and its orthogonal subspace:

$$\mathbf{X} = \pi_\eta^\perp \mathbf{X} + \pi_\eta \mathbf{X} = \pi_\eta^\perp \mathbf{X} + \left(\frac{\left\|\eta^\top \mathbf{X}\right\|_2}{\|\eta\|_2^2}\right) \eta \ \mathrm{dir}(\eta^\top \mathbf{X}), \qquad (5.10)$$

with $\pi_\eta^\perp = \left(I_n - \frac{\eta\eta^\top}{\|\eta\|_2^2}\right)$ and $\mathrm{dir}(\eta^\top \mathbf{X}) = \frac{\eta^\top \mathbf{X}}{\left\|\eta^\top \mathbf{X}\right\|_2}$ The proof of the decomposition is given in Equation (B.1) in Appendix B. With this decomposition, three parts of $\mathbf{X}$ are random: the

orthogonal projection of $\mathbf{X}$ on η, $\pi_\eta^\perp \mathbf{X}$, the test statistic $\left\|\eta^\top \mathbf{X}\right\|_2$ and the direction $\text{dir}(\eta^\top \mathbf{X})$. Then the p-value (5.9) is over-conditioned by fixing $\pi_\eta^\perp \mathbf{X}$ and $\text{dir}(\eta^\top \mathbf{X})$:

$$\begin{aligned} &p(\mathbf{x}; \{C_k(\mathbf{x}), C_{k'}(\mathbf{x})\}) \\ &\quad = \mathbb{P}_{\mathcal{H}_0}\Big(\left\|\eta(C_k(\mathbf{x}), C_{k'}(\mathbf{x}))^\top \mathbf{X}\right\|_2 \geq \left\|\eta(C_k(\mathbf{x}), C_{k'}(\mathbf{x}))^\top \mathbf{x}\right\|_2 \Big| C_k(\mathbf{x}), C_{k'}(\mathbf{x}) \in \mathcal{C}(\mathbf{X}), \blacksquare \Big) \end{aligned} \tag{5.11}$$

with $\blacksquare = \left\{\pi_\eta^\perp \mathbf{X} = \pi_\eta^\perp \mathbf{x}, \text{dir}(\eta^\top \mathbf{X}) = \text{dir}(\eta^\top \mathbf{x})\right\}$.

Theorem 1 of Gao et al. (2024), which is a specific case with $\eta_i = \frac{\mathbb{1}_{\{i \in C_k\}}}{|C_k|} - \frac{\mathbb{1}_{\{i \in C_{k'}\}}}{|C_{k'}|}$, can then be adapted for a general contrast η as a function of C_k and $C_{k'}$.

Theorem 1. *For any realization* $\mathbf{x}$ *of* $\mathbf{X}$*, where* $\mathbf{X} := (X_i)_{i \in [|n|]}$ *with* $X_i \sim \mathcal{N}_m(\mu_i, \sigma^2 I_m)$*, for any non overlapping groups of observations* $C_k, C_{k'} \in [|n|]$*, and for any contrast vector* $\eta := \eta(C_k, C_{k'}) \in \mathbb{R}^n$*,*

$$p(\mathbf{x}, \{C_k, C_{k'}\}) = 1 - \mathbb{F}\left(\left\|\eta^\top \mathbf{x}\right\|_2; \sigma \left\|\eta\right\|_2, S(\mathbf{x}, \{C_k; C_{k'}\})\right) \tag{5.12}$$

where $\mathbb{F}(.; c, S)$ *denotes the cumulative distribution function of a scaled Chi distribution with* m *degrees of freedom* $(c.\chi_m)$ *truncated to the set* S *and*

$$S(\mathbf{x}, \{C_k, C_{k'}\}) = \{\phi \geq 0 : C_k, C_{k'} \in \mathcal{C}(\tilde{\mathbf{x}}(\phi))\} \tag{5.13}$$

where

$$\tilde{\mathbf{x}}(\phi) = \pi_\eta^\perp \mathbf{x} + \frac{\phi}{\|\eta\|_2^2} \; \eta \; dir(\eta^\top \mathbf{x}) \tag{5.14}$$

is the perturbed data. Furthermore, if $\mathcal{H}_0$ *is true, then*

$$\mathbb{P}_{\mathcal{H}_0}\Big(p(\mathbf{X}; \{C_k, C_{k'}\}) \leq \alpha \;\Big|\; C_k, C_{k'} \in \mathcal{C}(\mathbf{X})\Big) = \alpha, \quad \text{for all } \alpha \in (0, 1). \tag{5.15}$$

That is, rejecting $\mathcal{H}_0 : \eta^\top \boldsymbol{\mu} = 0_m$ *whenever* $p(\mathbf{x}; \{\mathcal{C}_k, \mathcal{C}_{k'}\})$ *is below* α*, controls the selective type I error rate (Definition 1, Gao et al. (2024)) at level* α*.*

The proof of Theorem 1 is reported in Appendix B.1. Theorem 1 gives the explicit formula of the p-value which follows a truncated scaled Chi distribution. To compute the p-value, the set S defined in (5.13) must be explicit and depends on the perturbed data $\tilde{\mathbf{x}}(\phi)$. By fixing $\blacksquare$ in Equation (5.11), the decomposition of $\mathbf{X}$ in Equation (5.10) becomes $\mathbf{X} = \pi_\eta^\perp \mathbf{X} + \left(\frac{\|\eta^\top \mathbf{X}\|_2}{\|\eta\|_2^2}\right) \eta \, \text{dir}(\eta^\top \mathbf{x})$ where only the statistic $\left\|\eta^\top \mathbf{X}\right\|_2$ is random. The perturbed data can be rewritten as:

$$\begin{aligned} \tilde{\mathbf{x}}(\phi) &= \pi_\eta^\perp \mathbf{x} + \left(\frac{\phi}{\|\eta\|_2^2}\right) \eta \, \text{dir}(\eta^\top \mathbf{x}) \\ &= \mathbf{x} - \frac{\left\|\eta^\top \mathbf{x}\right\|_2}{\|\eta\|_2^2} \eta \, \text{dir}(\eta^\top \mathbf{x}) + \frac{\phi}{\|\eta\|_2^2} \eta \, \text{dir}(\eta^\top \mathbf{x}) \\ &= \mathbf{x} + \frac{\phi - \left\|\eta^\top \mathbf{x}\right\|_2}{\|\eta\|_2^2} \eta \, \text{dir}(\eta^\top \mathbf{x}). \end{aligned} \tag{5.16}$$

Equation (5.16) shows that the matrix $\mathbf{x}$ is perturbed allowing the two tested clusters to move further apart or closer together in the direction of the test statistic.

5.2.2.2 How to compute the p-value?

To compute the p-values, the set S must be explicitly specified. Gao et al. (2024) rely on the characteristics of Hierarchical Agglomerative Clustering (HAC) to specify the set S while Chen and Witten (2023) over-condition the p-value and use the characteristic of the K-means to explicit the p-value.

Explicit p-value for Hierarchical Clustering. Gao et al. (2024) address the problem of mean comparison between two clusters (see Example 1) under the assumption $\Sigma = \sigma^2 I_m$. For the HAC method with the squared Euclidean distance, a strong result in Lemma 1 of Gao et al. (2024) is that preserving Clusters C_k and $C_{k'}$ is equivalent to preserving the dendrogram of the HAC up to the $(n-K+1)$th step (where the dendrogram is cut to obtain K clusters). Lemma 1 in Gao et al. (2024) also shows that all the distances between the merged pair of clusters are preserved for the disturbed data $\tilde{\mathbf{x}}(\phi)$. Let D be the linkage measure used in the HAC method. Thus, with Lemma 1, the set S can be rewritten as the set of perturbations that leave the merging pairs of clusters in HAC unchanged. In other words, the distance between two merging pairs remains minimal among all pairs that can be merged by Theorem 2 in Gao et al. (2024):

$$S_{\text{HAC}} = \bigcap_{\{G,G'\}\in\mathcal{L}(\mathbf{x})} \left\{ \phi \geq 0 : D(G, G'; \tilde{\mathbf{x}}(\phi)) > \max_{l_{G,G'}(\mathbf{x}) \leq t \leq u_{G,G'}(\mathbf{x})} D\left(W_1^{(t)}(\mathbf{x}), W_2^{(t)}(\mathbf{x}); \mathbf{x}\right) \right\} \tag{5.17}$$

where $\mathcal{L}(\mathbf{x})$ is the set of all the pairs of clusters that do not merged in the dendrogram, $l_{G,G'}(\mathbf{x})$ (resp. $u_{G,G'}(\mathbf{x})$) is the first (resp. the last) step that Clusters G and G' could be merged, and $W_1^{(t)}(\mathbf{x})$ and $W_2^{(t)}(\mathbf{x})$ are the two merging groups at step t into the dendrogram. This means that S_{HAC} only contains the perturbations that have successfully maintained the dendrogram until there are only K clusters left. This set is the intersection of $\mathcal{O}(n^2)$ equations. Furthermore, Lance and Williams (1967) defined a relation that linearly allows decomposing the distance between two groups by the distances of the already merged groups. This decomposition only works for Average, Weighted, Ward, Centroid, and Median linkages. Therefore, for the squared Euclidean distance, the distance between two points can be expressed as a quadratic function in ϕ. Then, the set S_{HAC} is rewritten as a set of quadratic inequalities to be solved. This allows obtaining a computational cost of $\mathcal{O}(|\mathcal{M}(\mathbf{x})| + n^2 \log(n))$, where $\mathcal{M}(\mathbf{x})$ is the set of steps where inversion occurs in the dendrogram of $\mathbf{x}$. For single linkage, the decomposition by Lance and Williams (1967) does not allow for the same computational shortcuts, and Gao et al. (2024) proposes an explicit calculation of the set S_{HAC}.

Explicit p-value for K-means. Chen and Witten (2023) focus on the K-means algorithm and rely on the same statistical hypothesis as Gao et al. (2024), using the identical construction of the contrast vector as in Example 1. In order to obtain an explicit p-value, Chen and Witten (2023) condition the p-value by requiring that the clustering is entirely preserved across all steps of the K-means algorithm. Thus, the set S in Theorem 1 is reformulated as follows:

$$S_{\text{kmeans}} = \left\{ \phi \geq 0 : \bigcap_{t=0}^{T} \bigcap_{i=1}^{n} \left\{ C_i^{(t)}(\tilde{\mathbf{x}}(\phi)) = C_i^{(t)}(\mathbf{x}) \right\} \right\} \tag{5.18}$$

where T is the number of steps in the K-means algorithm and $C_i^{(t)}(\mathbf{x})$ is the cluster of sample i at step t. The conditions $C_i^{(t)}(\tilde{\mathbf{x}}(\phi)) = C_i^{(t)}(\mathbf{x})$ for all $t \in [|T|]$ and $i \in [|n|]$ can be rewritten as a quadratic equation in ϕ to solve, as the K-means algorithm uses the squared Euclidean

distance. Each equation depends on the clustering assignment of the previous step in the K-means algorithm. Then by fixing all the steps of the K-means algorithm, Chen and Witten (2023) provide an explicit computation of S_{kmeans}. Therefore, according to Proposition 5 in Chen and Witten (2023), the set S_{kmeans} can be computed in $\mathcal{O}\left(KT(n+p)+nKT\log(nKT)\right)$ operations.

Estimation of p-value via MC with Importance sampling. When the set $\mathcal{S}(\mathbf{x};\{C_k,C_{k'}\})$ is not tractable, the p-value $p(\mathbf{x};\{C_k,C_{k'}\})$ can be estimated by Monte Carlo sampling. Following Gao et al. (2024), the standard Monte Carlo can be used from Theorem 1:

$$p(\mathbf{x};\{C_k,C_{k'}\}) = \frac{\mathbb{E}\left[\mathbb{1}_{\{\phi\geq\left\|\eta^\top\mathbf{x}\right\|_2,C_k,C_{k'}\in\mathcal{C}(\tilde{\mathbf{x}}(\phi))\}}\right]}{\mathbb{E}\left[\mathbb{1}_{\{C_k,C_{k'}\in\mathcal{C}(\tilde{\mathbf{x}}(\phi))\}}\right]} \approx \frac{\sum_{q=1}^{Q}\mathbb{1}_{\{\phi_q\geq\left\|\eta^\top\mathbf{x}\right\|_2,C_k,C_{k'}\in\mathcal{C}(\tilde{\mathbf{x}}(\omega_q))\}}}{\sum_{q=1}^{Q}\mathbb{1}_{\{C_k,C_{k'}\in\mathcal{C}(\tilde{\mathbf{x}}(\omega_q))\}}} \tag{5.19}$$

for $\phi\sim(\sigma\left\|\eta\right\|_2)\cdot\chi_m$ and $\tilde{\mathbf{x}}(\phi)$ defined in Equation (5.16). The p-value can be estimated by naively sampling ϕ as $\phi_1,...,\phi_Q\overset{ind}{\sim}(\sigma\left\|\eta\right\|_2)\cdot\chi_m$. But in most cases, the quantity $\left\|\eta^\top\mathbf{x}\right\|_2$ lies in the tail of the distribution of ϕ. Thus, the Monte Carlo approximation of the p-value is poor for a finite values of Q. Monte Carlo with Importance Sampling (MC-Importance Sampling) approach can be used to solve this problem. Instead of sampling according to a distribution containing $\left\|\eta^\top\mathbf{x}\right\|_2$ in its tail, the idea is to sample around $\left\|\eta^\top\mathbf{x}\right\|_2$ and give more importance to the samples that are most likely to appear in the original distribution. Gao et al. (2024) propose to consider $\omega_1,...,\omega_Q\overset{ind}{\sim}\mathcal{N}(\left\|\eta^\top\mathbf{x}\right\|_2,\sigma^2\left\|\eta\right\|_2^2)$ and to give importance by scaling each sample ω_q by the factor $\pi_q=\frac{f_1(\omega_q)}{f_2(\omega_q)}$ where $f_1(\cdot)$ is the density function of $(\sigma^2\left\|\eta\right\|_2)\cdot\chi_m$ and $f_2(\cdot)$ is the density function of $\mathcal{N}(\left\|\eta^\top\mathbf{x}\right\|_2,\sigma^2\left\|\eta\right\|_2^2)$. With MC-Importance Sampling, the p-value can be approximated as:

$$p(\mathbf{x};\{C_k,C_{k'}\}) \approx \frac{\sum_{q=1}^{Q}\pi_q\mathbb{1}_{\{\omega_q\geq\left\|\eta^\top\mathbf{x}\right\|_2,C_k,C_{k'}\in\mathcal{C}(\tilde{\mathbf{x}}(\omega_q))\}}}{\sum_{q=1}^{Q}\pi_q\mathbb{1}_{\{C_k,C_{k'}\in\mathcal{C}(\tilde{\mathbf{x}}(\omega_q))\}}}. \tag{5.20}$$

This approximation corresponds to the proportion of perturbations that move the clusters away (i.e. $\phi\geq\left\|\eta^\top\mathbf{x}\right\|_2$) among those that preserve the clustering, adjusted by the π_q weights of the preferential sampling. This solution can be used for any clustering method (even for HAC and K-means algorithms). For example, the complete linkage in the HAC algorithm does not give explicit p-value and the estimation of the p-value can be performed by MC-Importance Sampling estimation in Equation (5.20), as proposed by Gao et al. (2024).

5.2.2.3 Beyond spherical covariance

Previously, we have assumed that the data follow a Gaussian distribution with a spherical covariance matrix. In practice, the assumption of independence between variables is rarely realistic. Both Gao et al. (2024) and Chen and Witten (2023) extend their results when Σ is a known general covariance matrix. They also address the case where an estimation of the unknown spherical variance is used.

Known general covariance matrix Σ. Gao et al. (2024) generalize their results for independent $X_i \sim \mathcal{N}_m(\mu_i, \Sigma)$ with $\mu_i \in \mathbb{R}^m$ and a known positive definite matrix Σ. They transform $\mathbf{X}$ into $\mathbf{X}\Sigma^{-\frac{1}{2}}$ where $X_i\Sigma^{-\frac{1}{2}} \sim \mathcal{N}(\mu_i\Sigma^{-\frac{1}{2}}, I_m)$ to come back to the spherical case. The p-value from Gao et al. (2024) is adapted as:

$$
\begin{aligned}
&p_\Sigma(\mathbf{x}; \{C_k, C_{k'}\}) \\
&= \mathbb{P}_{\mathcal{H}_0}\left(\left\|\eta(C_k(\mathbf{x}), C_{k'}(\mathbf{x}))^\top \mathbf{X}\Sigma^{-\frac{1}{2}}\right\|_2^2 \geq \left\|\eta(C_k(\mathbf{x}), C_{k'}(\mathbf{x}))^\top \mathbf{x}\Sigma^{-\frac{1}{2}}\right\|_2^2 \middle| \; C_k(\mathbf{x}), C_{k'}(\mathbf{x}) \in \mathcal{C}(\mathbf{X}), \blacksquare_\Sigma\right) \\
&= 1 - \mathbb{F}\left(\left\|\eta^2\mathbf{x}\Sigma^{-\frac{1}{2}}\right\|_2; \|\eta\|_2, S_\Sigma\left(\mathbf{x}; \{C_k, C_{k'}\}\right)\right), \qquad (5.21) \\
&\text{where } \blacksquare_\Sigma = \left\{\pi_\eta^\perp \mathbf{X} = \pi_\eta^\perp \mathbf{x}, \operatorname{dir}\left(\eta^\top \mathbf{X}\Sigma^{-1/2}\right) = \operatorname{dir}\left(\eta^\top \mathbf{x}\Sigma^{-1/2}\right)\right\},
\end{aligned}
$$

and $S_\Sigma(\mathbf{x}; \{C_k, C_{k'}\}) = \left\{\phi \geq 0 : C_k, C_{k'} \in \mathcal{C}\left(\pi_\eta^\perp \mathbf{x} + \phi\left(\frac{\eta}{\|\eta\|_2^2}\right)\operatorname{dir}\left(\eta^\top \mathbf{x}\Sigma^{-\frac{1}{2}}\right)\Sigma^{\frac{1}{2}}\right)\right\}$. The same transformation and result exist for the p-value developed by Chen and Witten (2023).

González-Delgado et al. (2023) propose a procedure which generalizes the test to any type of covariance between samples and variables as the matrix normal model (Horn and Johnson, 2012), $\mathbf{X} \sim \mathcal{MN}_{n\times m}(\boldsymbol{\mu}, U, \Sigma)$ where $\boldsymbol{\mu}$ is a matrix $n \times m$ of means, $U \in \mathbb{R}^{n\times n}$ is the matrix of covariance between samples, Σ is the matrix of covariance between variables. Then they adapt the p-value of Gao et al. (2024) in the case of mean difference comparison by changing the test statistic using a specific norm $\|w\|_V = \sqrt{w^\top V_{C_k,C_{k'}}^{-1} w}$, $w \in \mathbb{R}^m$ where $V_{C_k,C_{k'}} \in \mathbb{R}^{m\times m}$ contains the information about the scale matrices U and Σ, such that $V_{C_k,C_{k'}} = \mathbf{D}_{C_k,C_{k'}}(U_{C_k,C_{k'}} \otimes \Sigma)\mathbf{D}_{C_k,C_{k'}}$, with $\otimes$ denotes the Kronecker product between matrices. The matrix $U_{C_k,C_{k'}}$ is the principal sub-matrix of U which contains the rows and columns in $C_k \cup C_{k'}$ and

$$
\mathbf{D}_{C_k,C_{k'}} = \left(\underbrace{\frac{1}{|C_k|}I_m, ..., \frac{1}{|C_k|}I_m}_{|C_k| \text{ times}}, \; \underbrace{-\frac{1}{|C_{k'}|}I_m, ..., -\frac{1}{|C_{k'}|}I_m}_{|C_{k'}| \text{ times}}\right).
$$

Then the p-value becomes

$$
\begin{aligned}
p_{V_{C_k,C_{k'}}}(\mathbf{x}; \{C_k, C_{k'}\}) &= \mathbb{P}_{\mathcal{H}_0}\left(\left\|\eta^\top \mathbf{X}\right\|_V \geq \left\|\eta^\top \mathbf{x}\right\|_V \middle| C_k, C_{k'} \in \mathcal{C}(\mathbf{X}), \blacksquare_V\right) \\
&= 1 - \mathbb{F}'\left(\left\|\eta^\top \mathbf{x}\right\|_V; S_V(\mathbf{x}; \{C_k, C_{k'}\})\right), \qquad (5.22) \\
\text{where } \blacksquare_V &= \left\{\pi_\eta^\perp \mathbf{X} = \pi_\eta^\perp \mathbf{x}, \operatorname{dir}_V\left(\eta^\top \mathbf{X}\right) = \operatorname{dir}_V\left(\eta^\top \mathbf{x}\right)\right\},
\end{aligned}
$$

with $\operatorname{dir}_V(w) = \frac{w}{\|w\|_V}$ and $\mathbb{F}'(t; S_V)$ is the cumulative function distribution of a χ_m distribution truncated to the set S, and

$$
S_V(\mathbf{x}; \{C_k, C_{k'}\}) = \left\{\phi \geq 0 : C_k, C_{k'} \in \mathcal{C}\left(\pi_\eta^\perp \mathbf{x} + \phi\left(\frac{\eta}{\|\eta\|_2^2}\right)\operatorname{dir}_V\left(\eta^\top \mathbf{x}\Sigma^{-\frac{1}{2}}\right)^\top \Sigma^{\frac{1}{2}}\right)\right\}.
$$

Gao et al. (2024) is a sub-case of this test when $U = I_n$ and $\Sigma = \sigma^2 I_m$.

Impact of variance estimation. Gao et al. (2024) and Chen and Witten (2023) have also extended their results to the case of an unknown spherical covariance matrix $\Sigma = \sigma^2 I_m$. Theorem 4 in Gao et al. (2024) specifies that if a consistent over-estimator of the variance σ^2 is used (i.e., $\lim_{n\to\infty} \mathbb{P}_{\mathcal{H}_0} = \left(\hat{\sigma}(\mathbf{X}^{(n)}) \geq \sigma \mid C_k^{(n)}, C_{k'}^{(n)} \in \mathcal{C}(\mathbf{X}^{(n)})\right) = 1$, where "$(n)$" specifies the

dependence in the sample size), then the test asymptotically controls the type I error rate. They do not recommend a specific estimator of the variance but in practice use the global variance estimation

$$\hat{\sigma}_{\text{all}}(\mathbf{x}) = \sqrt{\frac{1}{m(n-1)} \sum_{i=1}^{n} \sum_{j=1}^{m} (x_{ij} - \bar{\mathbf{x}}_j)^2} \tag{5.23}$$

where $\bar{\mathbf{x}}_j$ is the empirical mean of the j-th variable of $\mathbf{x}$. If data contains no signal (no true clusters), then the variance should be well estimated. In other cases (presence of true clusters), the variance should be overestimated, and then Theorem 4 stands for this estimator. Chen and Witten (2023) introduce an extension for the K-means algorithm with unknown σ^2. They propose an estimator of the variance based on the median:

$$\hat{\sigma}_{\text{MED}}(x) = \left\{ \underset{i\in[|n|],j\in[|m|]}{\text{median}} \left(x_{ij} - \underset{i\in[|n|]}{\text{median}}(x_{ij}) \right)^2 / M_{\chi^2_1} \right\}^{1/2} \tag{5.24}$$

where $M_{\chi^2_1}$ is the median of the χ^2_1 distribution. In their Theorem 3.1, González-Delgado et al. (2023) generalize also their result to the case when one of the two matrices U, Σ is estimated (the other is still assumed to be known). In our case, we consider the matrix U as known and the estimation of Σ is given by

$$\hat{\Sigma}(\mathbf{X}) = \frac{1}{n-1} \left(\mathbf{X} - \bar{\mathbf{X}} \right)^\top U^{-1} \left(\mathbf{X} - \bar{\mathbf{X}} \right) \tag{5.25}$$

where $\bar{\mathbf{X}} = \mathbf{1}_n \otimes \frac{1}{n} \sum_{i=1}^{n} X_i$. Plugging an over-estimator of the covariance matrix is not enough to control the type I error rate: extra assumptions on the values of $\boldsymbol{\mu}$ and U must be met. Assumption 3.1 in González-Delgado et al. (2023) states that $\boldsymbol{\mu}$ is composed of exactly K different values. If U is non-diagonal, Assumption 3.2 requires that the proportion of pairs of observations for a given pair of means approach the product of individual proportion when both observations are far away. Assumption 3.3 imposes some properties on the matrix U, which, for example, the diagonal, compound symmetry and the auto-regressive covariance matrix structures satisfy. González-Delgado et al. (2023) have shown that these three assumptions are sufficient for the estimator (5.25) to over-estimate Σ asymptotically.

Accounting for unknown variance in the spherical case. One of the limits of the previous methods is that the variance must be known or estimated on the same data. Yun and Foygel Barber (2023) propose a method that avoids estimating the covariance matrix, which is assumed to be spherical ($\Sigma = \sigma^2 I_m$). In this work, instead of comparing means of clusters, they test a stronger hypothesis where each sample from the two compared clusters has the same mean:

$$\mathcal{H}_0(C_k(\mathbf{X}), C_{k'}(\mathbf{X})) : \mu_i = \mu_{i'}, \text{ for all } i, i' \in C_k(\mathbf{X}) \cup C_{k'}(\mathbf{X}). \tag{5.26}$$

This hypothesis can be rewritten as $\mathcal{H}_0(C_k(\mathbf{X}), C_{k'}(\mathbf{X})) : \boldsymbol{\zeta}(C_k(\mathbf{X}), C_{k'}(\mathbf{X}))^\top \boldsymbol{\mu} = 0$ where $\boldsymbol{\zeta}$ is a contrast matrix of size $n \times n$ such that

$$\boldsymbol{\zeta} = \left(I_n - \frac{w w^\top}{\|w\|^2} - \sum_{i \in [|n|] \setminus C_k \cup C_{k'}} e_i e_i^\top \right)$$

with $w_i := \frac{\mathbb{1}_{\{i \in C_k \cup C_{k'}\}}}{|C_k \cup C_{k'}|}$ and e_i is the ith canonical vector of $\mathbb{R}^n$. This contrast only depends on C_k and $C_{k'}$ but it cannot be used directly in Theorem 1 without adapting it to matrix contrast. But Theorem 1 needs to know the variance, which is not known in this model.

To generalize to an unknown variance, Yun and Foygel Barber (2023) use the same strategy to decompose $\mathbf{X}$ as projections of $\mathbf{X}$ onto orthogonal subspaces

$$\mathbf{X} = \mathcal{P}_0\mathbf{X} + \mathcal{P}_1\mathbf{X} + \mathcal{P}_2\mathbf{X} \tag{5.27}$$

where $\mathcal{P}_0 = \frac{\eta\eta^\top}{\|\eta\|_2^2}$ is the rank-one projection matrix that captures the difference in mean between C_k and $C_{k'}$, and with η defined in Equation (5.3). The matrix

$$\mathcal{P}_1 = \left(I_{C_k} - \frac{\mathbb{1}_{\{i\in C_k\}}\mathbb{1}_{\{i\in C_k\}}^\top}{|C_k|}\right) + \left(I_{C_{k'}} - \frac{\mathbb{1}_{\{i\in C_{k'}\}}\mathbb{1}_{\{i\in C_{k'}\}}^\top}{|C_{k'}|}\right)$$

captures differences among points within C_k and among points within $C_{k'}$, where I_{C_k} is the diagonal matrix of size $n \times n$ where the diagonal vector corresponds to a vector composed by $\mathbb{1}_{\{i\in C_k\}}$. The matrix $\mathcal{P}_2$ is the projection matrix onto the orthogonal space of $\mathcal{P}_0$ and $\mathcal{P}_1$ such that $\mathcal{P}_2 = I_n - \mathcal{P}_0 - \mathcal{P}_1$. For this procedure, they use the test statistic

$$\mathcal{T}(\mathcal{C}(\mathbf{X}), \mathbf{X}) = (|C_k(\mathbf{X})| + |C_{k'}(\mathbf{X})| - 2)\,\frac{\|\mathcal{P}_0\mathbf{X}\|_F^2}{\|\mathcal{P}_1\mathbf{X}\|_F^2} \tag{5.28}$$

where $\|\cdot\|_F$ is the Frobenius norm. Then, the p-value is

$$\begin{aligned} &p(\mathbf{x};\{C_k, C_{k'}\}) \\ &\quad = \mathbb{P}_{\mathcal{H}_0}\left((|C_k| + |C_{k'}| - 2)\frac{\|\mathcal{P}_0\mathbf{X}\|_F^2}{\|\mathcal{P}_1\mathbf{X}\|_F^2} \geq (|C_k| + |C_{k'}| - 2)\frac{\|\mathcal{P}_0\mathbf{x}\|_F^2}{\|\mathcal{P}_1\mathbf{x}\|_F^2} \mid \mathcal{C}(\mathbf{x}) = \mathcal{C}(\mathbf{X}), \blacksquare_{\text{YB}}\right) \end{aligned} \tag{5.29}$$

$$\begin{aligned} \text{with } \blacksquare_{\text{YB}} = \Big\{ &\|\mathcal{P}_0\mathbf{X}\|_F^2 + \|\mathcal{P}_1\mathbf{X}\|_F^2 = \|\mathcal{P}_0\mathbf{x}\|_F^2 + \|\mathcal{P}_1\mathbf{x}\|_F^2\,, \\ &\frac{\mathcal{P}_0\mathbf{X}}{\|\mathcal{P}_0\mathbf{X}\|_F} = \frac{\mathcal{P}_0\mathbf{x}}{\|\mathcal{P}_0\mathbf{x}\|_F},\ \frac{\mathcal{P}_1\mathbf{X}}{\|\mathcal{P}_1\mathbf{X}\|_F} = \frac{\mathcal{P}_1\mathbf{x}}{\|\mathcal{P}_1\mathbf{x}\|_F},\ \mathcal{P}_2\mathbf{X} = \mathcal{P}_2\mathbf{x}\Big\}. \end{aligned} \tag{5.30}$$

Then Theorem 2 from Yun and Foygel Barber (2023) gives that the p-value in Equation (5.29) is equal to

$$p(\mathbf{x}; \{C_k, C_{k'}\}) = 1 - \mathbb{F}_{F_{m,(|C_k|+|C_{k'}|-2)m}}\left((|C_k| + |C_{k'}| - 2)\frac{\|\mathcal{P}_0\mathbf{x}\|_F^2}{\|\mathcal{P}_1\mathbf{x}\|_F^2}, S_{\text{YB}}\right) \tag{5.31}$$

based on the cumulative distribution function of a Fisher distribution $F_{m,(|C_k|+|C_{k'}|-2)m}$ truncated on $S_{\text{YB}} = \{\phi > 0 : \mathcal{C}(\mathbf{x}) = \mathcal{C}(\tilde{\mathbf{x}}_{\text{YB}}(\phi))\}$ and the perturbed data are

$$\tilde{\mathbf{x}}_{\text{YB}}(\phi) = \left(\sqrt{\frac{\phi}{\blacktriangle_{k,k'} + \phi}}\frac{\mathcal{P}_0\mathbf{x}}{\|\mathcal{P}_0\mathbf{x}\|_F} + \sqrt{\frac{\blacktriangle_{k,k'}}{\blacktriangle_{k,k'} + \phi}}\frac{\mathcal{P}_1\mathbf{x}}{\|\mathcal{P}_1\mathbf{x}\|_F}\right) \cdot \sqrt{\|\mathcal{P}_0\mathbf{x}\|_F^2 + \|\mathcal{P}_1\mathbf{x}\|_F^2} + \mathcal{P}_2\mathbf{x}. \tag{5.32}$$

They distinguish two cases to compute the p-value. If the clustering is only composed of $K = 2$ clusters, the set S_{YB} can be rewritten as a subset of S_{HAC} (resp. $S_{K\text{means}}$), and they use the explicit characterization of S proposed by Gao et al. (2024) for HAC (resp. Chen and Witten (2023) for K-means) clustering method. For $K \geq 3$, they use an estimation of the p-value using the MC-Importance Sampling approach. A major limitation of this method is to be constrained to a spherical covariance matrix. Indeed, Cochran's theorem used to compute the p-value cannot be used for general covariance matrix and the transformation used by Gao et al. (2024) and Chen and Witten (2023) cannot be exploited as the covariance matrix is unknown.

5.3 Numerical comparisons with known spherical covariance

The methods for post-clustering inference presented in Section 5.2 have been developed very recently and a comprehensive comparison between these methods is still lacking. We propose to address this gap through numerical simulations. We consider Example 1 to compare multivariate means of clusters. Initially, the study aims to verify whether these methods control the type I error rate. Once the faulty methods are identified, the study proceeds with an analysis of the statistical power of the methods still in contention. In particular, the study aims to measure the impact on statistical power and computation time of an inference method conditioned with an explicit expression of the p-value compared to an estimation via MC-Importance Sampling.

5.3.1 Settings

5.3.1.1 Simulation settings

Let $\mathbf{X}$ be a $n \times m$ matrix corresponding to n independent observations of m features. We assume that for $i = 1, ..., n$, $X_i \sim \sum_{k=1}^{K} \frac{1}{K} \mathcal{N}_m(\nu_k, \Sigma_k)$. We assume that there exists an unknown partition of the n observations into K clusters, $\mathcal{C}^* = \{C_1^*, ..., C_K^*\}$. Let ν_k be the theoretical means of Cluster C_k^*. We consider two different settings with $K = 2$ and $K = 3$, respectively. These settings are illustrated in Figure 5.1.

Setting 1 In the first setting, $\nu_1 = (0_{m/2}, 0_{m/2})^\top$ and $\nu_2 = (a\mathbf{1}_{m/2}, 0_{m/2})$ where 0_s is a s-size vector of 0 and $\mathbf{1}_s$ a s-size vector of 1. Let Σ be the common $m \times m$ covariance matrix to all clusters ($\Sigma_1 = \Sigma_2 = \Sigma$) which is an auto-regressive matrix defined by $\Sigma_{ij} = \sigma^2 \rho^{|i-j|}$. The covariance matrix is spherical ($\Sigma = \sigma^2 I_m$) when $\rho = 0$. Then, for this setting, the difference between the mean of the true clusters is $\|\nu_1 - \nu_2\|_2 = \sqrt{\frac{m}{2}} \cdot a$ with η defined in Equation (5.3). Figure 5.1-A shows an example of a data set generated from Setting 1 with $n = 500$, $m = 2$ and $a = 5$.

Setting 2 The second setting allows us to challenge methods to compare two clusters in the presence of additional clusters. For this setting, we fix the number of variables to $m = 3$. The means of the clusters are $\nu_1 = (\frac{-a}{2}, 0, 0)^\top$, $\nu_2 = (\frac{a}{2}, 0, 0)^\top$ and $\nu_3 = (0, \frac{\sqrt{3}a}{2}, 0)^\top$. We define $\Sigma_1 = \Sigma_2 = \Sigma_3 = \sigma^2 I_3$. Then, the true difference in means between any two clusters is a. This construction allows to circumvent the problem of switching cluster labels inherent in clustering methods. Figure 5.1-B shows the first two dimensions of a data set generated from Setting 2 with $n = 500$ and $a = 5$. Note that Setting 1 is equivalent for $a = 0$, $\rho = 0$ and $m = 3$ to Setting 2.

5.3.1.2 Compared procedures

The post-clustering test procedures which are compared are summarized in Table 5.1. The naive method (presented in Section 5.1) and the data thinning procedure (see Section 5.2.1) are used with Wald's test as $\mathcal{T}(\eta, \mathbf{X}) = \left\|\eta^\top \mathbf{X}\right\|_2$ with η defined in Equation (5.3) and Σ known, to be comparable to conditional approaches. Data thinning is applied with $\varepsilon = 0.7$ (the justification for this parameter choice is provided in Section 5.3.3). Conditional approaches discussed in Section 5.2.2 depend on the clustering methods. For this study, the clustering methods used are Hierarchical Agglomerative Clustering (HAC) with Ward's linkage, the K-means algorithm (KM), and the clustering based on Gaussian Mixture Model (GMM) with the same spherical covariance matrix for each component.

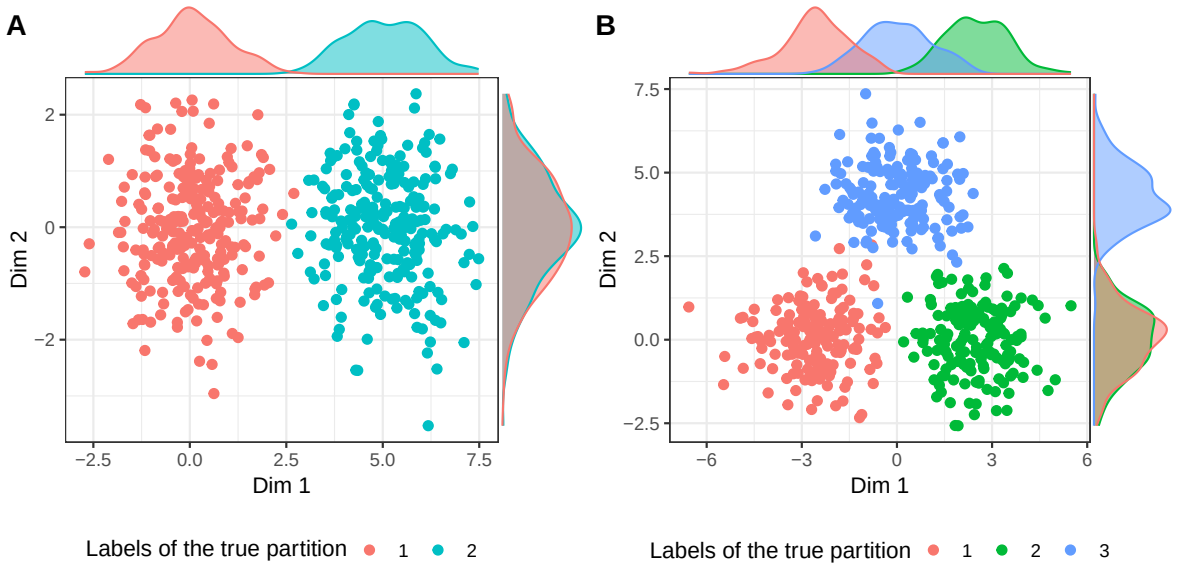

Figure 5.1 – **A** (left): Example of data generated from Setting 1 with $n = 500$, $m = 2$ and $a = 5$. **B** (right): The first two dimensions of data generated from Setting 2 with $n = 500$ and $a = 5$.

Name	Method	Reference	Clustering	Variance	MC-IS	Implementation
Naive-HAC Naive-KM Naive-GMM	Naive	-	HAC K-means GMM	Σ known	-	-
DT-HAC DT-KM DT-GMM	Data thinning	Neufeld et al. (2024a)	HAC K-means GMM	Σ known	- - -	Based on `datathin`

Cond-HAC
Cond-HAC-IS
Cond-KM
 	
Cond-KM-IS
Cond-GMM-IS
	Conditional Approaches with known variance	Gao et al. (2024)
		Gao et al. (2024)
		Chen and Witten (2023)
		
		Gao et al. (2024)
		Gao et al. (2024)
			HAC
			HAC
			K-means
			
			K-means
			GMM
				$\sigma^2 I_m$ and Σ known	-
					✓
					-
					
					✓
					✓
						`clusterpval`
						`clusterpval`
						`KmeansInference`
						
						`clusterpval`
						`clusterpval`

Table 5.1 – Summary of the methods which are compared in a multivariate context. The covariance matrix is theoretically defined as general (Σ) or spherical ($\sigma^2 I_m$). Methods can be explicit ('-' in column 'MC-IS') or estimated by Monte Carlo Importance Sampling (MC-IS) with $Q = 1000$ draws ('✓' in column 'MC-IS').

5.3.2 Evaluation of type I error rate

We start by assessing the methods' proper type I error control. For this, we use Setting 1 with $a = 0$ so that the data form a single cluster. We set $n \in \{10, 20, 50, 100, 200, 500\}$ and $m \in \{2, 10, 50, 100\}$, with $\Sigma = I_m$. Figure 5.2 represents the empirical cumulative distribution function (ecdf) of the p-values obtained under the null hypothesis for the HAC method. For conciseness, only the values of $n \in \{10, 100, 500\}$ and $m \in \{2, 100\}$ are reported but similar results are obtained. As expected and already reported e.g. by Gao et al. (2024), the naive method does not control the type I error rate. Even for small values of n, the naive method detects a difference between empirical cluster means as signal. Other methods control the type I error rate since their ecdf are very close to the identity function, which corresponds to the uniform distribution. The same conclusion can be made for all methods with K-means and GMM clusterings (see Figure B.1).

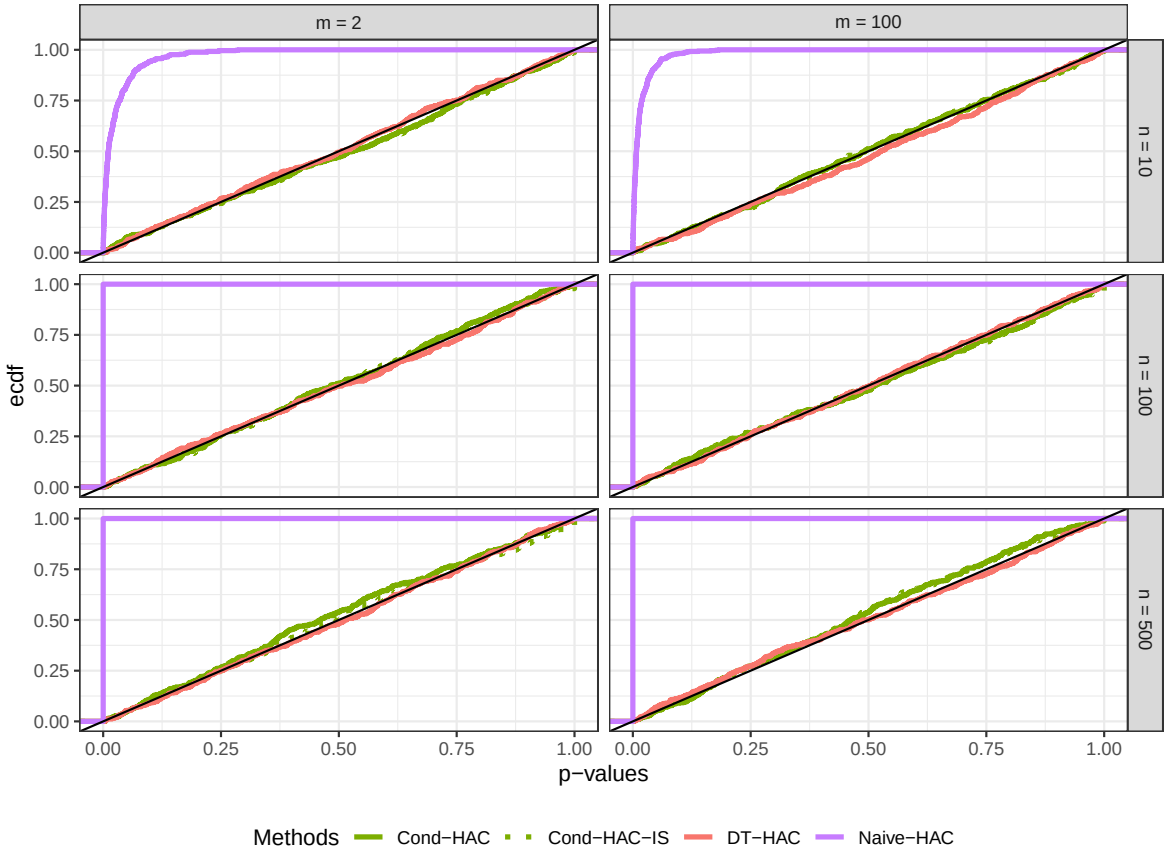

Figure 5.2 – Empirical cumulative distribution function (ecdf) of the p-values through 500 experiments under the null hypothesis. Each panel corresponds to a combination of a value of n and m. Methods are computed using hierarchical clustering (HAC) with Ward's linkage.

Exact conditional methods give uniform p-values, while the estimation by MC-Importance Sampling can be more conservative (see the clustering obtained from GMM in Figure B.1), often yielding a large proportion of p-values equal to 1, indicating that only perturbations that move clusters far apart preserve them (see Equation (5.20)). This suggests that clusters likely come from the same distribution.

We have checked that all these results are still valid in the presence of a third (untested) cluster. This point is illustrated in Figure B.2, which is based on experiments in Setting 2 with $K = 3$ and $a = 0$.

5.3.3 Evaluation of statistical power

To evaluate the statistical power of the methods, Setting 1 is parameterized with a variation of the signal $a \in \{0, 1, ..., 10\}$ while keeping the variations of the number of individuals n and variables m, with $\Sigma = \sigma^2 I_m$ ($\rho = 0$). For each experiment, means of clusters C_1 and C_2 are compared. The statistical power at level 0.05 is estimated as the proportion of experiments whose p-value is below 0.05. We can notice that Gao et al. (2024) and Chen and Witten (2023) examine conditional statistical power, i.e., only considering cases where both tested clusters are correctly identified. In contrast, our framework considers both clustering and statistical testing to compare all methods. For weak signal ($a \in (1, 4)$), the proportion of simulations recovering the true partition is low in practice (see Figures B.14 and B.15). Thus,

to refine the analysis, we computed the Adjusted Rand Index (ARI)(Hubert and Arabie, 1985) between the estimated clustering and the true partition of observations. This metric compares two partitions and ranges from -1 to 1. An ARI close to 1 indicates strong agreement between the two partitions. An ARI of 0 suggests that the partitions could be obtained by chance. Negative values indicate that the two scores systematically disagree.

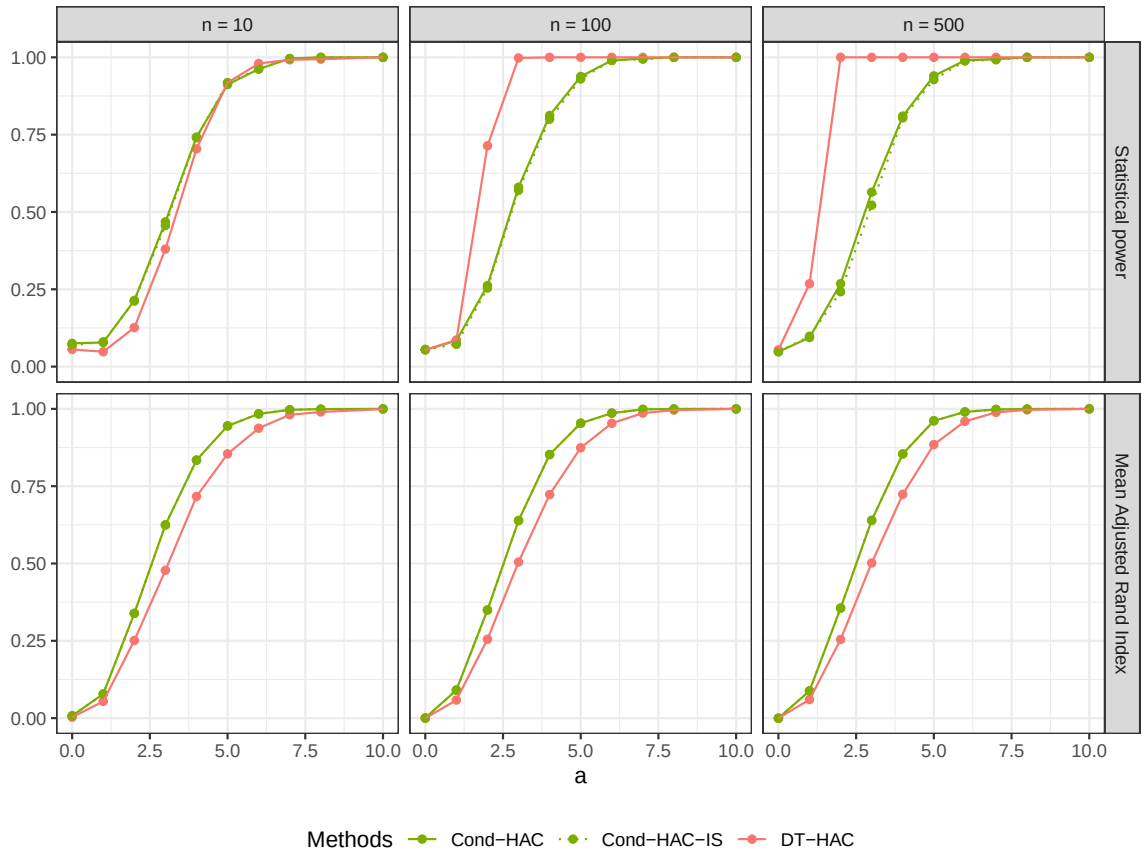

Figure 5.3 – Statistical power for the multivariate test and Adjusted Rand Index using HAC clustering method for Setting 1 with $m = 2$.

Figure 5.3 presents the statistical power of the test and the ARI of the clustering depending on the signal for Setting 1. Only HAC with $m = 2$ are reported because the order of the curves is similar for other clustering methods and values of m. The naive method is not included since it is not valid, as shown in the previous section.

Recall that while data thinning splits information between clustering and testing, conditional approaches use the entire information to estimate the clustering. However the latter need to introduce an extra conditioning at the testing step. Our goal here is to determine which strategy achieves the highest statistical power. Our results indicate that data thinning is more powerful than conditional methods. This was not obvious *a priori*, since the test could loose power due to a clustering less accurate than if it had been obtained from all the data. We also note that the statistical power of the conditional approaches appears to vary only slightly with increasing sample size while data thinning improves its statistical power. These results correspond to $\varepsilon = 0.7$, which in our experience leads to the best the trade-off between information used in clustering and testing (as discussed below in a dedicated paragraph).

With $Q = 1000$ draws, the conditional approach of Gao et al. (2024) estimated by MC-Importance Sampling is as powerful as the explicit version of the method, using the HAC

clustering. All these remarks also apply to the other clustering methods (see Figure B.3), where data thinning remains the most powerful method.

As above, we have checked that these results are still valid in the presence of a third (untested) cluster. This point is illustrated in Figure B.4, which is based on experiments in Setting 2 with $K = 3$. The observed lack of power for the K-means algorithm for a strong signal is a consequence of the clustering partition, which is not perfectly recovered in this setting.

Importance sampling can be more efficient than closed-form p-values for conditional tests. The specific case of K-means clustering is further studied in Figure 5.4. Perhaps surprisingly, the conditional approach with p-values estimated by MC-Importance Sampling turns out to be more powerful than the exact test proposed by Chen and Witten (2023) (see also Figure B.3). This can be explained as follows. To establish an explicit formulation for the conditional test applied to the K-means method, Chen and Witten (2023) had to condition on all successive steps of the K-means algorithm, as explained in Section 5.2.2.2. This results in a conditioning event that is much smaller than the one involved in the MC-Importance sampling test (compare Equations (5.18) and (5.13)).

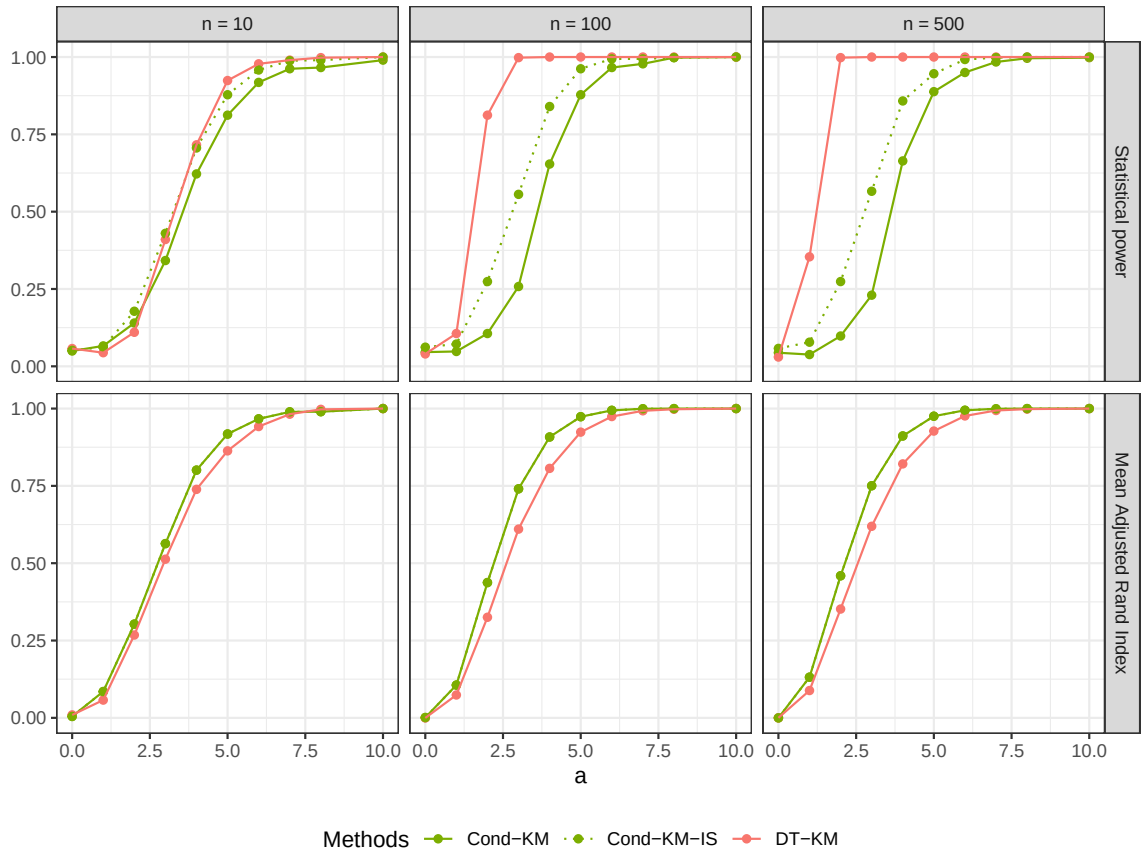

Figure 5.4 – Statistical power and ARI for the multivariate test using K-means clustering method on Setting 1 with $m = 2$.

In theory, one possible advantage of exact p-value is to be faster to compute than a Monte Carlo estimation. We have compared the computation time for explicit p-value and the MC-Importance Sampling estimation with $Q \in \{100, 200, 500, 1000\}$. The simulations were conducted on a computer equipped with an Intel Core i7 processor running at 2.80 GHz, 8

cores, 16 GB of RAM. None of the two method is always faster than the other one, and the

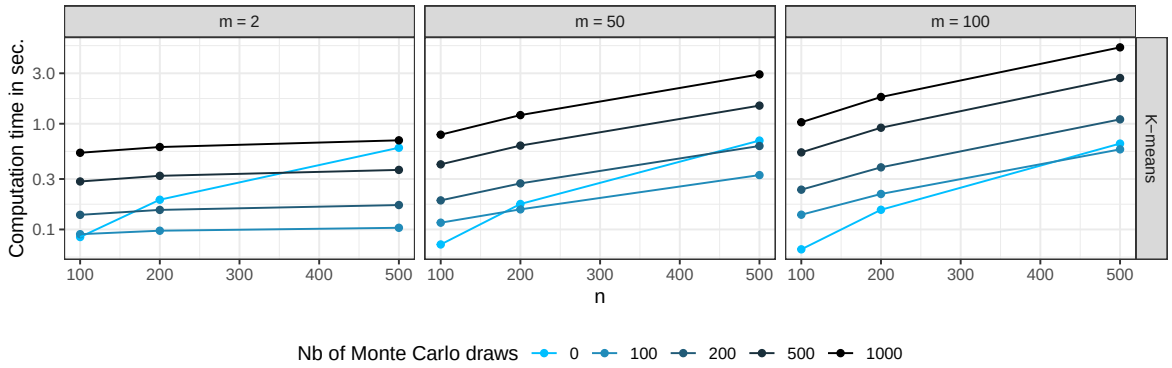

Figure 5.5 – Analysis of computation times for the conditional approaches using K-means algorithm for Setting 1. "0 draws" corresponds to the exact version of the test developed by Chen and Witten (2023). Other values of draws correspond to the estimated p-value by MC-Importance Sampling proposed by Gao et al. (2024) (see Equation (5.20))

MC-Importance Sampling method is even faster than the exact one for small m and large n. The same observations hold for Setting 2 (see Figure B.5). Overall, in the particular case of K-means, the MC-Importance Sampling conditional approach seems to be preferable to the exact conditional approach.

What value ε should be used for the thinning parameter? The data thinning method developed by Neufeld et al. (2024a) is parameterized by the thinning parameter $\varepsilon \in (0, 1)$. It can be interpreted as the proportion of information contained in $\mathbf{X}$ used for clustering, while $1 - \varepsilon$ corresponds to the proportion of information used for the test. The choice of ε is therefore important when using this method. This section illustrates the impact of this parameter on the statistical results of controlling the type I error rate and statistical power. We have comparing the output of the data thinning method with $\varepsilon \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$. The following results are presented using HAC, but similar trends are observed for K-means and GMM. First, we have checked that for all values of ε, the method controls the type I error rate (see Figure B.6).

Statistical power and ARI are presented in Figures 5.6 for Setting 1 and $n = 100$ (for other values of n and Setting 2, interpretations are similar). Smaller values of ε make it difficult to identify the correct clusters. Despite having sufficient information for the test, the individuals are too mixed to obtain accurate test results. Conversely, the clustering is accurate for $\varepsilon = 0.9$, but there is not enough information for the test to be powerful. The power is maximal for intermediate values of ε for all signal values. The precise value of $\varepsilon \in [0.5, 0.7]$ does not affect the power of the test, but impacts the accuracy of clustering. Based on these results, we chose the value $\varepsilon = 0.7$ because this value maximizes the statistical power and clustering recovery.

5.4 Numerical comparisons with unknown spherical covariance or auto-regressive covariance

In the previous section, the methods have been tested with known spherical covariance matrix $\Sigma = \sigma^2 I_m$. This section aims to assess how the methods behave, considering some correlations between variables or an unknown covariance matrix. On this latter point, Neufeld et al. (2024a) and Hivert et al. (2024b) have already shown that if the variance is poorly

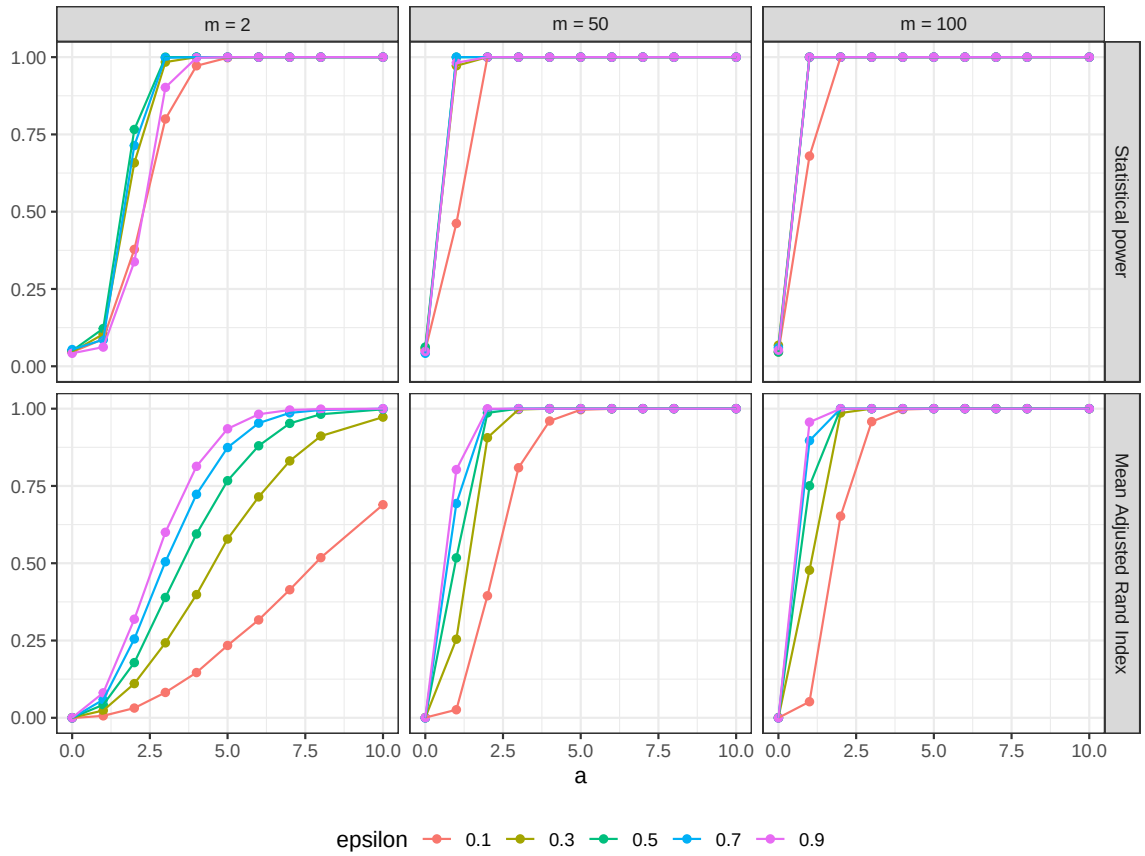

Figure 5.6 – Statistical power and ARI for several values of ε in the data thinning method for Setting 1 with $n = 100$.

estimated, then $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$ obtained by data thinning procedure on $\mathbf{X}$ are not independent, leaeding to an invalid test. Thus, we will focus on conditional approaches.

5.4.1 Impact of the estimation of σ^2 in the spherical case

5.4.1.1 Setting and methods

This analysis is made by simulating datasets from Setting 1 with $n = 500$ and $m \in \{2, 10\}$. The covariance matrix is spherical ($\Sigma = \sigma^2 I_m$) with $\sigma^2 \in \{0.5, 1, 3\}$. Table 5.2 summarizes methods used for this comparison. The YFB method is not available in a package. We have therefore used the functions, only implemented for HAC, proposed in the simulations provided with (Yun and Foygel Barber, 2023). In this section, only HAC clustering is used for simulations. The conditional approach with known variance is used as an oracle by plugging the true value of Σ. Since González-Delgado et al. (2023) is a generalization of Gao et al. (2024), both methods using the oracle yield the same result. In Section 5.3, the results for the conditional approach of Gao et al. (2024) shows that the estimation with Important Sampling gives the same results that the exact p-value. Thus, for this analysis, only the exact p-value is reported for this method. For the conditional approach with unknown variance (Yun and Foygel Barber, 2023), both exact and estimated p-values are studied. The estimators of the

global variance $\hat{\sigma}^2_{all}$ (all) (see Equation (5.23)) and the intra-cluster variance (intra)

$$\hat{\sigma}^2_{\text{intra}} = \sum_{k=1}^{K} \frac{|C_k|}{nm} \sum_{i \in C_k} \|x_i - \bar{\mathbf{x}}_k\|_2^2 \,, \tag{5.33}$$

with $\bar{\mathbf{x}}_k = \frac{1}{|C_k|} \sum_{i \in C_k} x_i$, are plugged in spherical assumption, as well as the estimator $\hat{\Sigma}$ of a general dependence (GD) defined in Equation (5.25), with $U = I_n$ known.

Name	**Method**	**Reference**	**Variance** $\Sigma = \sigma^2 I_m$	**Implementation**
Cond-oracle	Conditional Approaches with **known** variance	Gao et al. (2024)	σ^2 known	`clusterpval`
Cond-all		Gao et al. (2024)	$\hat{\sigma}^2_{all}$ plugged	`clusterpval`
Cond-intra		Gao et al. (2024)	$\hat{\sigma}^2_{intra}$ plugged	`clusterpval`
Cond-oracle		González-Delgado et al. (2023)	Σ known	`PCIdep`
Cond-GD-all		González-Delgado et al. (2023)	$\hat{\Sigma}$ plugged	`PCIdep`
YFB	Conditional Approaches with **unknown** variance	Yun and Foygel Barber (2023)	σ^2 unknown	Simulations in Yun and Foygel Barber (2023)

Table 5.2 – Summary of the methods used to analyze the impact of the variance estimation in a spherical case. All methods use explicit p-values except for YFB-IS, which uses the p-value estimated by MC-Importance Sampling with $Q = 1000$ draws.

5.4.1.2 Results

Since the results are similar for the three tested values of σ^2, only those for $\sigma^2 = 1$ are reported in this manuscript. Yun and Foygel Barber (2023) have already numerically studied these methods for $n = 30$, $m = 2$, and $\sigma^2 = 1$, but they only compared a subset of the methods. Figure 5.7 (column $\rho = 0$) shows the ecdf curve of the p-values for each method under the null hypothesis while Figure 5.8 (column $\rho = 0$) shows the statistical power.

Under $\mathcal{H}_0$ (with no clustering structure), the estimator $\hat{\sigma}^2_{\text{intra}}$ underestimates the variance. As m increases, this estimator comes closer to the true value of σ^2, due to the reduced marginal effect of clustering. As demonstrated by Gao et al. (2024) and González-Delgado et al. (2023), plugging in a variance sub-estimator does not allow the method to control type I error (see Figure 5.7). The conditional approach with unknown variance (Yun and Foygel Barber (2023)) controls the type I error rate in our settings for exact p-values. Their version with MC-Importance Sampling method requires defining the sampled distribution's standard deviation γ. We used the same value $\gamma = 0.05$ as in their simulations. However, our results (which are described in more detail in Appendix B.3.3) indicate that this default choice leads to more and more conservative p-values as the number of variables m increases. In particular, around 90% of the Monte-Carlo samples lead to a p-value equal to 1 when $m = 100$ for $\gamma = 0.05$, and other choices for γ did not lead to properly calibrated p-values (see Figure B.8 and B.9). Therefore, the issue of tuning this hyperparameter should be addressed in order to make this method applicable.

Regarding statistical power, knowing the variance remains the condition for obtaining the most powerful method. The method of YFB (with unknown variance) is more powerful than plugging an estimator of the variance into the Gao et al. (2024)'s method. The MC-Importance Sampling estimation of the YFB method needs more draws than $Q = 1000$ to achieve the same statistical power level as the exact YFB test. The González-Delgado et al. (2023)'s method is less powerful than the one of Gao et al. (2024) since estimating Σ is more complex in spherical case than estimating σ^2.

5.4.2 Impact of dependence between variables

5.4.2.1 Setting and methods

The analysis is still conducted by simulating datasets from Setting 1 with $n = 500$ and $m \in \{2, 10\}$. In this case, the covariance matrix Σ is an auto-regressive matrix with $\rho \in \{0, 0.3, 0.5\}$ and $\sigma^2 = 1$. Table 5.3 summarizes methods used for this comparison. As noted in Section 5.2.2, only González-Delgado et al. (2023) give guarantees on the conditional approach using an over-estimation of Σ. Then, intra-covariance estimation

$$\hat{\Sigma}_{intra} = \sum_{k=1}^{K} \frac{|C_k|}{n} \left(\mathbf{x}_{C_k} - \bar{\mathbf{x}}_k\right)^\top \left(\mathbf{x}_{C_k} - \bar{\mathbf{x}}_k\right), \tag{5.34}$$

with $\mathbf{x}_{C_k}$ the matrix restricted to observations in C_k, is used without any guarantees. The conditional approach with unknown variance from Yun and Foygel Barber (2023) is tested even if the method supposes spherical covariance matrices.

Name	Method	Reference	Covariance Σ	Implementation
Cond-oracle	Conditional Approaches with **known** covariance	Gao et al. (2024)	Σ known	`clusterpval`
Cond-oracle		González-Delgado et al. (2023)	Σ known	`PCIdep`
Cond-GD-all		González-Delgado et al. (2023)	$\hat{\Sigma}$ plugged (5.25)	`PCIdep`
Cond-intra		González-Delgado et al. (2023)	$\hat{\Sigma}_{intra}$ plugged	`PCIdep`
YFB	Conditional Approaches with **unknown** variance	Yun and Foygel Barber (2023)	$\sigma^2 I_m$ unknown	Simulations in Yun and Foygel Barber (2023)

Table 5.3 – Summary of the methods used in the analysis of the impact of some correlations between variables

5.4.2.2 Results

Figure 5.7 shows the ecdf curves of the p-values under the null hypothesis. As expected, only methods taking into account the correlation between variables (the oracle approach with the known covariance matrix and the estimation $\hat{\Sigma}_{all}$) control the type I error rate. The estimator $\hat{\Sigma}_{intra}$ (see Equation (5.34)) underestimates the covariance, giving an invalid test. In the case of dependence between variables, the test by Yun and Foygel Barber (2023) does not control the type I error rate. While this result is consistent with the fact that this test has been developed specifically for spherical covariance, it indicates that this test is not robust to departures from the strong assumption of spherical covariance. Figure 5.8 illustrates the statistical power only for the valid methods. In the case of dependence, only the Cond-oracle and Cond-GD-all methods are evaluated. The same conclusion as before is drawn: overestimation of the covariance matrix results in a loss of statistical power.

5.5 Conclusion

The multivariate comparison of clusters is a central challenge in post-clustering inference. Two main groups of methods have been proposed to address this problem: information partitioning and conditional approaches. Our comparison of these methods, based on simulations of Gaussian data, shows that information partitioning (illustrated by data thinning) offers better statistical power than conditional approaches in the case of Gaussian data with a known spherical covariance matrix.

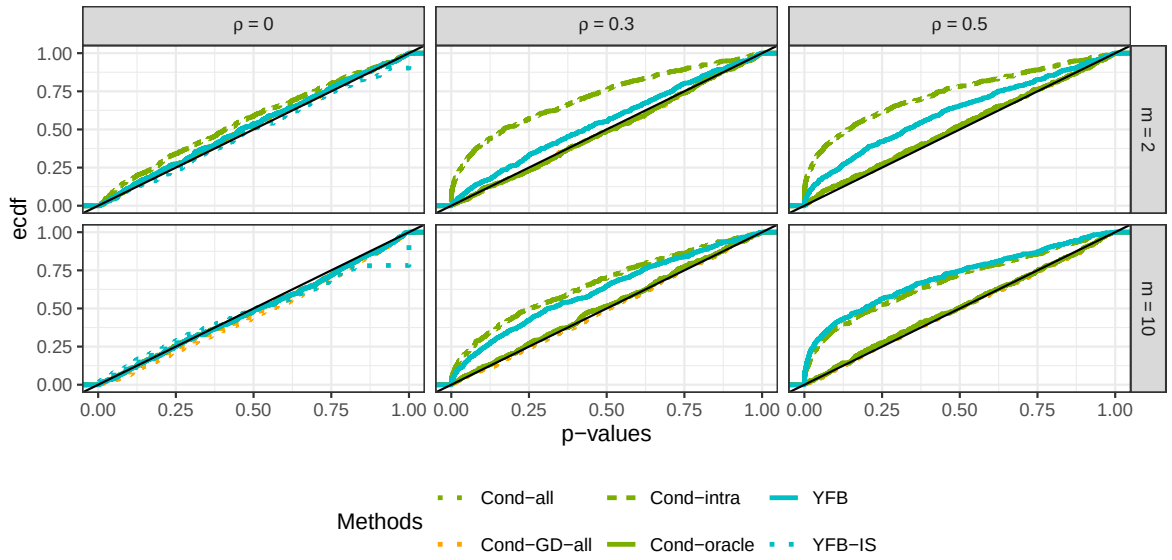

Figure 5.7 – ECDF of p-value under the null hypothesis for Setting 1 with $n = 500$, $m \in \{2, 10\}$. Each panel corresponds to the true value ρ of covariance. When $\rho > 0$, the methods "Cond-all" and "YFB-IS" are not evaluated.

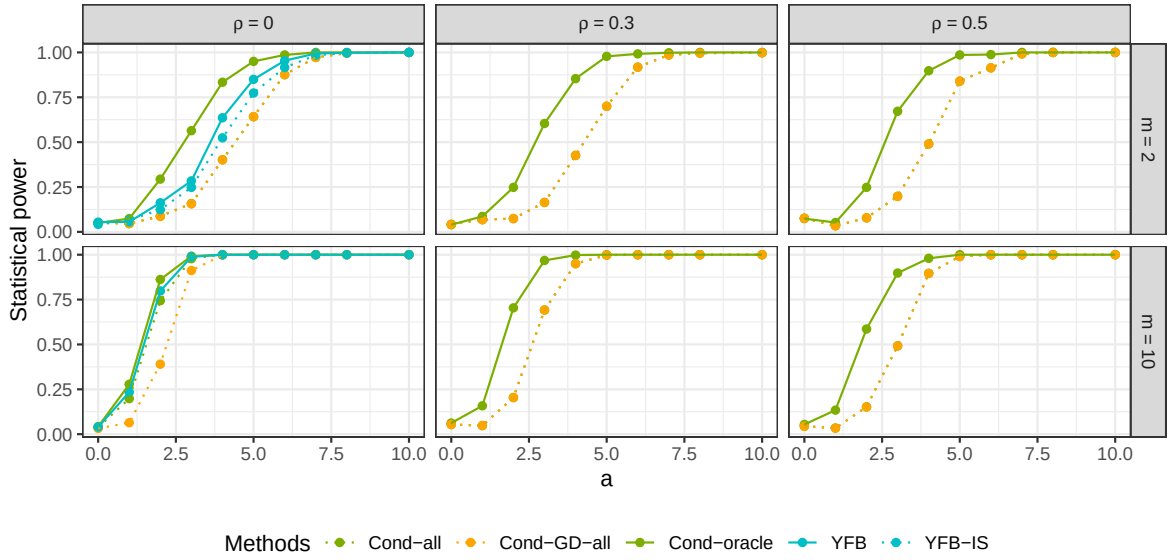

Figure 5.8 – Statistical power for Setting 1 with $n = 500$, $m \in \{2, 10\}$. Each panel corresponds to the true value ρ of covariance. Only methods offering proper control of type I error in Figure 5.7 are evaluated.

However, these methods are still limited by their parametric nature. Data thinning, although more flexible with respect to distribution types, requires its nuisance parameters (i.e. the covariance matrix in the Gaussian case) to be known. Neufeld et al. (2024a) have shown that miss-specifying the nuisance parameters of distribution implies $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$ are not independent. In this case, controlling the type I error rate is not guaranteed, as illustrated by Hivert et al. (2024b). The first conditional approaches also require the covariance matrix to be known. While solutions exist to bypass this constraint, such as covariance matrix estimation

or approaches adapted to unknown variance in the spherical case (Yun and Foygel Barber, 2023), these alternatives are less powerful. The use of estimators remains valid as long as the covariance matrix is overestimated (Gao et al., 2024; González-Delgado et al., 2023).

The inference part of the data thinning procedure has the advantage of being agnostic with respect to the clustering method, while conditional approaches provide explicit p-values for certain methods such as HAC and K-means. For other algorithms, estimation by MC-Importance Sampling increases computation time. However, for K-means, estimation by MC-Importance Sampling can in fact be more powerful and faster than its explicit calculation, demonstrating the limits of seeking an explicit calculation of the p-values by adding conditioning events.

Although these methods are promising, their application to real data remains very challenging because the assumption of a Gaussian distribution is often too restrictive. To date, only data thinning can be extended to other distributions, provided the associated nuisance parameters are known. This requirement limits its practical utility since the over or under estimation invalidates the method, as shown by (Hivert et al., 2024b).

Chapter 6

Univariate methods: review and numerical comparison

6.1 Introduction

In Chapter 5, the post-clustering inference question aims at comparing the multivariate means of two clusters. This chapter focuses on performing a marginal comparison between two clusters. This problem remains a post-clustering inference issue where (i) a multivariate clustering of observations is performed, and (ii) based on this clustering, two groups are compared along one of the dimensions.

This issue can arise in various applications. A commonly encountered case is in the analysis of scRNAseq data (see Lähnemann et al. (2020)), where gene expression is measured for several cells of a biological tissue. After a clustering of the cells, the genes having a difference of expression between clusters must be determined to characterize cell types. This step is called the detection of marker genes (see Section 1.1.2.2).

Like most methods in the literature on the subject, we consider a Gaussian framework. Let $\mathbf{X} = (X_i)_{i\in[|n|]}$ be the $n \times m$ matrix of samples. We assume that X_i are independent and for each $i \in [|n|]$, $X_i \sim \mathcal{N}_m(\mu_i, \Sigma)$ with $\mu_i \in \mathbb{R}^m$ and the covariance matrix $\Sigma \in \mathbb{R}^{m\times m}$. The mean vectors μ_i are grouped together in the matrix $\boldsymbol{\mu} = (\mu_i)_{i\in[|n|]} \in \mathbb{R}^{n\times m}$. Let $\mathcal{C}$ be a clustering method and $\mathcal{C}(\mathbf{X}) = \{C_1(\mathbf{X}), ..., C_K(\mathbf{X})\}$ be the associated partition of $\mathbf{X}$ into K clusters. The test aims to compare the behavior between some clusters on the marginal $j \in [|m|]$ of $\boldsymbol{\mu}$. Let $\boldsymbol{\mu}^{[j]} := (\mu_{1j}, ..., \mu_{nj})^\top$. In this chapter, we focus on the comparison between two clusters. Firstly, we define the test for two groups G_k and $G_{k'}$ for $k, k' \in [|K|], k \neq k'$ defined before seeing data. The null hypothesis is

$$\mathcal{H}_0 : \eta(G_k, G_{k'})^\top \boldsymbol{\mu}^{[j]} = 0, \tag{6.1}$$

where the contrast vector $\eta(G_k, G_{k'})$ depends on the two groups G_k and $G_{k'}$ defined before seeing the data $\mathbf{X}$.

Let $\mathcal{T}(\eta, \mathbf{X}_{\cdot j})$ be a test statistic that assesses the null hypothesis and depends on the contrast vector η and $\mathbf{X}_{\cdot j}$ the jth columns of $\mathbf{X}$. In this case, the test is not data-driven, (i.e. $\mathbf{X}$ is not used to estimate the clustering), the p-value for the observed data $\mathbf{x}$ is

$$p(\mathbf{x}) = \mathbb{P}_{\mathcal{H}_0}\left(\mathcal{T}(\eta(G_k, G_{k'}), \mathbf{X}_{\cdot j}) \geq \mathcal{T}(\eta(G_k, G_{k'}), \mathbf{x}_{\cdot j})\right) \tag{6.2}$$

and the test controls the type I error rate.

Example 2 (Marginal comparison in mean of two groups)**.** *Let G_k and $G_{k'}$ be two groups. Then the contrast vector $\eta(G_k, G_{k'})$ only depends on Groups G_k and $G_{k'}$ and can be written as Equation (5.3). To test the difference in mean on a variable $j \in [|m|]$ between the two groups, the null hypothesis is defined in Equation (6.1), which is equivalent to*

$$\mathcal{H}_0^{[j]} : \bar{\boldsymbol{\mu}}_{G_k}^{[j]} = \bar{\boldsymbol{\mu}}_{G_{k'}}^{[j]} \tag{6.3}$$

where $\bar{\boldsymbol{\mu}}_{G_k}^{[j]} = \frac{1}{|G_k|}\sum_{i\in G_k}\mu_{ij}$. *The test statistic* $\mathcal{T}(\eta, X_{.j}) = |\eta^\top \mathbf{X}_{.j}|$ *is considered. Then the p-value is*

$$p(\mathbf{x}) = \mathbb{P}_{\mathcal{H}_0}\left(|\eta(G_k, G_{k'})^\top \mathbf{X}_{.j}| \geq |\eta(G_k, G_{k'})^\top \mathbf{x}_{.j}|\right). \tag{6.4}$$

Considering $X_i \sim \mathcal{N}_m(\mu_i, \Sigma)$, *the statistic test* $|\eta(G_k, G_{k'})^\top \mathbf{X}_{.j}|$ *follows a Gaussian distribution under the null hypothesis as* $|\eta(G_k, G_{k'})^\top \mathbf{X}_{.j}| \overset{\mathcal{H}_0}{\sim} \mathcal{N}(0, \Sigma_{jj}\|\eta\|_2^2)$, *with* Σ_{jj} *the variance of the* j*th variable.*

Example 2 is a valid test when the compared groups are defined before observing the data. However, if the compared clusters are data-driven, i.e., Clusters $C_k(\mathbf{X})$ and $C_{k'}(\mathbf{X})$ are compared, the null hypothesis becomes

$$\mathcal{H}_0 : \eta(C_k(\mathbf{X}), C_{k'}(\mathbf{X}))^\top \boldsymbol{\mu}^{[j]} = 0. \tag{6.5}$$

The p-value in Equation (6.2) becomes

$$p(\mathbf{x}) = \mathbb{P}_{\mathcal{H}_0}\left(\mathcal{T}(\eta(C_k(\mathbf{X}), C_{k'}(\mathbf{X})), \mathbf{X}_{.j}) \geq \mathcal{T}(\eta(C_k(\mathbf{X}), C_{k'}(\mathbf{X})), \mathbf{x}_{.j})\right). \tag{6.6}$$

The distribution of the test statistic becomes unidentifiable because $\eta(C_k(\mathbf{X}), C_{k'}(\mathbf{X}))$ is random. The naive procedure is to wrongly consider that the contrast vector is not random and compute the p-value.

To address this problem, several solutions have recently been published, which can be divided into two categories: 1) methods based on information partitioning and 2) conditional approaches, as described in Chapter 5. Among the methods of the first category for marginal testing, Zhang et al. (2019) have proposed a procedure based on data splitting, commonly used in supervised learning. Leiner et al. (2023); Neufeld et al. (2024a) and Dharamshi et al. (2024) have proposed an information partitioning solution that can be applied to both global and marginal mean comparisons. In the second category, the methods condition on the clustering event $\{C_k(\mathbf{x}), C_{k'}(\mathbf{x}) \in \mathcal{C}(\mathbf{X})\}$ to control the type I error rate. Based on Gao et al. (2024), Chen and Gao (2023) and Hivert et al. (2024a) have adapted the previous results for the marginal mean comparison, while Bachoc et al. (2023) have proposed a specific solution for convex clustering. As in Chapter 5, the present work aims to comprehensively review the state-of-the-art methods, followed by a quantitative comparison to identify their limitations and explore potential extensions.

This chapter follows the same organization as Chapter 5. Section 6.2 is devoted to a review of post-clustering inference methods for marginal testing. A comparison through numerical experiments between methods under a known spherical covariance matrix assumption ($\Sigma = \sigma^2 I_m$) is addressed in Section 6.3. In Section 6.4, the numerical comparison is extended for a known general covariance matrix Σ or in the case where Σ is unknown.

6.2 Review of methods

6.2.1 Information partitioning

Information partitioning encompasses a set of methods that separate the information into two parts: one for clustering and the other for performing the statistical test. Naturally, data splitting is a method that addresses this goal. We described this procedure in Algorithm 1 (see Section 1.3.4). As explained and illustrated in that section, this procedure does not address the double dipping problem. Zhang et al. (2019) use this method for the marginal mean comparison to detect marker genes in scRNA-seq analysis. In practice, they use a Support

Vector Machine classification model to fit hyperplane using the clustering $\mathcal{C}(\mathbf{X}^{(1)})$ as labels. Then, they can predict the label of $\mathbf{X}^{(2)}$ and conduct a marginal test to determine which genes are markers between two clusters. This testing procedure compares the distributions of the two clusters, which are assumed to be truncated normal distributions.

The methods proposed by Leiner et al. (2023) and Neufeld et al. (2024a) can be employed independently of the clustering method and the test used. Specially, these methods (described in Section 5.2.1) can address the comparison of marginal means between two clusters. The data thinning procedure of Neufeld et al. (2024a) is described as follows. The data thinning procedure for the Gaussian distribution is applied for each observations $X_i \sim \mathcal{N}_m(\mu_i, \Sigma)$ to obtain two independent data sets such that

$$X_i^{(1)} \sim \mathcal{N}_m(\varepsilon\mu_i, \varepsilon\Sigma) \quad \text{and} \quad X_i^{(2)} \sim \mathcal{N}_m((1-\varepsilon)\mu_i, (1-\varepsilon)\Sigma). \tag{6.7}$$

The contrast vector is computed on $\mathbf{X}^{(1)}$ which is independent from $\mathbf{X}^{(2)}$. Then the p-value becomes

$$\mathbb{P}_{\mathcal{H}_0}\left(\mathcal{T}(\eta(C_k(\mathbf{x}^{(1)}), C_{k'}(\mathbf{x}^{(1)})), \mathbf{X}_{.j}^{(2)}) \geq \mathcal{T}(\eta(C_k(\mathbf{x}^{(1)}), C_{k'}(\mathbf{x}^{(1)})), \mathbf{x}_{.j}^{(2)})\right). \tag{6.8}$$

Following Example 2, the test statistic for the data thinning procedure can be

$$\mathcal{T}(\eta(C_k(\mathbf{x}^{(1)}), C_{k'}(\mathbf{x}^{(1)})), \mathbf{X}_{.j}^{(2)}) = \left|\eta(C_k(\mathbf{x}^{(1)}), C_{k'}(\mathbf{x}^{(1)}))^\top \mathbf{X}_{.j}^{(2)}\right|$$

which follows

$$\mathcal{N}\left((1-\varepsilon)\ \eta\left(C_k(\mathbf{x}^{(1)}), C_{k'}(\mathbf{x}^{(1)})\right)^\top \boldsymbol{\mu}^{[j]}, (1-\varepsilon)\Sigma_{jj}\left\|\eta(C_k(\mathbf{x}^{(1)}), C_{k'}(\mathbf{x}^{(1)}))\right\|_2^2\right).$$

6.2.2 Conditional approach inspired from Gao et al. (2024)

6.2.2.1 p-value through over-conditioning

In the case of a marginal comparison of means, the conditional method proposed by Gao et al. (2024) has been adapted by Hivert et al. (2024a) and Chen and Gao (2023). In both contributions, for a fixed variable $j \in [|m|]$, the null hypothesis in Equation (6.5) is considered and the test statistic is

$$\mathcal{T}(\eta, \mathbf{X}_{.j}) = |\eta^\top \mathbf{X}_{.j}|. \tag{6.9}$$

While Hivert et al. (2024a) use the conditional event $\{C_k(\mathbf{x}), C_{k'}(\mathbf{x}) \in \mathcal{C}(\mathbf{X})\}$ as Gao et al. (2024), Chen and Gao (2023) condition the p-value by a stronger event, since they fix the entire clustering partition as $\{\mathcal{C}(\mathbf{x}) = \mathcal{C}(\mathbf{X})\}$. In both methods, as in Gao et al. (2024), an orthogonal decomposition of $\mathbf{X}$ along η is used to obtain perturbed data $\tilde{\mathbf{x}}(\phi, j)$ for the tested variable $j \in [|m|]$ and for the perturbation ϕ which is defined as:

$$\tilde{\mathbf{x}}(\phi, j) = \mathbf{x} + \frac{\phi - \eta^\top \mathbf{x}_{.j}}{\|\eta\|_2^2}\ \eta\left(\frac{\Sigma_{.j}}{\Sigma_{jj}}\right)^\top \tag{6.10}$$

with $\Sigma_{.j}$ the jth columns of Σ and Σ_{jj} the variance of the variable j.

Following Gao et al. (2024), random decomposition elements that are not the statistic test are fixed in the p-value. Then, the p-value is:

$$p(\mathbf{x}; \{C_k, C_{k'}\}) = \mathbb{P}_{\mathcal{H}_0}\left(|\eta(C_k(\mathbf{x}), C_{k'}(\mathbf{x}))\mathbf{X}_{.j}| \geq |\eta(C_k(\mathbf{x}), C_{k'}(\mathbf{x}))^\top \mathbf{x}_{.j}|\ \middle|\mathcal{C}(\mathbf{x}) = \mathcal{C}(\mathbf{X}), \blacksquare\right) \tag{6.11}$$

$$\blacksquare = \left\{\mathbf{X} - \frac{\eta\Sigma_{.j}^\top(\eta^\top\mathbf{X})_j}{\|\eta\|_2^2\,\Sigma_{jj}} = \mathbf{x} - \frac{\eta\Sigma_{.j}^\top(\eta^\top\mathbf{x})_j}{\|\eta\|_2^2\,\Sigma_{jj}}\right\}$$

where, under the null hypothesis (6.5), $\eta(C_k(\mathbf{x}), C_{k'}(\mathbf{x}))^\top \mathbf{X}_{\cdot j} \sim \mathcal{N}\left(0, \Sigma_{jj} \left\|\eta(C_k(\mathbf{x}), C_{k'}(\mathbf{x}))\right\|_2^2\right)$. The method of Hivert et al. (2024a) is a sub-case of Chen and Gao (2023) since Hivert et al. (2024a) develop the methods in the case of independent variables. Then the value of $\Sigma_{\cdot j}/\Sigma_{jj}$ is a vector with 1 to the jth element and 0 otherwise and the perturbation of clusters is only made on the jth variable.

Moreover, Hivert et al. (2024a) show that the direct conditional test is not robust if there are clusters in-between the two tested clusters. The perturbation can disturb the clustering without mixing the two tested clusters and under-evaluate the separability between the tested clusters. To get around this problem, the authors suggest to merge p-values of adjacent in-between clusters. They define marginal in-between clusters as

$$C^j_{k:k'} := \left\{C_l, l \in [|K|] \mid \overline{\mathbf{X}}_{C_l,j} \in \left[\min(\overline{\mathbf{X}}_{C_k,j}, \overline{\mathbf{X}}_{C_{k'},j}), \max(\overline{\mathbf{X}}_{C_k,j}, \overline{\mathbf{X}}_{C_{k'},j})\right]\right\} \tag{6.12}$$

with $\overline{\mathbf{X}}_{C_k,j} = \frac{1}{|C_k|}\sum_{i\in C_k} X_{ij}$ the empirical mean of the cluster k on variable j. Then, the merged p-value is defined as the harmonic mean of the in-between p-values. This merging function, recommended by Vovk and Wang (2020), is robust to dependencies between p-values. Then

$$p^j_{k:k'} = \min\left(e\log(|C^j_{k:k'}|-1)\frac{|C^j_{k:k'}|-1}{\sum_{v=1}^{|C^j_{k:k'}|-1}\frac{1}{p^j_v}}, 1\right) \tag{6.13}$$

with p^j_v the p-value for the v-th test for adjacent clusters in $C^j_{k:k'}$.

6.2.2.2 How to compute the p-value?

To compute the p-value, Chen and Gao (2023) use the explicit computation proposed by Gao et al. (2024) and Chen and Witten (2023) for the HAC (with the squared Euclidean distance) and the K-means methods respectively (see Section 5.2.2.2 for more details). They do not develop an estimated p-value by Monte Carlo with Importance Sampling.

On the other hand, Hivert et al. (2024a) only use the estimation of the p-value by MC-Importance sampling. They aim to use in practice the Euclidean distance rather than the squared Euclidean distance which is one of the key to compute the exact p-value for the HAC clustering. The estimated p-value is inspired by the solution proposed by Gao et al. (2024) and reported in Section 5.2.2.2. Let $\omega_1, ..., \omega_Q \overset{ind}{\sim} \mathcal{N}\left(\left|\eta^\top \mathbf{x}^{[j]}\right|, \Sigma_{jj}\left\|\eta\right\|_2^2\right)$ and let $\pi_q = \frac{f_1(\omega_q)}{f_2(\omega_q)}$ be the importance sampling probabilities with $f_1(\cdot)$ is the density function of $\mathcal{N}(0, \Sigma_{jj}\left\|\eta\right\|_2^2)$ and $f_2(\cdot)$ is the density function of $\mathcal{N}(\left|\eta^\top \mathbf{x}^{[j]}\right|, \Sigma_{jj}\left\|\eta\right\|_2^2)$. Then, the p-value can be estimated as

$$p(\mathbf{x}; \{C_k, C_{k'}\}) \approx \frac{\sum_{q=1}^{Q} \pi_q \mathbb{1}_{\{|\omega_q| \geq \left|\eta^\top \mathbf{x}^{[j]}\right|, C_k, C_{k'} \in \mathcal{C}(\tilde{\mathbf{x}}(\omega_q, j))\}}}{\sum_{q=1}^{Q} \pi_q \mathbb{1}_{\{C_k, C_{k'} \in \mathcal{C}(\tilde{\mathbf{x}}(\omega_q, j))\}}}. \tag{6.14}$$

With the HAC using the squared Euclidean distance and in the case of the diagonal covariance matrix, the method of Hivert et al. (2024a) gives an estimation of the p-value of Chen and Gao (2023), since for the HAC method, both conditions $C_k, C_{k'} \in \mathcal{C}(\tilde{x}(\phi))$ and $\mathcal{C}(x) = \mathcal{C}(\tilde{x}(\phi))$ are equivalent (see Lemma 1 of Gao et al. (2024)).

6.2.2.3 Estimation of the covariance matrix

The theoretical development of the marginal conditional approach supposes to know the covariance matrix Σ, but in practice, it is often unknown. Hivert et al. (2024a), assuming independence between the variables, propose to estimate the variance of the variable j, Σ_{jj} using only observations from the two compared clusters. This plug-in variance is defined as:

$$\hat{\Sigma}_{jj} = \frac{1}{|C_k| + |C_{k'}| - 1} \sum_{i \in C_k \cup C_{k'}} \left(X_{ij} - \bar{X}_{C_k \cup C_{k'}, j} \right)^2 \tag{6.15}$$

with $\bar{X}_{C_k \cup C_{k'}, j} = \frac{1}{|C_k| + |C_{k'}|} \sum_{i \in C_k \cup C_{k'}} X_{ij}$. They argue that this estimator may underestimate the variance of $\mathbf{X}_{.j}$ in some instances, particularly when clustering introduces significant artificial differences. Through numerical experiments, they demonstrate that overestimating the variance leads to conservative p-values while underestimating it fails to control the type I error rate. However, these observations have not been theoretically validated.

In the context of the merged test, where the p-value is given by Equation (6.13), the variance estimator for variable j is restricted to observations contained within $C^j_{k:k'}$ (see Equation (6.12)), such that:

$$\hat{\Sigma}_{jj} = \frac{1}{\left| C^j_{k:k'} \right| - 1} \sum_{i \in C^j_{k:k'}} \left(X_{ij} - \bar{X}_{C^j_{k:k'}, j} \right)^2$$

with $\bar{X}_{C^j_{k:k'}, j} = \frac{1}{|C^j_{k:k'}|} \sum_{i \in C^j_{k:k'}} X_{ij}$.

Chen and Gao (2023) do not address the issue of variance estimation in their theoretical development, nor in their simulated data where the covariance matrix is assumed to be known. In their scRNAseq application, they employ the sample covariance matrix estimator $\hat{\Sigma} = \frac{1}{n-1} \left(\mathbf{X} - \bar{\mathbf{X}} \right)^\top \left(\mathbf{X} - \bar{\mathbf{X}} \right)$, without discussing the guarantees associated with this estimator.

6.2.3 Post convex clustering inference for a marginal test

Lee et al. (2016) have introduced an exact post-selection inference procedure in the lasso regression framework. Bachoc et al. (2023) have adapted this development to obtain a post-clustering inference method for a partition obtained from a convex clustering procedure (Pelckmans et al., 2005), being the closest method to a lasso regression problem. The first step of this method involves computing the uni-dimensional convex clustering for all variables for a fixed value of λ. For the variable $j \in [|m|]$, the first step is to solve

$$\hat{B}_{.j}(\mathbf{X}_{.j}) = \underset{B_{.j} \in \mathbb{R}^n}{\arg\min} \; \frac{1}{2} \left| \mathbf{X}_{.j} - B_{.j} \right|_2^2 + \lambda \sum_{\substack{i,i'=1 \\ i<i'}}^{n} \left| B_{ij} - B_{i'j} \right| .$$

The associated uni-dimensional clustering $\mathcal{C}^{[j]}(\mathbf{X}_{.j})$ is then defined as follows: observations i and i' are in the same cluster $C_k^{[j]}$ if $B_{ij} = B_{i'j}$.

The second step of this procedure aims to only obtain one partition of the n observations. Let $\{\mathcal{C}^{[j]}(\mathbf{X}_{.1}), ...\mathcal{C}^{[m]}(\mathbf{X}_{.m})\}$ be the unidimensional clusterings obtained on the first step. Estimating clustering on a single variable preserves the order of individuals on that variable and induces an order relationship between clusters. Thus, unidimensional clusterings are considered as ordinal data where the clustering $\mathcal{C}^{[j]}(\mathbf{X}_{.j})$ is a n-size vector of L modalities (corresponding to the cluster labels). In ordinal variable, modalities $\{r_1, ..., r_L\}$ have an order relationship with each other $r_1 \prec r_2 \prec ... \prec r_L$. Each clustering has a different number of

modalities L depending on the value of λ, the signal and the dispersion into the associated variable. To obtain a partition of observations considering all unidimensional clusterings, a clustering method for ordinal data O is computed as $\mathcal{C} := O(\mathcal{C}^{[1]}, ..., \mathcal{C}^{[m]})$. A set of solutions is available and discussed in Appendix B.2 below. We adopt the rank normalization approach described by Marden (1996). Let Y_{ij} be a categorical variable with L known ordered categories encoded as $1, \dots, L$. The normalized ordinal data is given by $\tilde{Y}_{ij} = \frac{Y_{ij}-1}{L-1} \in [0, 1]$. This transformation makes $\tilde{Y}_{ij}$ continuous, allowing to use the Euclidean distance and clustering methods HAC with Ward linkage or K-means.

Thus, the procedure consists of two clustering steps: the first for each variable; the second on the m clusterings obtained in the first step. The p-value is conditioned by the clustering and the data order. This condition is rewritten in order to apply the polyhedral lemma of Lee et al. (2016) in order to obtain an exact conditional p-value. The test statistic is based on a truncated cumulative distribution function of a Gaussian distribution, which follows a Uniform distribution under the null hypothesis. Then, the conditional p-value can be explicitly computed. Note that, Bachoc et al. (2023) over-condition the p-value to obtain an exact p-value. The same strategy is used by Gao et al. (2024) with an other conditioning event (as described in Section 6.2.2).

6.2.4 Multimodality test

The multimodality test proposed by Hivert et al. (2024a), uses the Dip test of Hartigan and Hartigan (1985) to address the question of marginal comparison of clusters. We consider the null hypothesis that the global distribution of the tested clusters (and the in-between clusters) is unimodal. This method assumes that clusters incorrectly split under the null hypothesis are sufficiently close for the marginal distribution of their union to remain unimodal. Hivert et al. (2024a) use this test by selecting marginally in-between observations in the two tested clusters $C^{[j]}_{k:k'}$ in Equation (6.12). The Dip statistic is defined by

$$\text{dip}(F) = \min_{G \in \mathcal{U}} \sup_{x} |F(x) - G(x)| \tag{6.16}$$

with F the cumulative distribution function of the data and $\mathcal{U}$ the class of unimodal distributions. The associated p-value is

$$p := \mathbb{P}\left(\text{dip}\left(U_{n_{k:k'}}\right) \geq \text{dip}\left(F_{X^{[j]}_{k:k'}}\right)\right). \tag{6.17}$$

Here, $U_{n_{k:k'}}$ represents the empirical cumulative distribution function drawn from $n_{k:k'} := \left|C^{j}_{k:k'}\right|$ samples of the standard uniform distribution and $F_{X^{[j]}_{k:k'}}$ is the observed empirical cumulative marginal distribution function for the samples in $C^{j}_{k:k'}$ for the variable j. Hartigan and Hartigan (1985) have demonstrated that the Uniform distribution has the largest asymptotic Dip statistic among the set of unimodal distributions. If the data distribution has at least two modes, the procedure concludes that there is a difference between the two tested clusters. However, this method is only valid when the marginal distribution of each cluster is unimodal and does not account for clusters with marginal multimodal distributions.

6.3 Numerical comparisons for a spherical covariance matrix

Most of the literature on post-clustering inference relies on marginal tests after performing multidimensional clustering (see Example 2). This section aims to compare these marginal post-clustering inference methods numerically. Statistical performances are evaluated by controlling the type I error rate and then the statistical power.

6.3.1 Simulation setting

To analyze the methods, Settings 1 and 2 described in Section 5.3.1.1 are used to explore methods and their limits. However, only Setting 2 is reported, as the interpretations of methods in Setting 1 do not differ from those in Setting 2. Table 6.1 summarizes the marginal distances between the $K = 3$ clusters. The first feature allows to distinguish the three clusters, the second gives the same marginal distribution for Clusters C_1 and C_2, while the third carries no signal. The setting is computed with $m = 3$, $n \in \{10, 20, 50, 100, 200, 500\}$ but only $n = 100$ is reported in this manuscript as other values of n do not affect the interpretations.

	C_1 vs C_2	C_1 vs C_3	C_2 vs C_3
$\eta^\top \boldsymbol{\mu}^{[1]}$	a	$a/2$	$a/2$
$\eta^\top \boldsymbol{\mu}^{[2]}$	0	$\sqrt{3}a/2$	$\sqrt{3}a/2$
$\eta^\top \boldsymbol{\mu}^{[3]}$	0	0	0

Table 6.1 – True marginal distance between means of two clusters in Setting 2.

For marginal comparison, Setting 2 does not provide interchangeable comparisons. Therefore, cluster label switches can falsify interpretations. We fix this issue with a solution based on the knowledge from the simulation setting. Since $X_i \sim \mathcal{N}(\mu_i, \sigma^2 I_m)$, we can calculate the true mean associated with Cluster C_k, $\bar{\boldsymbol{\mu}}_{C_k(\mathbf{x})} = \frac{1}{|C_k(\mathbf{x})|} \sum_{i \in C_k(\mathbf{x})} \mu_i$, enabling correct cluster labeling.

Three cases stand out: (i) there is no signal in the data, no clustering to find, and marginally the tests are under the null hypothesis (Global null hypothesis), (ii) there is a signal in the data and clustering can be found, but marginally, the test is under the null hypothesis (Marginal null hypothesis), and (iii) there is a signal in the data, clustering can be found, and the test is under the alternative hypothesis. Case (i) is simulated with $a = 0$, while the other cases are simulated with $a \in \{1, 2, 3, 4, 5, 6, 8, 10\}$.

Name	**Method**	**References**	**Shape of Σ (known)**	**Computation of p-value**	**Implementation**
Naive	Naive	-	General	Explicit	-
TN	Data splitting	Zhang et al. (2019)	General	Explicit	`tn_test`
DT	Data thinning	Neufeld et al. (2024a)	General	Explicit	Based on `datathin`
CG	Conditional Approaches	Chen and Gao (2023)	General	Explicit	`CADET`
H		Hivert et al. (2024a)	Diagonal	IS	`VALIDICLUST`
H-merge		Hivert et al. (2024a)	Diagonal	IS & Merged	`VALIDICLUST`
poclin		Bachoc et al. (2023)	General	Explicit	`poclin`
Dip	Multimodality test	Hivert et al. (2024a)	Diagonal	Explicit	`VALIDICLUST`

Table 6.2 – Summary of the methods used to compare marginal methods.
Shape of Σ: "General" means that any covariance structure can be used for the covariance matrix Σ. "Diagonal" means that the covariance structure is diagonal s.t. $\Sigma = \text{diag}(\{\sigma_j^2\}_{j \in [|m|]})$.
Computation: "IS" stands for an estimation by MC-Importance Sampling. "Merged" refers to the merge test in the presence of intermediate clusters.

The methods used are compiled in Table 6.2. To compare with the conditional methods of Chen and Gao (2023) and Hivert et al. (2024a), the Z test (described in Example 2) is used for the naive method and data thinning (see Section 6.2.1). The data thinning method is parameterized by $\varepsilon = 0.7$, for the same reasons outlined in Section 5.3.3. Chen and Gao (2023) only proposed an explicit test for a general covariance matrix, while Hivert et al. (2024a) proposed an estimated test by MC-Importance Sampling for independent variables

(see Section 6.2.2). The MC-Importance Sampling is computed with $Q = 1000$ draws. For Setting 2 with $K = 3$, the merged version of the test (described in Equation (6.13)) is different from the direct test. The conditional approach of Bachoc et al. (2023) for convex clustering, presented in Section 6.2.3, is employed for $n = 100$ with $\lambda \approx 0.0136$ (obtained according to the practical recommendation of the authors).

For this comparison, even if we supposed that the data splitting method (Zhang et al., 2019) is not valid for post clustering inference (see Section 1.3.4) and the multimodality test of Hivert et al. (2024a) do not take into account the clustering, both methods are compared. All these methods test clusters obtained via Hierarchical Ascendant Clustering (with Ward linkage), except for the method of Bachoc et al. (2023), which is used with convex clustering on each dimension and aggregate ordinal clusterings with HAC on rescaled data by normalized rank method (see Section 6.2.3).

6.3.2 Evaluation of type I error rate

Firstly, the methods are tested under the global null hypothesis. In this case, Setting 2 is simulated with $a = 0$. Figure 6.1 shows the ecdfs of the p-values for the global null hypothesis (on the column "a = 0") and comparisons for Variables 2 and 3 (in rows). All comparisons across all variables yield the same interpretations as those presented here. As expected, and shown by Hivert et al. (2024a) and Chen and Gao (2023) and in Section 5.3.2, the naive method does not control the type I error rate. The TN method also does not control the type I error rate: as discussed in Section 6.2.1, data splitting does not provide a direct solution for post-clustering inference. All other methods control the type I error rate. Note that the multimodality test proposed by Hivert et al. (2024a) is conservative because the Gaussian distribution of samples is "more" unimodal than the uniform distribution (which is the distribution under the null hypothesis in the Dip test). The merged test proposed by Hivert et al. (2024a) is more conservative than the direct test because of the merging formula (Equation (6.13)).

Secondly, the methods are tested on comparisons under the marginal null hypothesis. It concerns the comparison of Clusters C_1 and C_2 on the second marginal (first row of Figure 6.1) and all comparisons on the third marginal (whose results are grouped on the second row of Figure 6.1), where Setting 2 is simulated with a variation of $a \in \{1, 2, 3, 4, 5, 6, 8, 10\}$. All previous conclusions hold. Additionally, the naive method controls the type I error rate when there is enough signal. Indeed, in these cases, the clusters are easily recovered when the signal in the data is sufficiently strong (see the ARI curves in Figure 6.2). As a result, the marginal distributions of the obtained clusters are similar, allowing the naive p-value to be under the null hypothesis. But this result in a specific case does not prove that the naive method is valid.

Furthermore, for intermediate signal values such as $a \in \{2, 3, 4\}$, the conditional approaches of Chen and Gao (2023) and Hivert et al. (2024a) (direct and merged test) and the data thinning method (Neufeld et al., 2024a) have p-values stochastically smaller than $\mathcal{U}[0, 1]$, indicating that they do not seem to control the type I error rate. However, as seen in Figure 6.2, the clusterings are not easily recovered for these signal values. The unsupervised nature of the problem leads issue to identify the label of the clusters and then identify the true hypothesis of the comparison. Moreover, the obtained clusters are marginally different enough for their comparison not to be considered under the null hypothesis. Thus, the methods remain valid.

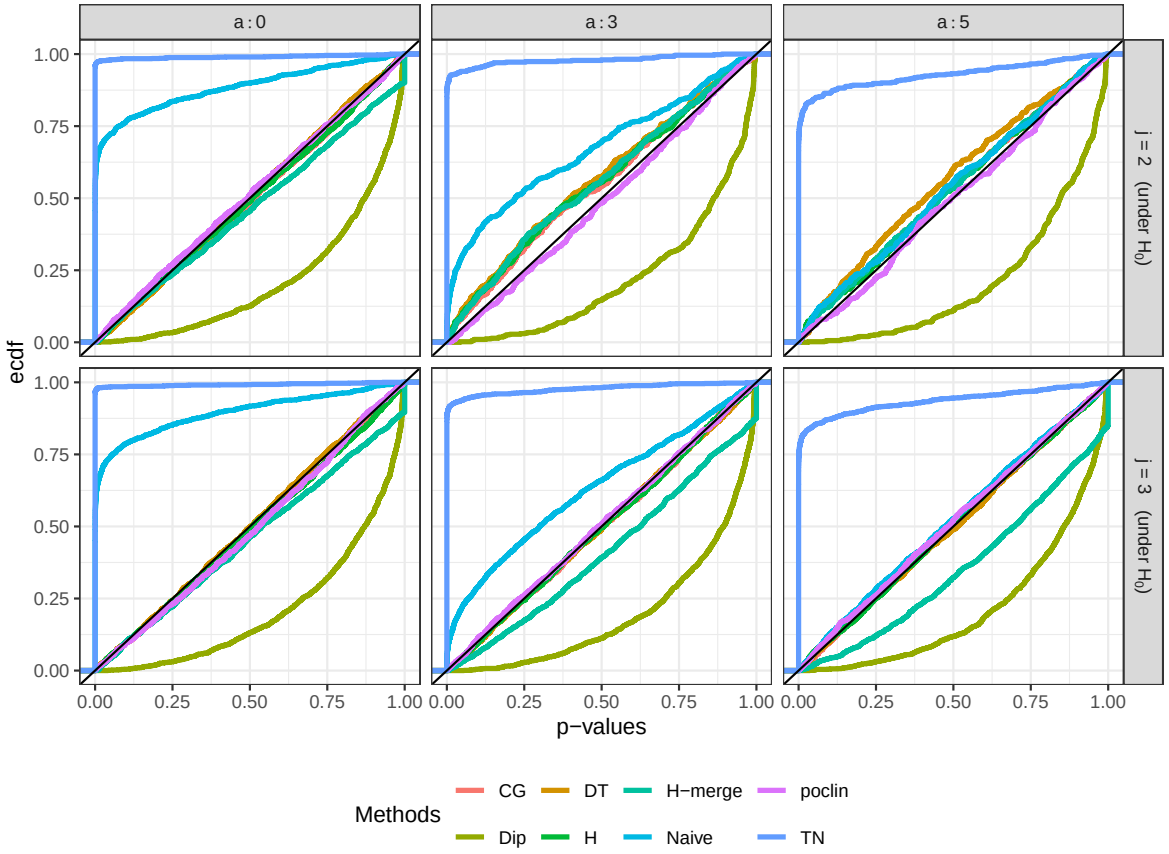

Figure 6.1 – Evaluation of the type I error rate: ecdf of p-values in the case of the marginal null hypothesis of Setting 2 with $n = 100$. The value of signal a drives the distinction between the true clusters. On the second variable, only the comparison between C_1 and C_2 is under the null hypothesis. On the third variable, all comparisons are under the null hypothesis.

6.3.3 Statistical power

The statistical power of the methods is analyzed in this section. Setting 2 is simulated by varying the signal value $a \in \{1, 2, 3, 4, 5, 6, 8, 10\}$. The comparisons under the alternative hypothesis are all those of the first marginal and the comparison 'C_1 vs C_3' and 'C_2 vs C_3' for the second variable (see Table 6.1 for details and values of the theoretical true distances). Figure 6.3 represents the power as a function of the initial signal a. Each panel groups the comparisons with the same theoretical distance value for each variable. As already observed for multivariate methods analyzed in Section 5.3.3, data thinning is more powerful than other methods. The conditional approach of Chen and Gao (2023) and the direct test of Hivert et al. (2024a) have the same statistical power, since using the HAC algorithm, both methods are equal (as explained in Section 6.2.2). Note that implementing the merged test by Hivert et al. (2024a) requires using an estimator of the variance restricted to observations contained in compared clusters. Thus, the method loses statistical power due to this estimation. The increase in statistical power for the comparisons C_1 vs C_3 and C_2 vs C_3 on the first marginal (such that the true difference in means is $a/2$) is due to a better estimation of the variance than the estimation made for the comparison C_1 vs C_2 (such that the true difference in means is a). The multimodality test needs enough distance between modalities in the distribution of samples as observed by Hivert et al. (2024a). In Setting 2, the distance between modalities is

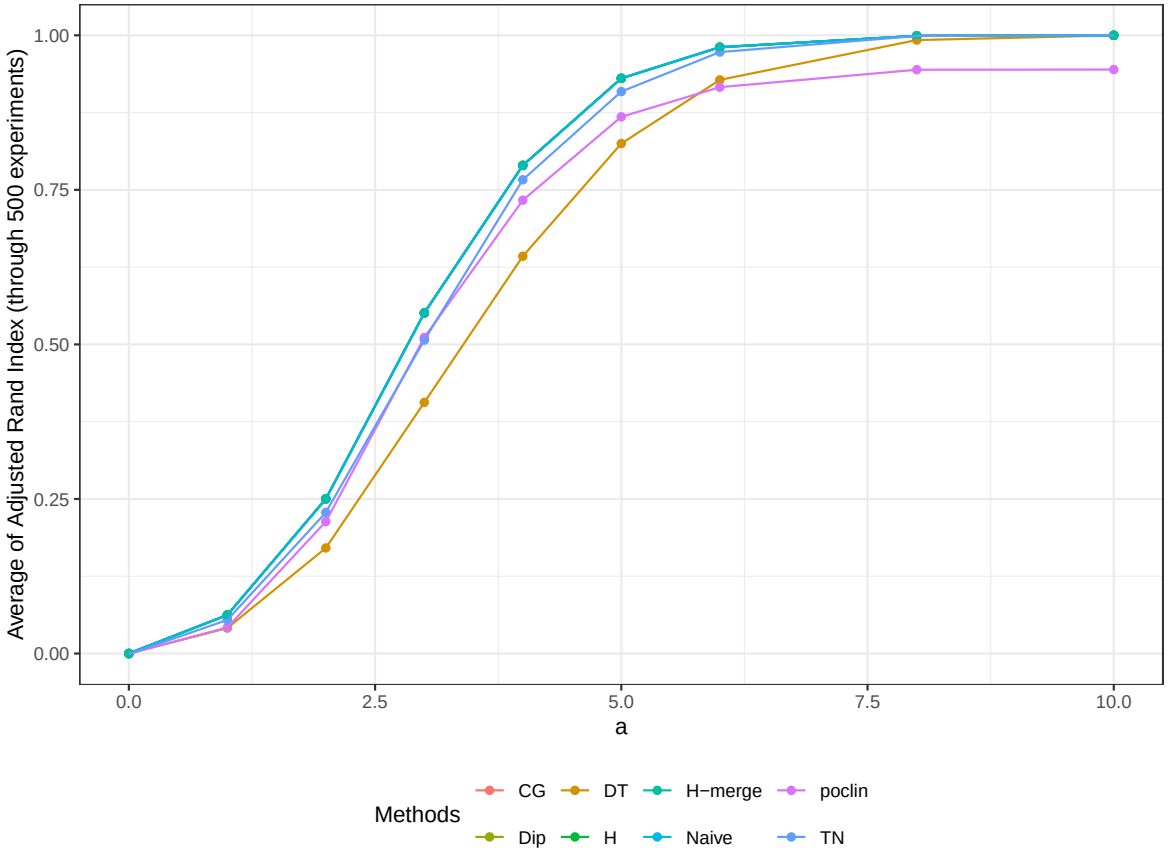

Figure 6.2 – Average ARI as a function of the signal a for $n = 100$ using Setting 2. The CG, Dip, H, H-merge, and Naive curves overlap as the same clustering method is applied to the entire dataset. In contrast, TN and DT estimate the clustering based on only a subset of the data, limiting their ability to recover the correct partition accurately. The poclin method applies convex clustering to each dimension followed by HAC on rescaled unidimensional ordinal clusterings, leading to information loss due to this double clustering process.

$a/2$ for all the comparisons on the first variable while $\sqrt{3}a/2$ for the second variable explaining the difference in statistical power. The conditional approach proposed by Bachoc et al. (2023) exhibits limited statistical power. If the marginal variable considered distinguishes all clusters (e.g., $j = 1$), the method can detect signal if the clusters are separated enough. On the contrary, when the variable fails to distinguish at least two clusters (e.g., $j = 2$), then the method fails to detect signal in other comparisons.

6.4 Numerical comparison for general covariance matrix Σ and its estimation

6.4.1 Setting and methods

Previous methods require knowledge of the covariance matrix. Furthermore, the methods proposed by Hivert et al. (2024a) require the covariance matrix to be diagonal, a theoretical scenario rarely encountered in practice. This section aims to investigate the performance of

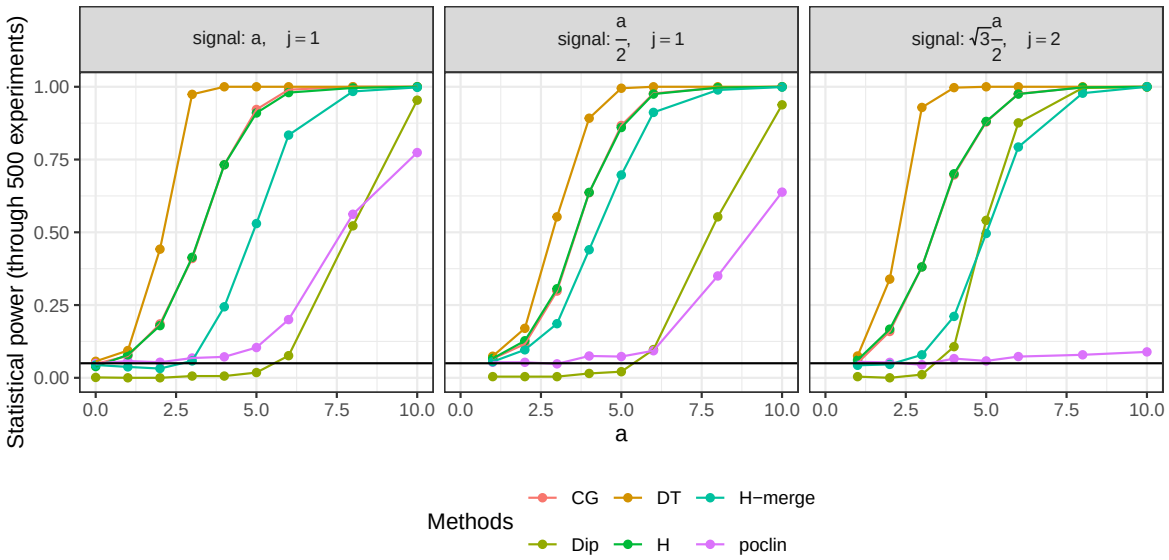

Figure 6.3 – Statistical power of methods using Setting 2 for different distance values between means of clusters. Panels represent the true signal between clusters and the tested variable. Corresponding comparisons for each panel are reported in Table 6.1.

methods in more realistic scenarios where the covariance structure is less constrained and/or unknown.

The simulations are again derived from Setting 2, as described previously. We define a single dependence between variables 1 and 3 as follows:

$$\Sigma = \begin{pmatrix} 1 & 0 & \rho \\ 0 & 1 & 0 \\ \rho & 0 & 1 \end{pmatrix},$$

with $\rho \in \{0, 0.3, 0.5\}$. The case $\rho = 0$ allows for evaluating the estimators compared to previous analyses.

This analysis only explores the valid and powerful methods seen earlier. They are summarized in Table 6.3. The methods of Chen and Gao (2023), and Data thinning accept a general but known covariance matrix. Both methods are used with the true value of Σ as well as its global estimation $\hat{\Sigma}_{\text{all}} = \frac{1}{n-1}(\mathbf{X} - \bar{\mathbf{X}})^\top(\mathbf{X} - \bar{\mathbf{X}})$. The methods are challenged by misspecifying the dependence between variables using I_m and $\hat{\sigma}^2 I_m$, with $\hat{\sigma}^2$ being the global estimation of the variance (see Equation (5.23)). The conditional methods of Hivert et al. (2024a) require independence between observations. Thus, the test is misspecified for cases $\rho \neq 0$. The test for multimodality is non-parametric. It is included to study how dependence affects its results.

6.4.2 Results

Firstly, the type I error rate is evaluated under the global null hypothesis (with $a = 0$). Given the assumption of the global null hypothesis, where variables 1 and 3 are interchangeable and variable 2 is independent of others (interpretations align with Section 6.3.2), only results for $j = 3$ are presented Figure 6.4, on the first column "$a : 0$". Under dependence, using misspecification of Σ (wrong assumption of independence), Chen and Gao (2023), data thinning and the conditional approach by Hivert et al. (2024a) do not control the type I error

rate. Neufeld et al. (2024a) demonstrates that incorrect estimation of the nuisance parameter (here Σ) renders the thinned datasets dependent, potentially invalidating the test. But, as $\hat{\Sigma}_{\text{all}}$ under the null hypothesis is a good estimator, the test controls the type I error rate. Other methods (Dip test, data thinning with Σ and Chen and Gao (2023), with Σ and $\hat{\Sigma}_{\text{all}}$) observe the previously noted control of the type I error rate.

Secondly, the type I error rate assessment is evaluated under the marginal null hypothesis. Variable 2 is independent of other variables, and as observed in Section 6.3.2, the analysis of ecdf introduces some uncertainties due to clustering. Therefore, only marginal tests on variable 3 are reported (see Figure 6.4). The interpretations of the methods are similar to those in the previous point. The methods become conservative in the presence of a sufficient signal ($a \geq 5$) due to good clustering (see Figure 6.5), as observed with the naive method in Section 6.3.2. Additionally, note that the method of Chen and Gao (2023) becomes conservative with increasing a when using $\hat{\sigma}^2 I_m$. This method appears poorly calibrated in this case.

Thirdly, statistical power is studied for valid and powerful methods, and presented in Figure 6.6. Estimating Σ reduces the power of data thinning and the method of Chen and Gao (2023). This effect is particularly pronounced for data thinning. The use of $\hat{\Sigma}_{\text{all}}$ impacts the separation of the data, and thus the clustering and statistical testing. Consequently, the clustering results lose power with an estimation of Σ (see Figure 6.5). This seems to explain the loss of power of data thinning with $\hat{\Sigma}_{\text{all}}$. A more in-depth study would be necessary to fully understand the impacts of the covariance matrix and its estimation on data thinning. Note that, as expected, the Dip test is not impacted by the correlation between variables.

Name	**Method**	**References**	**Shape of Σ (known)**	**Σ used**	**Implementation**
DT DT-estim DT-Id DT-sph-estim	Data thinning	Neufeld et al. (2024a)	General	Σ $\hat{\Sigma}_{\text{all}}$ I_m $\hat{\sigma}^2 I_m$	Based on `datathin`
CG CG-estim CG-Id CG-sph-estim	Conditional Approaches	Chen and Gao (2023)	General	Σ $\hat{\Sigma}_{\text{all}}$ I_m $\hat{\sigma}^2 I_m$	`CADET`
H H-estim H-merge-estim	Conditional Approaches	Hivert et al. (2024a)	Diagonal	$\text{diag}(\{\sigma_j^2\}_{j\in[\|m\|]})$ $\text{diag}(\{\hat{\sigma}_j^2\}_{j\in[\|m\|]})$ $\text{diag}(\{\hat{\sigma}_j^2\}_{j\in[\|m\|]})$	`VALIDICLUST`
Dip	Multimodality test	Hivert et al. (2024a)	Diagonal	-	`VALIDICLUST`

Table 6.3 – Summary of the methods used to compare marginal methods under dependence. Shape of Σ: "General" means that any covariance structure can be used for the covariance matrix. "Diagonal" means that the covariance structure is diagonal s.t. $\Sigma = \text{diag}(\{\sigma_j^2\}_{j\in[|m|]})$. The methods can account for different values of Σ used. Σ corresponds to the true known covariance matrix. $\hat{\Sigma}_{\text{all}}$ corresponds to its estimation using all samples (see Equation (5.25) with $U = I_n$). I_m corresponds to using the identity matrix. $\hat{\sigma}^2 I_m$ corresponds to a spherical estimation (using estimation of σ^2 in Equation (5.23). '-' means that the variance is not required in the test. Only the marginal variance is required for the methods of Hivert et al. (2024a). σ_j corresponds to the known marginal variance. $\hat{\sigma}_j^2$ corresponds to its estimation only for the observations contained within the tested clusters (see Equation (6.15)).

6.5 Conclusion

This chapter studies the problem of marginal testing between two clusters, a classical case of post-clustering inference. A comparison between methods has been made on Gaussian data to be consistent with theoretical assumption of these methods. The conclusion is consistent with those of Chapter 5: data thinning proves more powerful than conditional approaches when the data distribution is known.

However, knowledge of the covariance matrix remains critical for achieving well calibrated and powerful tests. For data thinning, the estimation of this matrix leads to reduced cluster recovery and diminished test power. In contrast, conditional tests rely on clustering performed on the entire dataset, with only the testing phase of the procedure being affected, limiting the loss in statistical power.

Theoretical gaps persist in the development of marginal conditional approaches inspired by Gao et al. (2024), particularly concerning the impact of using covariance matrix estimations in conditional tests. Moreover, Chen and Gao (2023) do not consider the estimation of the p-value using MC-Importance Sampling for an arbitrary but known covariance matrix. This procedure cannot be applied to any clustering method. Moreover, the work of Yun and Foygel Barber (2023), which considers unknown variance in the context of multivariate testing, has yet to be extended to the case of marginal mean comparison between two clusters.

The conditional approach for convex clustering proposed by Bachoc et al. (2023) exhibits limited power. Furthermore, introducing a third untested cluster can make the test overly

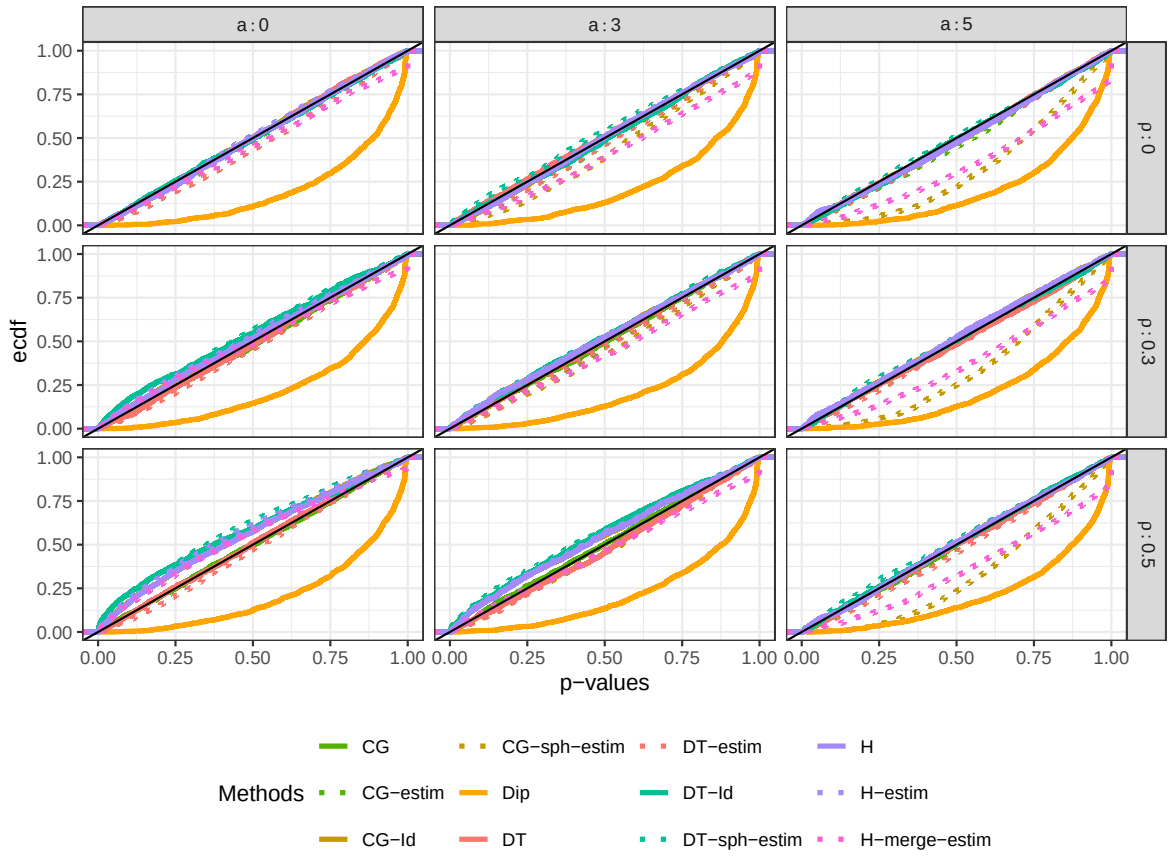

Figure 6.4 – Evaluation of the type I error rate: ecdf of p-values for methods studied in case of dependence and summarized in Table 6.3. Each panel corresponds to a value of ρ and a.

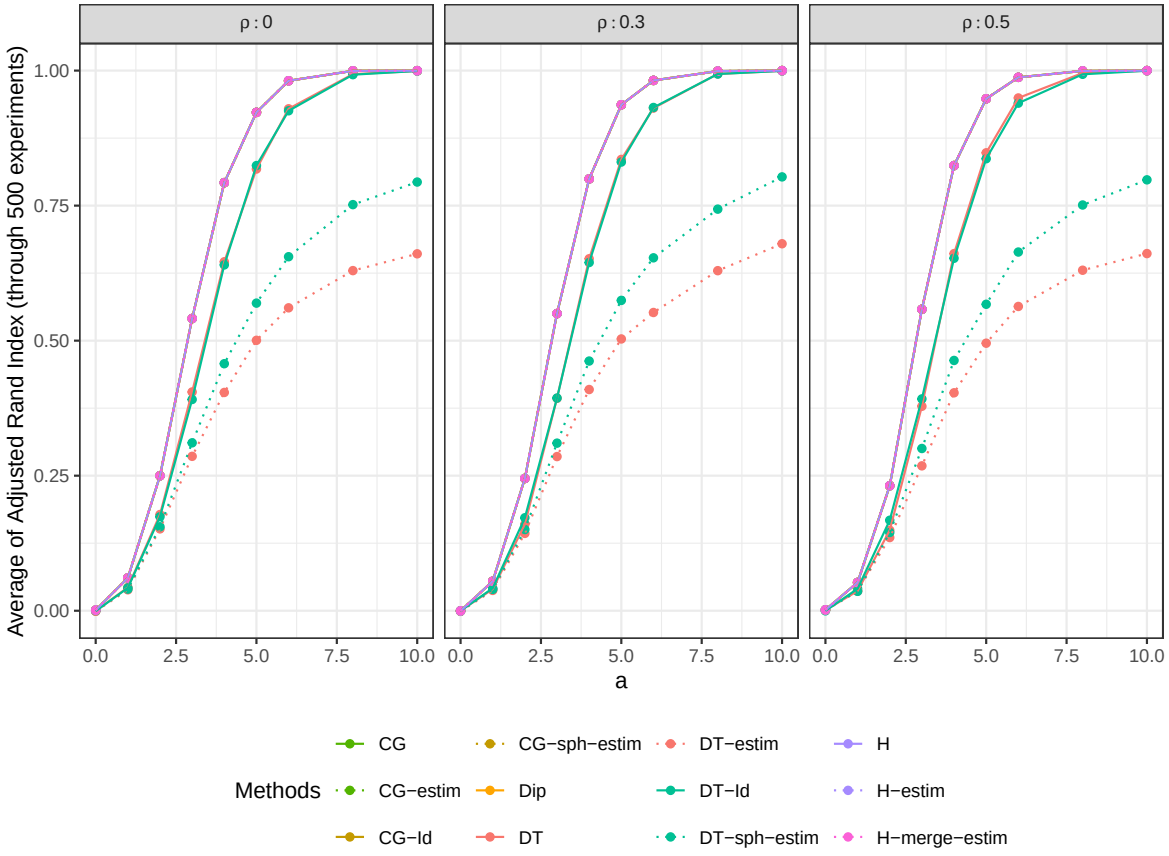

Figure 6.5 – ARI of valid marginal methods in the analysis of dependence. Only clusterings obtained with data thinning are impacted by the covariance matrix.

conservative. These results can be explained by fact that conditioning on the ordering of each input variable is very demanding.

The multimodal test proposed by Hivert et al. (2024a) is an interesting solution for post-clustering inference. However, it is only relevant when clusters are unimodal. Otherwise, the test may incorrectly detect a difference. This bias is evident and was therefore not reported in this study.

Finally, an intrinsic challenge in the evaluation of post-clustering inference methods lies in knowing the true partition. In our simulations, even if the true partition is known, in practice, it is difficult to determine whether the comparison of clusters are truly under the null hypothesis, complicating the evaluation of test performance.

These methods are still not suitable for application to scRNAseq data due to two main limitations: (1) their restrictive parametric assumptions and (2) the complexity of interpreting the results to ensure both the validity and statistical power of the tests in this application. This point is further discussed in Section 8.3.

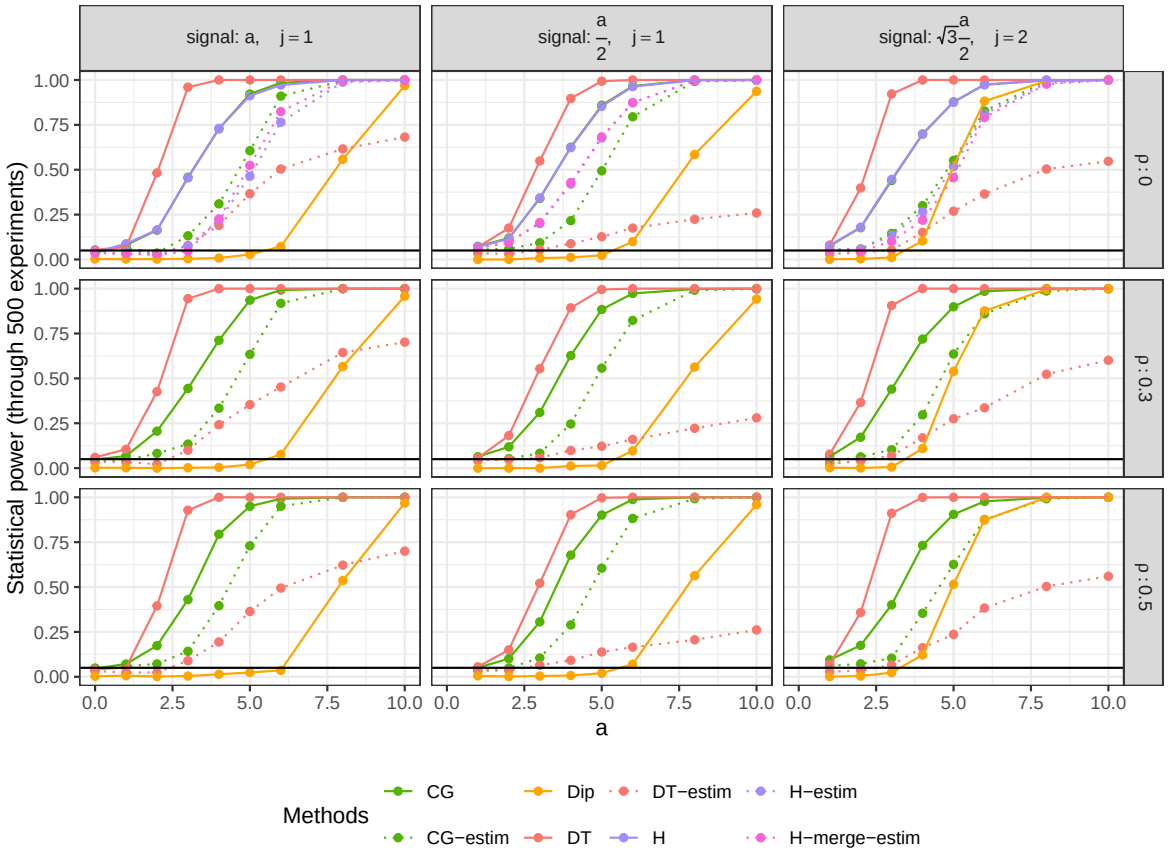

Figure 6.6 – Statistical power of marginal valid methods. Each panel corresponds to a combination of a marginal comparison and a value of ρ and methods. Methods of Hivert et al. (2024a) are only valid for $\rho = 0$.

Chapter 7

On the use of Gaussian mixtures in conditional approaches

In this chapter, we study the use of Gaussian Mixture Models (GMMs) for conditional post-clustering inference. Contrary to distance-based clustering algorithms such as HAC and K-means, GMMs provide a rich probabilistic model that can be exploited at the inference step. After a reminder on GMMs in Section 7.1, we present two ongoing attempts to extend the conditional approach of Gao et al. (2024) to the specific case of GMMs (Sections 7.2 and 7.3). Finally, we describe a preliminary numerical study to examine the impact of variance estimators provided by GMMs in Section 7.4. To streamline the analysis, we focus on the multivariate comparison of two clusters, leaving marginal comparisons for future exploration.

7.1 Gaussian Mixture Models (GMMs)

Gaussian Mixture Models (McLachlan, 2000; Bouveyron et al., 2019) are probabilistic models used to represent the underlying distribution of potentially complex data as a weighted combination of multiple multivariate normal distributions. These models are particularly well-suited for data that originate from multiple subpopulations, each of which can be modeled by a distinct Gaussian distribution. After estimating the unknown distribution of a sample by a Gaussian mixture distribution, a clustering of this sample can be deduced.

7.1.1 Clustering based on Gaussian Mixture Models

Let $\mathbf{X} = (X_1, ..., X_n)$ denote the considered sample. Clustering from mixture models is based on the idea that there exists an unknown class structure in our data that we want to discover. This sub-population structure is defined by latent variables $\mathbf{Z} = (Z_1, ..., Z_n)$ such that the probability that individual i belongs to Cluster $\mathcal{C}_k$ is $\mathbb{P}(Z_i = k) = \pi_k$, where $\pi_k \in (0, 1)$ and $\sum_{k=1}^{K} \pi_k = 1$. The individuals in the same sub-population are assumed to arise from the same probability distribution (here the same Gaussian distribution)

$$X_i \mid Z_i = k \sim \mathcal{N}(\mu_k, \Sigma_k). \tag{7.1}$$

Then, the distribution of X_i is a Gaussian mixture whose the density function is defined by

$$f(x|\theta) = \sum_{k=1}^{K} \pi_k \varphi(x \mid \mu_k, \Sigma_k), \tag{7.2}$$

where $\theta = (\pi_1, ..., \pi_K, \mu_1, ..., \mu_K, \Sigma_1, ..., \Sigma_K)$ is the parameter vector and $\varphi(\cdot|\mu, \Sigma)$ is the density function of the Gaussian law $\mathcal{N}_m(\mu, \Sigma)$.

In the context of Gaussian mixtures, 28 forms of mixtures are available. For each component $k \in [|K|]$, the decomposition of the covariance matrix is given by

$$\Sigma_k = L_k D_k^\top A_k D_k$$

where $L_k = |\Sigma_k|^{1/m}$, A_k is the diagonal matrix of the normalized eigenvalues of Σ_k, and D_k is the matrix of eigenvectors of Σ_k. Various constraints on these terms control the volume, shape, and orientation of the k-th component (Banfield and Raftery, 1993; Celeux and Govaert, 1995). By allowing the cluster proportions to be either variable or fixed ($\pi_k = 1/K$), a collection of 28 parsimonious and interpretable Gaussian mixture models can be derived. The `Rmixmod` package (Biernacki et al., 2006) implements these models, whose nomenclature we adopt. Among these models, let us highlight the pLI model, which specifies a common spherical covariance matrix, $\Sigma_k = \sigma^2 I_m$ for all $k \in [|K|]$, and fixed proportions $\pi_k = 1/K$. The pLC model assumes a general covariance matrix common to all components, $\Sigma_k = \Sigma$ for all $k \in [|K|]$, and fixed proportions $\pi_k = 1/K$. Finally, we consider the p_kLI and p_kLC models, which are the versions with free proportions π_k of pLI and pLC respectively.

For a fixed number of clusters K and a fixed shape, the estimation of the parameter vector θ is performed using a EM-type algorithm (see details in Section 7.1.2). Next, to select the best model, a model selection criterion as BIC (Schwarz, 1978) or ICL (Biernacki et al., 2000) is used. In this Chapter, this model selection step is not considered since the number of clusters K and the mixture shape are fixed. Finally, based on the estimator $\hat{\theta}$ of θ, the posterior probability of an individual i to belong to cluster k is estimated by

$$\hat{\tau}_{ik} = \mathbb{P}(Z_i = k \mid X_i, \hat{\theta}) = \frac{\hat{\pi}_k \varphi(X_i \mid \hat{\mu}_k, \hat{\Sigma}_k)}{\sum_{l=1}^{K} \hat{\pi}_l \varphi(X_i \mid \hat{\mu}_l, \hat{\Sigma}_l)}. \tag{7.3}$$

A clustering $\mathcal{C}$ is deduced through the Maximum A Posteriori (MAP) rule: the individual i is assigned to Cluster $C_{\hat{z}_i}$ with

$$\hat{z}_i = \underset{k \in [|K|]}{\arg\max}\ \hat{\tau}_{ik}. \tag{7.4}$$

7.1.2 EM-type algorithms

As previously mentioned, the estimation of the parameter vector θ is generally performed using the Expectation-Maximization (EM) algorithm (Dempster et al., 1977). This iterative algorithm is based on maximizing the expectation of the completed log-likelihood, conditional on $\mathbf{X}$ and a current value $\theta^{(t)}$:

$$\mathbb{E}\left[\ln f(\mathbf{X}, \mathbf{Z}|\theta)|\mathbf{X}, \theta^{(t)}\right], \tag{7.5}$$

where the completed likelihood is expressed as

$$f(\mathbf{X}, \mathbf{Z}|\theta) = \prod_{i=1}^{n} \prod_{k=1}^{K} \{\pi_k \varphi(X_i|\mu_k, \Sigma_k)\}^{\mathbb{1}_{\{Z_i=k\}}}.$$

After initializing the parameters $\theta^{(0)}$, the algorithm iterates between the following two steps. At iteration t, the two steps are:

- **Expectation Step (E-step)**: Compute the posterior probabilities $\tau_{ik}^{(t)}$ given the current parameters $\theta^{(t-1)}$ using the definition in Equation (7.3):

$$\tau_{ik}^{(t)} = \mathbb{P}(Z_i = k \mid X_i, \theta^{(t-1)}) \tag{7.6}$$

- **Maximization Step (M-step)**: Update the model parameters $\theta^{(t)}$ by maximizing the expected completed log-likelihood (see Equation (7.5)) with respect to the probabilities $\tau_{ik}^{(t)}$ computed in E-step. The proportions are updated as

$$\pi_k^{(t)} = \frac{1}{n} \sum_{i=1}^{n} \tau_{ik}^{(t)},$$

and the mean as

$$\mu_k^{(t)} = \frac{\sum_{i=1}^{n} \tau_{ik}^{(t)} X_i}{\sum_{i=1}^{n} \tau_{ik}^{(t)}}.$$

The covariance matrices are updated depending on the shape of the mixture (Celeux and Govaert, 1995). In the more general case $p_k L_k C_k$, the component covariance matrix is updated as

$$\Sigma_k^{(t)} = \frac{\sum_{i=1}^{n} \tau_{ik}^{(t)} \left(X_i - \mu_k^{(t)}\right) \left(X_i - \mu_k^{(t)}\right)^\top}{\sum_{i=1}^{n} \tau_{ik}^{(t)}}$$

Several extensions of the EM algorithm have been proposed for estimating GMM parameters. The Classification EM (CEM) algorithm (Celeux and Govaert, 1992) introduces an additional step between the E-step and M-step. This step consists of applying the MAP rule $\hat{z}_i^{(t)} = \underset{k \in [|K|]}{\arg\max}\, \tau_{ik}^{(t)}$, and using this value in the M-step instead of $\tau_{ik}^{(t)}$. The Stochastic EM (SEM) algorithm (Celeux, 1985) introduces a Stochastic Classification (S-step), which provides random assignments $\hat{z}_{ik}^{(t)}$ based on a random draw from a multinomial distribution with parameters $\tau_{ik}^{(t)}$. The M-step is then updated using $\hat{z}_{ik}^{(t)}$ instead of $\tau_{ik}^{(t)}$. We can also mention the SAEM algorithm (Delyon et al., 1999) and the MCEM algorithm (Wei and Tanner, 1990).

Note that the K-means procedure can be written as a mixture model with the pLI form by using the CEM algorithm to estimate the parameters.

7.2 Conditional test for GMM clustering

This section aims to explore how to use the test proposed by Gao et al. (2024) with a clustering based on Gaussian Mixture Model. Let $\mathcal{C}(\mathbf{X})$ denote the clustering obtained after applying the GMM procedure and $\theta^{(t)}$ be the estimation of the parameters vector θ at the t-th step of the EM algorithm. Under the same assumption as Gao et al. (2024), let $X_i \sim \mathcal{N}_m(\mu_i, \Sigma)$ for $i \in [|n|]$, where the covariance matrix Σ is assumed to be known. Assume the null hypothesis $\mathcal{H}_0 : \eta(C_k, C_{k'})^\top \boldsymbol{\mu} = 0_m$, where η is defined in Equation (5.3) and $\left\|\eta^\top \mathbf{X}\right\|_2$ is the associated test statistic. Gao et al. (2024) and González-Delgado et al. (2023) have described the conditional p-value associated with this test (see Equations (5.11) and (5.12) for the spherical case and (5.21) for the non-spherical case). In the context of a clustering based on GMM, Gao et al. (2024) proposes estimating the p-value using MC-Importance Sampling. The behavior of this test after a GMM clustering is illustrated for Setting 1 in Figures B.1 and B.3, using the parameterization described in Section 5.3, which involves a context with a spherical covariance matrix common to all individuals. The test controls the type I error rate conservatively (about 25% of p-values equal to 1 for $n = 500$) and the power of the test appears to be comparable to other conditional tests using HAC or K-means clusterings.

However, the use of MC-Importance Sampling to estimate the p-value requires to evaluate GMM clustering on each perturbed dataset, and the EM algorithm for estimating GMM parameters is computationally time-consuming. Figure 7.1 shows the computation time for the p-value as a function of the number of observations and variables. It is evident that execution time increases with the number of parameters. The data thinning procedure creates

two datasets $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$ from $\mathbf{X}$ where $X_i \sim \mathcal{N}(\mu_i, \Sigma)$. The GMM clustering is computed on $\mathbf{X}^{(1)}$ and the test of Wald is evaluated on $\mathbf{X}^{(2)}$. In comparison with the conditional approach, the data thinning method is faster.

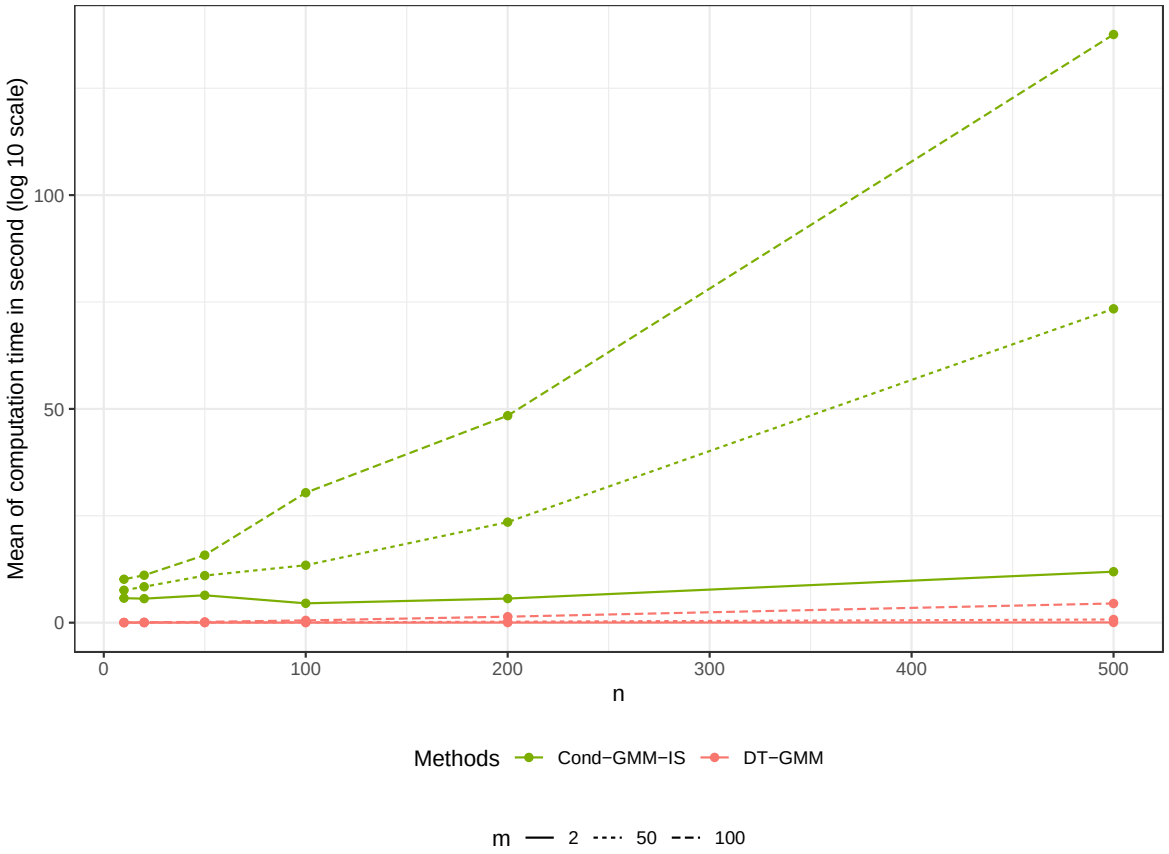

Figure 7.1 – Computation time of the conditional test by MC-Importance Sampling for GMM clustering using Setting 1 for simulations made in Section 5.3.

Thus, to save computational time, it would be beneficial to have an explicit calculation of the p-value in the case of GMM clustering. One idea is to take inspiration from Chen and Witten (2023) and exploit that the K-means algorithm is a CEM algorithm for the pLI shape of a GMM (see Section 7.1.2). Chen and Witten (2023) propose an explicit version of the conditional test for K-means by overconditioning the p-value. The set S_{kmeans} in Equation (5.18) can be rewritten as

$$\left\{ \phi \geq 0 : \bigcap_{t=0}^{T} \bigcap_{i=1}^{n} \left\{ \hat{z}_i^{(t)}(\tilde{\mathbf{x}}(\phi)) = \hat{z}_i^{(t)}(\mathbf{x}) \right\} \right\} \tag{7.7}$$

with $\hat{z}_i^{(t)}$ the estimated cluster to the individual i at Step t of the iterative clustering algorithm. In the CEM algorithm, it is possible to fix the clustering for all steps: at step $t \in [|T|]$ and for an individual $i \in [|n|]$,

$$\hat{z}_i^{(t)}(\tilde{\mathbf{x}}) = \hat{z}_i^{(t)}(\mathbf{x}) \iff \tau^{(t)}_{i\hat{z}_i^{(t)}(\mathbf{x})}(\tilde{\mathbf{x}}) > \tau_{il}^{(t)}(\tilde{\mathbf{x}}) \quad \forall l \neq \hat{z}_i^{(t)}(\mathbf{x}) \tag{7.8}$$

$$\iff \tilde{\pi}^{(t-1)}_{\hat{z}_i^{(t)}(\mathbf{x})} \varphi \left(\tilde{x}_i \mid \tilde{\mu}^{(t-1)}_{\hat{z}_i^{(t)}(\mathbf{x})}, \tilde{\Sigma}^{(t-1)}_{\hat{z}_i^{(t)}(\mathbf{x})} \right) > \tilde{\pi}_l^{(t-1)} \varphi \left(\tilde{x}_i \mid \tilde{\mu}_l^{(t-1)}, \tilde{\Sigma}_l^{(t-1)} \right) \quad \forall l \neq \hat{z}_i^{(t)}(\mathbf{x}) \tag{7.9}$$

where $\tilde{\mathbf{x}} := \tilde{\mathbf{x}}(\phi)$ and $\tilde{\pi}$, $\tilde{\mu}$ and $\tilde{\Sigma}$ denote the estimations of parameters based on $\tilde{\mathbf{x}}(\phi)$, to make reading easier. In particular, for the case pLI ($\Sigma_k = \sigma^2 I_m$ and $\pi_k = \frac{1}{K}$ for all $k \in [|K|]$), Equation (7.9) is equivalent to

$$\left\| \tilde{x}_i - \tilde{\mu}^{(\text{t-1})}_{\hat{z}_i^{(\text{t})}(\mathbf{x})} \right\|_2^2 > \left\| \tilde{x}_i - \tilde{\mu}_l^{(\text{t-1})} \right\|_2^2 \quad \forall l \neq \hat{z}_i^{(\text{t})}(\mathbf{x}). \tag{7.10}$$

Thus, in the CEM algorithm, the mean of the k-th component at step t is given by

$$\mu_k^{(\text{t})}(\mathbf{x}) = \frac{\sum_{i=1}^{n} \mathbb{1}_{\{\hat{z}_i^{(\text{t})}(\mathbf{x})=k\}} x_i}{n_k^{(t)}(\mathbf{x})} \quad \text{with} \quad n_k^{(\text{t})}(\mathbf{x}) = \sum_{i=1}^{n} \mathbb{1}_{\{\hat{z}_i^{(\text{t})}(\mathbf{x})=k\}}.$$

Equation (7.10) can then be rewritten as, $\forall l \neq \hat{z}_i^{(\text{t})}(\mathbf{x})$:

$$\left\| \tilde{x}_i - \sum_{i'=1}^{n} \left(\frac{\mathbb{1}_{\{\hat{z}_{i'}^{(\text{t-1})}(\tilde{\mathbf{x}})=\hat{z}_i^{(\text{t})}(\mathbf{x})\}}}{n^{(\text{t-1})}_{\hat{z}_i^{(\text{t})}(\mathbf{x})}(\tilde{\mathbf{x}})} \, \tilde{x}_{i'} \right) \right\|_2^2 > \left\| \tilde{x}_i - \sum_{i'=1}^{n} \left(\frac{\mathbb{1}_{\{\hat{z}_{i'}^{(\text{t-1})}(\tilde{\mathbf{x}})=l\}}}{n_l^{(\text{t-1})}(\tilde{\mathbf{x}})} \, \tilde{x}_{i'} \right) \right\|_2^2 \tag{7.11}$$

$$\implies \quad \left\| \tilde{x}_i(\phi) - \sum_{i'=1}^{n} \left(\frac{\mathbb{1}_{\{\hat{z}_{i'}^{(\text{t-1})}(\mathbf{x})=\hat{z}_i^{(\text{t})}(\mathbf{x})\}}}{n^{(\text{t-1})}_{\hat{z}_i^{(\text{t})}(\mathbf{x})}(\mathbf{x})} \, \tilde{x}_{i'}(\phi) \right) \right\|_2^2 > \left\| \tilde{x}_i(\phi) - \sum_{i'=1}^{n} \left(\frac{\mathbb{1}_{\{\hat{z}_{i'}^{(\text{t-1})}(\mathbf{x})=l\}}}{n_l^{(\text{t-1})}(\mathbf{x})} \, \tilde{x}_{i'}(\phi) \right) \right\|_2^2 \tag{7.12}$$

The transition from Equations (7.11) to (7.12) holds true as long as Equation (7.7) is satisfied with $\hat{z}_{i'}^{(\text{t-1})}(\mathbf{x}) = \hat{z}_{i'}^{(\text{t-1})}(\tilde{\mathbf{x}})$, and by extension $n_k^{(\text{t-1})}(\mathbf{x}) = n_k^{(\text{t-1})}(\tilde{\mathbf{x}})$ for all $k \in [|K|]$. Equation (7.12) depends on ϕ only through $\tilde{x}_i(\phi)$, allowing ϕ to be expressed as a solvable quadratic inequality since $\hat{z}_i^{(\text{t})}(\mathbf{x})$ is already known, for all $i \in [|n|]$ and all $t \in [|T|]$. This result is consistent with the findings of Chen and Witten (2023).

In the EM algorithm, consider the scenario where only the final clustering is maintained, $\mathcal{C}(\tilde{\mathbf{x}}(\phi)) = \mathcal{C}(\mathbf{x})$, corresponding to the following set, with the last step T of EM algorithm:

$$S_{EM} = \left\{ \phi \geq 0 : \bigcap_{i=1}^{n} \left\{ \hat{z}_i^{(\text{T})}(\tilde{\mathbf{x}}(\phi)) = \hat{z}_i^{(\text{T})}(\mathbf{x}) \right\} \right\}. \tag{7.13}$$

For an individual $i \in [|n|]$, Equations (7.8) to (7.9) hold for $t = T$. In particular, for the shape pLI, Equation (7.10) holds for $t = T$. In the EM algorithm,

$$\mu_k^{(\text{T})}(\mathbf{x}) = \sum_{i=1}^{n} \frac{\tau_{ik}^{(\text{T})}(\mathbf{x})}{\left(n_k^{[\tau]}(\mathbf{x}) \right)^{(\text{T})}} \, x_i \quad \text{with} \quad \left(n_k^{[\tau]}(\mathbf{x}) \right)^{(\text{T})} = \sum_{v=1}^{n} \tau_{vk}^{(\text{T})}(\mathbf{x}),$$

which gives

$$\left\| \tilde{x}_i - \sum_{i'=1}^{n} \frac{\tau^{(\text{T-1})}_{i' \hat{z}_i^{(\text{T})}(\mathbf{x})}(\tilde{\mathbf{x}})}{\left(n^{[\tau]}_{\hat{z}_i^{(\text{T})}(\mathbf{x})}(\tilde{\mathbf{x}}) \right)^{(\text{T-1})}} \, \tilde{x}_{i'} \right\|_2 > \left\| \tilde{x}_i - \sum_{i'=1}^{n} \frac{\tau_{i'l}^{(\text{T-1})}(\tilde{\mathbf{x}})}{\left(n_l^{[\tau]}(\tilde{\mathbf{x}}) \right)^{(\text{T-1})}} \, \tilde{x}_{i'} \right\|_2 \quad \forall l \neq \hat{z}_i^{(\text{T})}(\mathbf{x}). \tag{7.14}$$

In this case, S_{EM} does not allow expressing the inequality in terms of ϕ as seen in Equation (7.12). To achieve this, one would need to fix the $\tau_{ik}^{(\text{t})}(\tilde{\mathbf{x}}(\phi))$ at each step $t \in [|T|]$ as follows:

$$S'_{EM} = \left\{ \phi \geq 0 : \bigcap_{i=1}^{n} \left\{ \hat{z}_i^{(\text{T})}(\tilde{\mathbf{x}}(\phi)) = \hat{z}_i^{(\text{T})}(\mathbf{x}) \right\} \cap \bigcap_{t=1}^{T} \bigcap_{k=1}^{K} \bigcap_{i=1}^{n} \left\{ \tau_{ik}^{(\text{t})}(\tilde{\mathbf{x}}(\phi)) = \tau_{ik}^{(\text{t})}(\mathbf{x}) \right\} \right\}. \tag{7.15}$$

Nevertheless, as the event $\tau_{ik}^{(t)}(\tilde{\mathbf{x}}(\phi)) = \tau_{ik}^{(t)}(\mathbf{x})$ is only possible if $\mathbf{x} = \tilde{\mathbf{x}}(\phi)$, the set S'_{EM} is a singleton such that $S'_{EM} = \{\phi = \left\|\eta^\top \mathbf{x}\right\|_2\}$. Then, for the EM algorithm, the over-conditioning is too restrictive to obtain an explicit p-value.

Here, we have only considered the pLI shape and the EM algorithm. Another possible extension is to explore the CEM with other shape of covariance matrix. As the estimators $\pi_k^{(t)}$, $\mu_k^{(t)}$ and $\Sigma_k^{(t)}$ do not depend on τ_{ik}, an explicit p-value could be obtained without fixing τ_{ik} in the condition. This needs more investigation.

In conclusion, the conditional method of Gao et al. (2024) can be used in the GMM context with the p-value estimated by MC-Importance Sampling, although the computational time can be high. An explicit version of this test for GMM with the EM algorithm does not seem possible. However, based on K-means results, obtaining an explicit p-value could be possible using the CEM algorithm in cases other than pLI, but with an additional conditioning event that could affect the statistical power.

7.3 Use the posterior probabilities in contrast?

Let us consider the case of Gaussian data analysis where $X_i \sim \mathcal{N}(\mu_i, \Sigma)$ with $\Sigma = \sigma^2 I_m$ for simplification, though the following reasoning applies for a general covariance matrix Σ. The test proposed by Gao et al. (2024) is based on the null hypothesis $\mathcal{H}_0^{\mathcal{C}} : \eta(C_k, C_{k'})^\top \boldsymbol{\mu} = 0_m$, where η is defined in Equation (5.3) and recalled here:

$$\eta_i(C_k, C_{k'}) = \frac{\mathbb{1}_{i \in C_k}}{|C_k|} - \frac{\mathbb{1}_{i \in C_{k'}}}{|C_{k'}|}, \ \forall\ i \in [|n|]. \tag{7.16}$$

Since the estimator $\eta(C_k, C_{k'})^\top \mathbf{X}$ follows a Gaussian distribution $\mathcal{N}(0, \sigma^2 \left\|\eta(C_k, C_{k'})\right\|_2^2)$ under the null hypothesis, the test statistic $\left\|\eta(C_k, C_{k'})^\top \mathbf{X}\right\|_2$ is distributed as $\sigma^2 \left\|\eta\right\|_2 \ \chi_m$. These two quantities are defined under the null hypothesis provided that η is fixed (i.e., known before the data are observed).

In our case, the clustering $\mathcal{C}$ is derived from the GMMs via the MAP rule $\mathcal{C}(\mathbf{X}) := \text{MAP}(\tau(\mathbf{X}))$. The contrast vector is random, preventing us from knowing the distribution of the test statistic $\left\|\eta(C_k(\mathbf{X}), C_{k'}(\mathbf{X}))^\top \mathbf{X}\right\|_2$. Recall that Gao et al. (2024) conditions the p-value on the event $\{C_k(\mathbf{x}), C_{k'}(\mathbf{x}) \in \mathcal{C}(\mathbf{X})\}$, which allows fixing the contrast vector. The distribution of the conditional test statistic is then accessible by decomposing $\mathbf{X}$ to reveal the test statistic (see Equation (5.10)) and fixing the other parts of $\mathbf{X}$ (see the overconditioning in Equation (5.11)). Theorem 1 demonstrates how the distribution of the conditional test statistic is determined.

In the estimation of the GMM parameters via an EM algorithm, the mean of Cluster C_k is estimated by

$$\hat{\mu}_k = \sum_{i=1}^{n} \frac{\tau_{ik}}{n_k^{[\tau]}} X_i \text{ with } n_k^{[\tau]} = \sum_{v=1}^{n} \tau_{vk}.$$

It takes into account all individuals (including those not in C_k) by weighting them according to τ_{ik}, their posterior probability of belonging to Cluster C_k. This raises the question of how to use these mean estimators to construct a post-clustering test. Let $\left\|\xi(\tau)^\top \mathbf{X}\right\|_2$ be a test statistic associated with the means $\hat{\mu}_k$ and $\hat{\mu}_{k'}$, defined by

$$\xi_i(\tau) = \frac{\tau_{ik}}{n_k^{[\tau]}} - \frac{\tau_{ik'}}{n_{k'}^{[\tau]}}, \ \forall i \in [|n|]. \tag{7.17}$$

First, let us examine how $\left\|\xi(\tau)^\top \mathbf{X}\right\|_2$ behaves. Under $\mathcal{H}_0^{\mathcal{C}}$, $\xi(\tau)^\top \mathbf{X} \sim \mathcal{N}(\xi(\tau)^\top \boldsymbol{\mu}, \sigma^2 \left\|\xi(\tau)\right\|_2^2)$. This distribution is not known as long as $\xi(\tau)^\top \boldsymbol{\mu}$ is unknown, as there is no direct link between $\xi(\tau)^\top \boldsymbol{\mu}$ and $\eta(C_k, C_{k'})^\top \boldsymbol{\mu}$ (which is equal to zero under $\mathcal{H}_0^{\mathcal{C}}$). Thus the distribution of the test statistic $\left\|\xi(\tau)^\top \mathbf{X}\right\|_2$ is not known under $\mathcal{H}_0^{\mathcal{C}}$.

To circumvent this problem, we can consider a modified null hypothesis: $\mathcal{H}_0^\tau : \xi(\tau)^\top \boldsymbol{\mu} = 0_m$. This allows us to have

$$\xi(\tau)^\top \mathbf{X} \overset{\mathcal{H}_0^\tau}{\sim} \mathcal{N}(0, \sigma^2 \left\|\xi(\tau)\right\|_2^2)$$

and thus $\left\|\xi(\tau)^\top \mathbf{X}\right\|_2 \overset{\mathcal{H}_0^\tau}{\sim} (\sigma \left\|\xi(\tau)\right\|_2)\chi_m$. However, the interpretation of $\mathcal{H}_0^\tau$ is not clear.

Let us move beyond the interpretation of the null hypothesis. In practice, the posterior probabilities of membership are estimated from the data $\tau := \tau(\mathbf{X})$. Thus, the contrast vector $\xi(\tau(\mathbf{X}))$ is random. Consider the conditioning proposed by Gao et al. (2024), which requires that the compared clusters be fixed, i.e., $\{C_k(\mathbf{x}), C_{k'}(\mathbf{x}) \in \mathcal{C}(\mathbf{X})\}$. The p-value then becomes:

$$\mathbb{P}_{\mathcal{H}_0^\tau}\Big(\left\|\xi(\tau(\mathbf{X}))^\top \mathbf{X}\right\|_2 \geq \left\|\xi(\tau(\mathbf{X}))^\top \mathbf{x}\right\|_2 \;\Big|\; C_k(\mathbf{x}), C_{k'}(\mathbf{x}) \in \mathcal{C}(\mathbf{X})\Big) \tag{7.18}$$

$$\neq \mathbb{P}_{\mathcal{H}_0^\tau}\Big(\left\|\xi(\tau(\mathbf{x}))^\top \mathbf{X}\right\|_2 \geq \left\|\xi(\tau(\mathbf{x}))^\top \mathbf{x}\right\|_2 \;\Big|\; C_k(\mathbf{x}), C_{k'}(\mathbf{x}) \in \mathcal{C}(\mathbf{X})\Big). \tag{7.19}$$

The conditioning proposed by Gao et al. (2024) does not fix the contrast vector $\xi(\tau(\mathbf{X}))$, as specified by the transition from (7.18) to (7.19). The conditioning $\{C_k(\mathbf{x}), C_{k'}(\mathbf{x}) \in \mathcal{C}(\mathbf{X})\}$ only provides information about the ordering of some of the τ_{ik}, specifically $\tau_{ik} > \tau_{il}$ $\forall l \neq k$, $\forall i \in C_k$ and $\tau_{ik'} > \tau_{il}$ $\forall l \neq k'$, $\forall i \in C_{k'}$. The contrast vector can only be fixed by providing at least as much information as present in the vector itself for conditioning. Therefore, conditioning on the larger event $\{\tau(\mathbf{x}) = \tau(\mathbf{X})\}$ allows us to fix the contrast vector as follows:

$$\mathbb{P}_{\mathcal{H}_0^\tau}\Big(\left\|\xi(\tau(\mathbf{X}))^\top \mathbf{X}\right\|_2 \geq \left\|\xi(\tau(\mathbf{X}))^\top \mathbf{x}\right\|_2 \;\Big|\; \tau(\mathbf{x}) = \tau(\mathbf{X})\Big) \tag{7.20}$$

$$= \mathbb{P}_{\mathcal{H}_0^\tau}\Big(\left\|\xi(\tau(\mathbf{x}))^\top \mathbf{X}\right\|_2 \geq \left\|\xi(\tau(\mathbf{x}))^\top \mathbf{x}\right\|_2 \;\Big|\; \tau(\mathbf{x}) = \tau(\mathbf{X})\Big). \tag{7.21}$$

In this case, the p-value can be expressed using the decomposition of $\mathbf{X}$ as described by Gao et al. (2024) and by using ξ instead of η. However, the conditioning $\{\tau(\mathbf{x}) = \tau(\mathbf{X})\}$ is too restrictive, leading to conditioning by a singleton, as described in Section 7.2.

In conclusion, utilizing posterior probabilities from the GMM clustering does not appear to yield an interesting test due to the difficulty in interpreting the null hypothesis for exploiting the test statistic and the restrictive conditioning required, which makes the p-value non-informative. We can notice that to address issues in post-clustering inference using conditional approaches, the conditioning needs to incorporate at least as much information as is present in the contrast vector.

7.4 Variance estimation from GMM clustering in conditional tests

In the GMM clustering process, the EM algorithm provides estimators for both means $\hat{\mu}_k$ and covariance matrices $\hat{\Sigma}_k$. Here, we assume a setting where the covariance matrix is common to all individuals ($\Sigma_k = \Sigma$ $\forall k \in [|K|]$), which correspond to the framework required to apply the conditional test developed by Gao et al. (2024). In this context, a key question is whether the GMM covariance estimators are suitable for use in conditional inference tests, particularly in controlling type I error rate and statistical powers.

Theoretical works in Gao et al. (2024) and González-Delgado et al. (2023) show that overestimating covariance matrix can ensure valid type I error control in conditional tests.

This robustness to overestimation has been confirmed in numerical experiments, as discussed in Section 5.4. Conversely, underestimating the covariance, as seen with the intra-cluster estimator, leads to test that fail to control type I error rate effectively.

The estimator given by the GMM clustering process provides a natural candidate for covariance estimation in conditional tests since they leverage information from the clustering procedure taking into account all individuals weighted by the τ_{ik}. This estimator should be better than the intra-cluster estimator, which underestimates the variance under the null hypothesis due to artificial cluster separation and than the global estimator which overestimates the variance under the alternative hypothesis. The goal of this section is to clarify whether this GMM estimator strikes the right balance between avoiding underestimation while not overly inflating variance, thereby preserving test validity while retaining statistical power.

7.4.1 Simulation setting

Consider Setting 1 (see Section 5.3.1.1) with $n = 500$, $m = 2$, and $\Sigma = I_m$, simulating two groups. The statistical performance is evaluated for $a \in \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$, with 500 simulations for each value of a. The clustering is obtained by a Gaussian mixture model with the pLI shape (see Section 7.1.1) and $K = 2$. The variance estimator in this case is given by

$$\hat{\sigma}^2_{\text{GMM}} = \frac{1}{nm} \sum_{i=1}^{n} \sum_{k=1}^{K} \hat{\tau}_{ik} \, \|x_i - \hat{\mu}_k\|_2^2 \quad \text{with } \hat{\mu}_k = \sum_{i=1}^{n} \frac{\hat{\tau}_{ik}}{n_{k'}^{[\tau]}} \, x_i. \tag{7.22}$$

The inference method used is the conditional approach for a multivariate test (Gao et al., 2024), comparing Clusters C_1 and C_2 in Setting 1. The p-value is estimated via MC-Importance Sampling (see Equation (5.20)) with $Q = 1000$ draws, as there is no explicit characterization of the set S in this case (see Section 7.2). The method is applied using the variance estimator from Equation (7.22). It is compared to other plug-in estimators, including the oracle $\sigma^2 = 1$, the global (all) estimator is

$$\hat{\sigma}^2_{\text{all}} = \frac{1}{m(n-1)} \sum_{i=1}^{n} \|x_{ij} - \bar{\mathbf{x}}\|_2^2$$

and the intra-cluster (intra) estimator is

$$\hat{\sigma}^2_{\text{intra}} = \sum_{k=1}^{K} \frac{|C_k|}{nm} \sum_{i \in C_k} \|x_i - \bar{\mathbf{x}}_k\|_2^2 .$$

7.4.2 Statistical performance

The characterization of the type I error rate is conducted under the null hypothesis ($a = 0$). The ecdfs of the p-values are shown in Figure 7.2. As previously noted in Sections 5.3 and 5.4, the method with the known covariance matrix and the estimator 'all' control the type I error rate, while the estimator 'intra' does not control the type I error rate. The estimator given by the clustering based on GMM does not control the type I error rate as the obtained p-values are stochasticly smaller than the uniform. Note that the excess of p-value at 1 is explained in Section 5.3.

The results agree with the variance estimates obtained, as shown in Figure 7.3 where the average estimated values of σ^2 is a function of the signal a. Under $\mathcal{H}_0$, the global estimator provides a good variance estimate, while both the GMM and intra-cluster estimators underestimate the variance, which is consistent with previous observations. As expected, the GMM

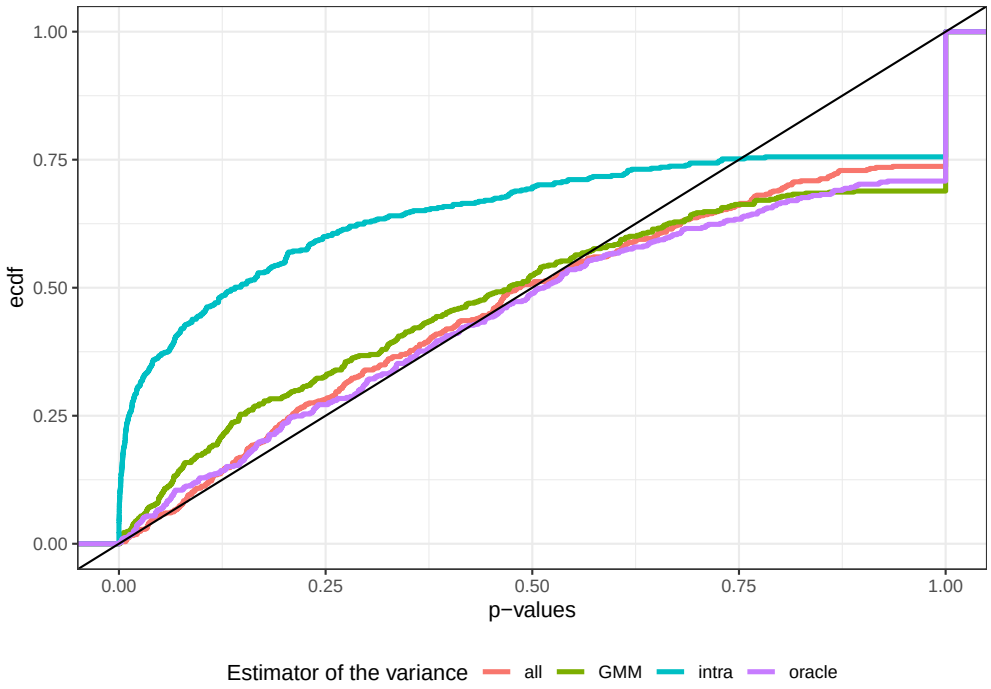

Figure 7.2 – ECDF of p-values under $\mathcal{H}_0$ using Setting 1 with $m = 2$, $n = 500$, $\Sigma = I_m$. The test is evaluated with the conditional approach of Gao et al. (2024) using the GMM clustering (with pLI assumption) and the p-values are estimated by MC-Importance Sampling with $Q = 1000$ draws (see Equation (5.20)). The test is computed with the known covariance matrix $\Sigma = I_m$ (oracle), and the plug-in estimators $\hat{\sigma}^2_{\text{all}}\, I_m$ (all), $\hat{\sigma}^2_{\text{intra}}\, I_m$ (intra) and $\hat{\sigma}^2_{\text{GMM}}\, I_m$ (GMM).

estimator underestimates variance less markedly than the intra-cluster estimator. When signal is present in the data, the global estimator tends to overestimate the variance, while the intra-cluster estimator underestimates it for $a \leq 5$. However, the GMM estimator provides an accurate variance estimate as soon as there is sufficient signal. Therefore, it is particularly interesting to examine the statistical power of the test using this estimator.

The statistical power is shown in Figure 7.4. The power of the test using the GMM estimator is comparable to that of the test using the oracle. This observation has to be tempered by the fact that the test using the GMM estimator is slightly anti-conservative (Figure 7.3). In contrast, the global estimator overestimates the variance, making the test overly conservative; it only becomes powerful when the signal between clusters is sufficiently strong. Finally, the power of the test using the intra-cluster estimator is not interpretable since this method is strongly anti-conservative.

In conclusion, the variance estimator provided by the GMM clustering process offers accurate estimates once there is some signal present, enabling a test that is as powerful as one using the true variance, at the price of a moderate anti-conservativeness under the null hypothesis. This suggests that the GMM estimator is an interesting candidate when the variance is unknown.

Note that the assumption of equal proportions π_k can be too restrictive in practice. We have explored the estimation of a Gaussian mixture model assuming $p_k LI$ shape (the π_k are free) while, in the simulation, the true proportion $\pi_k = 0.5$. In this case, we have observed that the method of Gao et al. (2024) does not control the type I error rate: indeed, it provides an excess of small p-values even in the case of known variance (see Figure B.10 in Appendix

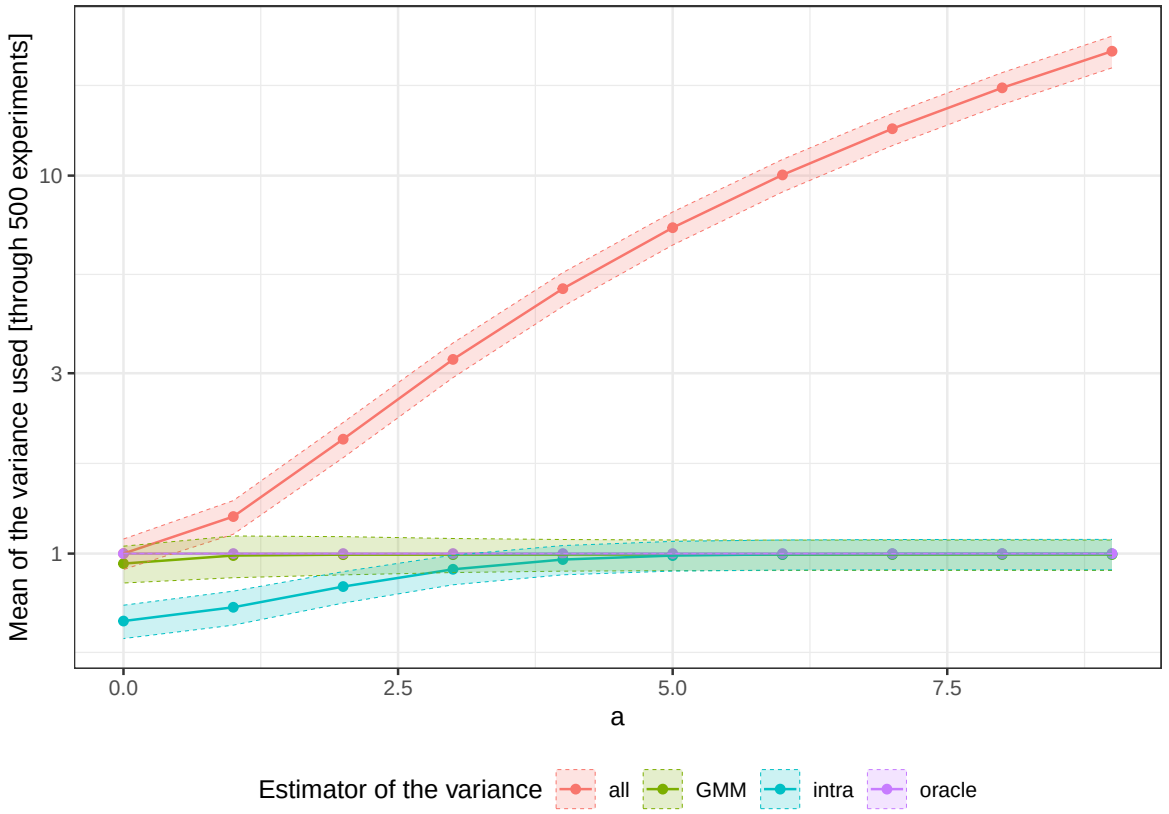

Figure 7.3 – Evolution of the estimation of σ in function of the signal. The value 'oracle' is the true value of σ and set to 1 for all experiments for all a. The variance is overestimated by the estimator 'all'. The estimator 'intra' under-estimates the variance for $a \leq 4$. The 'GMM' (in pLI assumption) estimator seems to be a good estimator since it underestimates the variance for $a \leq 1$.

B.3.4).

7.4.3 Non-spherical covariance matrix

This section extends the previous analysis to a non-spherical covariance context. We use the same simulation setting, adapting the covariance matrix to $\Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$ with $\rho \in \{0, 0.3, 0.5\}$. In this scenario, the GMM clustering is estimated with a pLC shape and $K = 2$. The covariance matrix estimator under this model is given by:

$$\hat{\Sigma}_{\text{GMM}} = \frac{1}{n} \sum_{k=1}^{K} \sum_{i=1}^{n} \tau_{ik} (x_i - \hat{\mu}_k)^\top (x_i - \hat{\mu}_k). \tag{7.23}$$

We compare this GMM estimator to several other estimators:
— the oracle covariance matrix Σ,
— the global estimator $\hat{\Sigma}_{\text{all}} = \frac{1}{n-1} (\mathbf{x} - \bar{\mathbf{x}})^\top (\mathbf{x} - \bar{\mathbf{x}})$,
— the intra-cluster estimator $\hat{\Sigma}_{\text{intra}} = \sum_{k=1}^{K} \frac{|C_k|}{n} (\mathbf{x}_{C_k} - \bar{\mathbf{x}}_{C_k})^\top (\mathbf{x}_{C_k} - \bar{\mathbf{x}}_{C_k})$.

Results are detailed in Appendix B.3.4. The findings align with those from the spherical case. Specifically, the GMM estimator performs comparably to the oracle in terms of statistical power, but it does not adequately control the type I error rate.

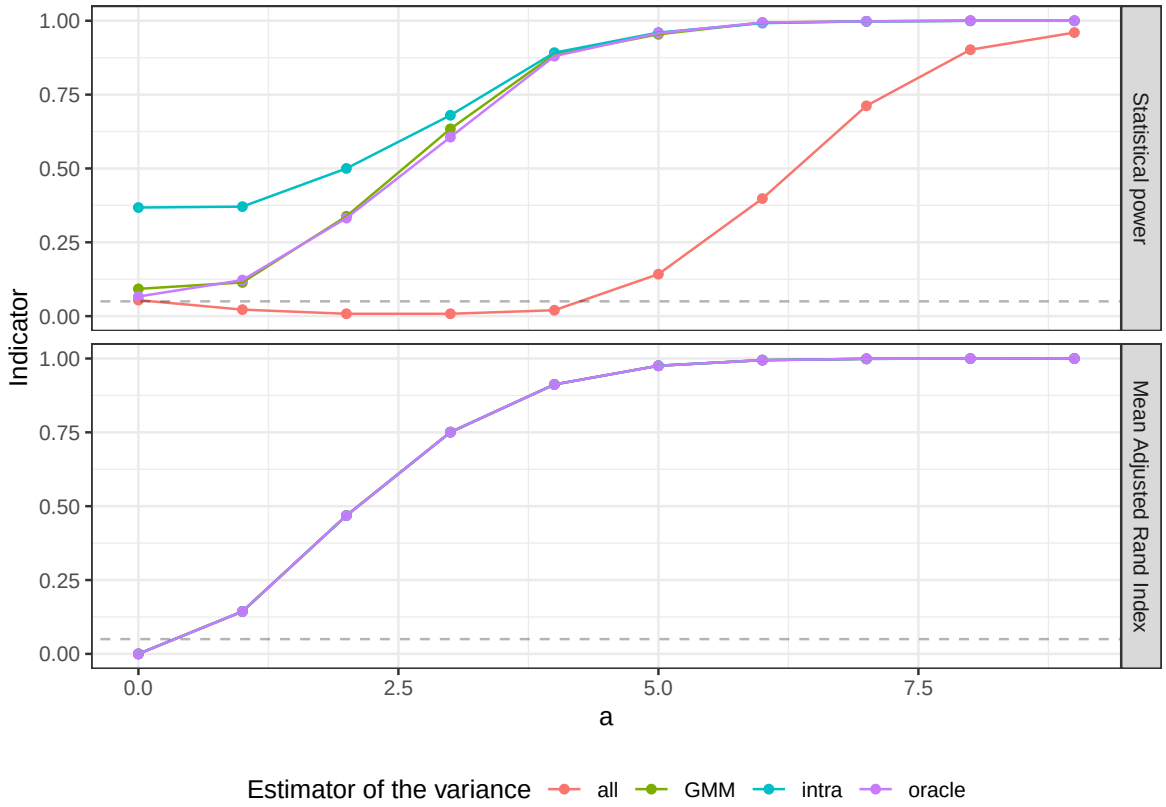

Figure 7.4 – Estimation of the statistical power and the Adjusted Rand Index in function of the signal a using Setting 1 with $m = 2$, $n = 500$, $\Sigma = I_m$. The test is evaluated with the conditional approach of Gao et al. (2024) using the GMM clustering (with pLI assumption) and the p-values are estimated by MC-Importance Sampling with Q =1000 draws (see Equation (5.20)). The test is computed with the known covariance matrix $\Sigma = I_m$ (oracle), and the plug-in estimators $\hat{\sigma}^2_{\text{all}} I_m$ (all), $\hat{\sigma}^2_{\text{intra}} I_m$ (intra) and $\hat{\sigma}^2_{\text{GMM}} I_m$ (GMM).

7.5 Conclusion

The estimation of a mixture model not only provides a clustering of the sample but also estimates the parameters of the cluster distributions. This chapter investigates the use of these parameters to adapt the conditional procedure proposed by Gao et al. (2024). Specifically, we explore the feasibility of obtaining an exact p-value when clustering is obtained via the MAP rule after an EM algorithm, how to incorporate the mean estimator into the test statistic, and whether the covariance matrix estimator could be used to enhance statistical performance.

The EM algorithm iteratively estimates mixture parameters based on posterior probabilities. However, the over-conditioning associated with these parameters renders the p-value uninformative, preventing the derivation of an exact p-value. A similar issue arises when incorporating the mean density estimators into the test statistic. Finally, plugging in the covariance matrix estimator yields a powerful test under $\mathcal{H}_1$ (comparable to the oracle), at the price of a moderate anti-conservativeness under $\mathcal{H}_0$.

Thus, adapting the Gao et al. (2024) procedure to mixture models appears challenging due to these challenges. However, leveraging the CEM algorithm for GMM estimation seems to be an interesting perspective. This approach could potentially yield an exact p-value, though with reduced power due to over-conditioning, as highlighted for the K-means exact

conditional p-value (Chen and Witten, 2023).

Chapter 8

Conclusions and perspectives

8.1 Conclusions

Performing clustering estimation followed by hypothesis testing on the same dataset introduces a *double dipping* issue. A number of post-clustering inference methods based on information partition or conditional inference have have been developed in parallel during the course of this thesis project to address this problem. Chapters 5 and 6 provide a review and comparison of existing methods using simulated Gaussian data, focusing respectively on the multivariate and marginal comparison of two-cluster means. Chapter 7 describes our attempts to exploit the Gaussian mixture models according to several points of view (clustering and estimators of parameters) for the conditional approach.

Our simulation shows that the data thinning method is more powerful than conditional approaches if the entire data distribution is known. As demonstrated numerically in Section 5.3.3, the thinning parameter ε, corresponding to the proportion of information containing in $\mathbf{X}^{(1)}$, has an impact on both clustering and statistical performance of the test. Although we expected the calibration of ε to be the principal issue in applying this method, knowledge of the data distribution (and the distribution parameters) is more problematic in practice. In our simulation, we see that the covariance matrix estimation decreases the statistical power of the data thinning procedure. Indeed, the misspecification of the data distribution (including parameters) can produce invalid tests or drastically reduce the statistical power.

Conditional approaches have been developed for Gaussian distribution. Some extensions give guarantees beyond the case of a fully known covariance matrix. In the spherical case, incorporating the unknown variance into the test statistic (Yun and Foygel Barber, 2023) provides a more powerful test than plugging a variance estimator into the test assuming a known variance. The extension assuming an unknown variance is only valid in the spherical case, and we will discuss how extending this method to the non-spherical case in Section 8.2.1. We want to outline the following point for this category of method: to gain precision and computation time, obtaining an exact p-value can lead to over-conditioning, resulting in a loss of statistical power. In the K-means method, using the exact p-value does not consistently reduce computation time and decreases statistical power. Thus, in the context of conditional approaches to post-clustering inference, systematically seeking an exact p-value is not necessary.

As part of the numerical evaluation of these methods, we have observed that clustering estimation can bring several difficulties. Cluster label switching can make it challenging to identify which comparisons to study. In addition, inherent to the issue of post-clustering inference, clustering estimation also impacts the evaluation of comparisons between methods, leading to potential bias in interpretations. To control the comparisons processed, Gao et al. (2024) only considers experiments where true clustering is found, while we jointly evaluate statistical performance with the indications of clustering performance using the Adjusted Rand Index. Both solutions fail to resolve these issues perfectly. In future evaluations of post-clustering inference methods, it is essential to take these issues into account and to take the necessary precautions regarding interpretation.

8.2 Discussion on assumptions of the covariance matrix

Post-clustering inference methods mostly assume that the covariance matrix is common to all individuals and known. This modeling assumption is rarely realistic. In this section, we examine what would happen if the covariance matrix were common but unknown and non-spherical. Next, we turn to the case where the covariance matrix is not common to all observations.

8.2.1 Unknown general Σ, common to all individuals

The method proposed by Gao et al. (2024) and Yun and Foygel Barber (2023) assume a common variance across all individuals. This assumption allows for the straightforward application of Cochran's theorem to establish the independence of Gaussian quantities, which holds under spherical conditions. Our numerical results reveal that the approach of Gao et al. (2024) with an estimator of the variance reduces the statistical power more than the Yun and Foygel Barber (2023)'s method (where the variance is unknown and not directly estimated). However, the approach of Yun and Foygel Barber (2023) is only valid under spherical assumptions, as seen in simulations in Section 5.4.

Extending the results of Yun and Foygel Barber (2023) to handle general covariance structures is an important question. In the case where the covariance matrix is both non-spherical and unknown, the existing techniques of Gao et al. (2024) and González-Delgado et al. (2023), which rely on the knowledge of Σ to use Cochran's theorem, become inapplicable. A theoretical tool analogous to Cochran's theorem for non-spherical Gaussian data is still needed to express the required independence to extend the method of Yun and Foygel Barber (2023) to a general covariance matrix.

8.2.2 Covariance matrix common to individuals within the same cluster

The current modeling in post-clustering inference assumes that the covariance matrix is common to all individuals. A more realistic modeling would consider that individuals within the same cluster share the same covariance matrix. This is one of the assumptions of mixture models. However, cluster membership is a latent variable that needs to be estimated in order to assign a covariance matrix to each individual. This estimation creates another form of circular analysis.

Hivert et al. (2024b) have already highlighted this issue for data thinning, and have demonstrated that this leads to inflated false positive rates. To address this, they propose a model where each individual has its own covariance matrix Σ_i. However, in practice, these matrices are often unknown, and the results are highly sensitive to how well Σ_i is estimated, which can lead to inflated FDR. A perspective to overcome this issue could be to use tools from Bayesian statistics.

In contrast to data thinning, conditional approaches can estimate clustering without requiring Σ and then use the estimation of the latent variable to assign a covariance matrix to each cluster. However, Theorem 1, which assumes a common covariance matrix across all individuals, would need to be adapted accordingly. Here again, the latent structure of clustering leads us to explore the Bayesian approach as a potential solution to this problem.

8.3 Post-clustering inference methods for scRNAseq data

The identification of marker genes in scRNAseq data analysis was the initial motivation for our study of post-clustering inference. Post-clustering inference methods are currently constrained by their parametric assumptions. Additionally, conditional approaches are tailored to

specific problems, such as comparing two clusters (multivariate or marginal). These methods remain far from addressing the issue of identifying marker genes. This section outlines the necessary steps and the key challenges identified for applying these methods in practice.

8.3.1 Adaptivity to non-Gaussian distributions

The modeling of scRNAseq data is a particularly challenging task. The widely used negative binomial modeling often encounters difficulties for estimating the dispersion parameter. To circumvent these parametric challenges, non-parametric tests are frequently used, such as the Wilcoxon test implemented in `Seurat` or the Kolmogorov-Smirnov test based on empirical distribution functions.

Conditional approaches are tied to Gaussian data, since Gaussianity is used to decompose $\mathbf{X}$ and obtain the distribution of the conditional p-value. On the other hand, data thinning is parametric in essence, as it requires the data distribution to be known. Therefore, closing the gap between existing methods and possible applications to scRNAseq data seems to be a particularly challenging task.

8.3.2 Adapting comparisons to identify marker genes

The marginal comparison of clusters discussed in Chapter 6 is motivatied in analyzing marker genes for scRNAseq data by identifying which variable carries a difference between two clusters. However, in practice, the identification of marker genes for a given cluster C_k involves testing this cluster against all others, which are grouped in $\bar{C}_k := \bigcup_{l \in [|K|] \setminus k} C_l$ (as implemented e.g. in `Seurat`). The data thinning procedure is independent of the clustering and statistic tests. Thus, this procedure can be applied to compare one cluster against all others (and more generally to all contrasts). In the conditional approach, the two-cluster comparison test can be used by considering $\bar{C}_k$ as a cluster. The test statistic is $\left\|\eta(C_k, \bar{C}_k)^\top \mathbf{X}\right\|_2$ and the conditioning becomes $\{C_k, \bar{C}_k \in \mathcal{C}(\mathbf{X})\}$, which can be simplified as $\{C_k \in \mathcal{C}(\mathbf{X})\}$ for this specific question. Thus, adapting the marginal conditioning methods seems relatively straightforward.

Theorem 1 is a generalization of the main result of Gao et al. (2024) to any contrast comparing two clusters. Extending this result to an arbitrary contrast $\eta(\mathcal{C}(\mathbf{X}))$ would allow to handle not only the above case where one cluster is compared to all others, but also the comparison of two groups of clusters. The latter could be interesting in marker gene identification when it is assumed that cell clusters have a biological hierarchical structure.

Nevertheless, even if the generalization of Theorem 1 would provide statistical guarantees for any contrast based on $\mathcal{C}(\mathbf{X})$ clustering, the question of calculating this p-value remains open. Exact p-values are based on the properties of specific clustering methods, and sometimes require over-conditioning. So, the first question is whether these developments are still valid or how to adapt them to the general problem. The MC-Importance Sampling approach seems more straightforward to adapt, as the conditioning is directly used without modification. For example, when comparing one cluster against all others, the specific conditioning $\{C_k \in \mathcal{C}(\mathbf{X})\}$ seems directly adaptable.

In conclusion, methodological developments are needed to adapt conditional approaches to the comparison induced by the identification of marker genes. Initially, we want to extend the conditional test to compare one cluster against all the others. Then, in a second phase, we envisage further developments to extend the method to any contrast. These perspectives have been here discussed within the marginal framework motivated by applications, but naturally extends to the multivariate framework.

8.4 Statistical guarantees on clustering with post-clustering inference methods

The multivariate tests discussed in Chapter 5 assess whether two clusters are truly distinct. While we can currently obtain $K(K-1)/2$ p-values for all possible pairwise comparisons between clusters, no practical use of these tests has been proposed. This section opens a discussion on the use of these tests.

A common challenge in clustering is to determine the number of clusters K, which is generally unknown in practice (Haslbeck and Wulff, 2020; Estivill-Castro, 2002). To address this, we can suggest the following procedure: starting from the top of a dendrogram obtained via hierarchical clustering, the procedure sequentially tests cluster merges to decide which pairs of clusters should remain distinct. If the null hypothesis is rejected, the clusters are considered separate, and the procedure continues to the next pair. If the null hypothesis is accepted, the clusters remain merged, and no further comparisons are made along that branch. The process repeats until all branches are evaluated.

Although this approach helps to select K, it involves the computational cost of constructing a dendrogram and repeating the tests. Importantly, this procedure would ensure that the resulting clusters are statistically guaranteed to be distinct, offering a more precise partitioning of the dendrogram than a horizontal cut in a standard HAC procedure. However, multiple testing corrections are necessary since they involve a series of tests. Yekutieli (2008) provides a multiple testing correction method (controlling the FDR) that fits this problem, where a sequence of tests (with an unknown number of tests a priori) is performed, each test dependent on the previous one in the hierarchy.

Appendix B

Appendix of post clustering inference part

B.1 Proof of Theorem 1

Proof. The proof of Theorem 1 is adapted from the proof of Theorem 1 of Gao et al. (2024) which is similar to the proof of Theorem 3.1 in Loftus and Taylor (2015), the proof of Lemma 1 in Yang et al. (2016), and the proof of Theorem 3.1 in Chen and Bien (2020).

For any $\eta \in \mathbb{R}^n$, we have

$$\begin{aligned}
\mathbf{X} &= \pi_\eta^\perp \mathbf{X} + (\mathbf{I}_n - \pi_\eta^\perp)\mathbf{X} \\
&= \pi_\eta^\perp \mathbf{X} + \frac{\eta\eta^\top}{\|\eta\|_2^2}\mathbf{X} \\
&= \pi_\eta^\perp \mathbf{X} + \frac{\left\|\eta^\top \mathbf{X}\right\|_2}{\|\eta^\top \mathbf{X}\|_2}\eta\frac{\eta^\top \mathbf{X}}{\|\eta\|_2^2} \\
&= \pi_\eta^\perp \mathbf{X} + \left(\frac{\left\|\eta^\top \mathbf{X}\right\|_2}{\|\eta\|_2^2}\right)\eta \mathrm{dir}(\eta^\top \mathbf{X}).
\end{aligned} \tag{B.1}$$

Substituting $\mathbf{X}$ into the definition of $p(\mathbf{x}; \{C_k, C_{k'}\})$ given by Equation (5.11) yields

$$p(\mathbf{x}; \{C_k, C_{k'}\}) = \mathbb{P}_{\mathcal{H}_0}\left(\left\|\eta^\top \mathbf{X}\right\|_2 \geq \left\|\eta^\top \mathbf{x}\right\|_2 \;\middle|\; C_k, C_{k'} \in \mathcal{C}\left(\pi_\eta^\perp \mathbf{x} + \left(\frac{\left\|\eta^\top \mathbf{X}\right\|_2}{\|\eta\|_2^2}\right)\eta \mathrm{dir}(\eta^\top \mathbf{x})\right), \blacksquare\right) \tag{B.2}$$

$$\blacksquare = \left\{\pi_\eta^\perp \mathbf{X} = \pi_\eta^\perp \mathbf{x},\ \mathrm{dir}(\eta^\top \mathbf{X}) = \mathrm{dir}(\eta^\top \mathbf{x})\right\}.$$

At this stage, only $\left\|\eta^\top \mathbf{X}\right\|_2$ is random by fixing the values of $\pi_\eta^\perp \mathbf{X}$ and $\mathrm{dir}(\eta^\top \mathbf{X})$. For Gaussian data such that $X_i \sim \mathcal{N}(\mu_i, \sigma^2 I_m)$ and $\mathbf{X} = (X_i)_{i=1,\ldots,n}$, all that remains is to show that this quantity is independent of the over-conditioning,

$$\left\|\eta^\top \mathbf{X}\right\|_2 \perp\!\!\!\perp \pi_\eta^\perp \mathbf{X}, \tag{B.3}$$

and that under $\mathcal{H}_0 : \eta^\top \boldsymbol{\mu} = 0_m$,

$$\left\|\eta^\top \mathbf{X}\right\|_2 \perp\!\!\!\perp \mathrm{dir}(\eta^\top \mathbf{X}). \tag{B.4}$$

Proof of (B.3): Recall that $\pi_\eta^\perp$ is the orthogonal projection matrix onto the subspace orthogonal to η. Thus, $\pi_\eta^\perp \eta = 0_n$. It follows from properties of the matrix normal and multivariate normal distributions, and the Cochran's theorem that $\pi_\eta^\perp \mathbf{X} \perp\!\!\!\perp \eta^\top \mathbf{X}$.

Proof of (B.4): It follows from $X_i \overset{ind}{\sim} \mathcal{N}(\mu_i, \sigma^2 I_m)$. Thus, under $\mathcal{H}_0 : \eta^\top \boldsymbol{\mu} = 0_m$,

$$\frac{\eta^\top \mathbf{X}}{\sigma \|\eta\|_2} \sim \mathcal{N}(0_m, \mathbf{I}_m), \tag{B.5}$$

and (B.4) follows from the independence of the length and direction of a standard multivariate normal random vector.

We now apply (B.3) and (B.4) to (B.2). This yields

$$p(\mathbf{x}; \{C_k, C_{k'}\}) = \mathbb{P}_{\mathcal{H}_0}\left(\left\|\eta^\top \mathbf{X}\right\|_2 \geq \left\|\eta^\top \mathbf{x}\right\|_2 \;\middle|\; C_k, C_{k'} \in \mathcal{C}\left(\pi_\eta^\perp \mathbf{x} + \left(\frac{\left\|\eta^\top \mathbf{X}\right\|_2}{\|\eta\|_2^2}\right) \eta \ \mathrm{dir}(\eta^\top \mathbf{x})\right)\right) \tag{B.6}$$

Under $\mathcal{H}_0 : \eta^\top \boldsymbol{\mu} = 0_m$,

$$\frac{\left\|\eta^\top \mathbf{X}\right\|_2^2}{\sigma^2 \left(\|\eta\|_2^2\right)} \sim \chi_m^2,$$

by (B.5). From the definition of $\mathcal{S}(\mathbf{x}; \{C_k, C_{k'}\})$ in Equation (5.13),

$$p(\mathbf{x}; \{C_k, C_{k'}) = \mathbb{P}_{\mathcal{H}_0}\left(\left\|\eta^\top \mathbf{X}\right\|_2 \geq \left\|\eta^\top \mathbf{x}\right\|_2 \;\middle|\; \left\|\eta^\top \mathbf{X}\right\|_2 \in \mathcal{S}(\mathbf{x}; \{C_k, C_{k'}\})\right). \tag{B.7}$$

Therefore, for $\phi \sim (\sigma \|\eta\|_2) \cdot \chi_m$, it follows from (B.7) that

$$\begin{aligned} p(\mathbf{x}; \{C_k, C_{k'}\}) &= \mathbb{P}\left(\phi \geq \left\|\eta^\top \mathbf{x}\right\|_2 \;\middle|\; \phi \in \mathcal{S}(\mathbf{x}; \{C_k, C_{k'}\})\right) \\ &= 1 - \mathbb{F}\left(\left\|\eta^\top \mathbf{x}\right\|_2; \sigma \|\eta\|_2, \mathcal{S}(\mathbf{x}; \{C_k, C_{k'}\})\right), \end{aligned} \tag{B.8}$$

where $\mathbb{F}(\cdot; c, \mathcal{S})$ denotes the cumulative distribution function of a scale Chi distribution $c \cdot \chi_m$ truncated to the set $\mathcal{S}$.

Proof of (5.15): It follows from the definition of $p(\cdot; \{C_k, C_{k'}\})$ in Equation (5.11) that for all $\mathbf{x} \in \mathbb{R}^{n \times m}$,

$$\mathbb{P}_{\mathcal{H}_0}\Big(p(\mathbf{X}; \{C_k, C_{k'}\}) \leq \alpha \mid C_k, C_{k'} \in \mathcal{C}(\mathbf{X}), \blacksquare\Big) = \alpha, \tag{B.9}$$

where $\blacksquare$ is defined in Equation (B.2). Therefore, we have

$$\begin{aligned} &\mathbb{P}_{\mathcal{H}_0}\Big(p(\mathbf{X}; \{C_k, C_{k'}\}) \leq \alpha \;\Big|\; C_k, C_{k'} \in \mathcal{C}(\mathbf{X})\Big) \\ &= \mathbb{E}_{\mathcal{H}_0}\Big[\mathbb{1}_{\{p(\mathbf{X}; \{C_k, C_{k'}\}) \leq \alpha\}} \;\Big|\; C_k, C_{k'} \in \mathcal{C}(\mathbf{X})\Big] \\ &= \mathbb{E}_{\mathcal{H}_0}\left[\mathbb{E}_{\mathcal{H}_0}\Big[\mathbb{1}_{\{p(\mathbf{X}; \{C_k, C_{k'}\}) \leq \alpha\}} \;\Big|\; C_k, C_{k'} \in \mathcal{C}(\mathbf{X}), \pi_\eta^\perp \mathbf{X}, \mathrm{dir}(\eta^\top \mathbf{X})\Big] \;\middle|\; C_k, C_{k'} \in \mathcal{C}(\mathbf{X})\right] \\ &= \mathbb{E}_{\mathcal{H}_0}\Big[\alpha \;\Big|\; C_k, C_{k'} \in \mathcal{C}(\mathbf{X})\Big] \\ &= \alpha, \end{aligned}$$

where the second equality follows from the law of total expectation, and the third equality follows from (B.9).

□

B.2 Clustering of ordinal data

The clustering of a sample described by ordinal variables requires specific algorithms, standard algorithms like K-means or HAC with Euclidean distance are not suitable. Clustering approaches based on ordinal data can be grouped into three categories. The first category consists of distance-based methods. Traditional clustering methods can be adapted for ordinal data using appropriate distances and criteria. The distance proposed by Walesiak (1999) is specifically designed for ordinal data, adapting the Kendall (1938) rank correlation coefficient. Marden (1996) suggests normalizing ordinal data into continuous ranks (values between 0 and 1) to apply distances suitable for continuous data. The second category focuses on model-based approaches using mixture models for multinomial distribution assumptions (Lubke and Neale, 2008; DeSantis et al., 2008; Jollois and Nadif, 2009) or the Binary Ordinal Search model (Biernacki and Jacques, 2016; Selosse et al., 2018). The third category includes tandem approaches where a dimensionality reduction technique first transforms ordinal data into continuous data, followed by the application of standard clustering methods for continuous data (Hwang et al., 2006; Iodice D'Enza and Palumbo, 2013; Van Buuren and Heiser, 1989). A large literature employs algorithms designed for categorical data, which disregard the order of categories (i.e., nominal data), to perform clustering on ordinal data. However, this approach is not considered here.

Among these methods, we focus on three strategies particularly suitable for clustering ordinal data: (1)- the distance by Walesiak (1999), (2)- the mixture model by Biernacki and Jacques (2016) and (3)- normalized ranks from Marden (1996). These methods are compared in the context of aggregate ordinal unidimensional clusterings but the results are not reported in this manuscript. However, they present certain limitations. Specifically, Methods (1) and (2) have quadratic complexity with respect to the number of modalities L and the number of variables m, making them challenging to apply. This is particularly problematic for high-dimensional data or when used in the procedure of Bachoc et al. (2023), where the number of categories is implicitly driven by λ, potentially exceeding the recommended $L \leq 8$ constraint (as suggested by Biernacki and Jacques (2016) due to computational cost). Therefore, in our simulation, we use Method (3) of normalized ranks. Let Y_{ij} be a categorical variable with L known ordered categories encoded as $1, \ldots, L$. The normalized ordinal data is given by $\tilde{Y}_{ij} = \frac{Y_{ij}-1}{L-1} \in [0, 1]$. This transformation makes $\tilde{Y}_{ij}$ continuous, allowing to use the Euclidean distance and clustering methods HAC with Ward linkage or K-means.

B.3 Comparison of multivariate methods

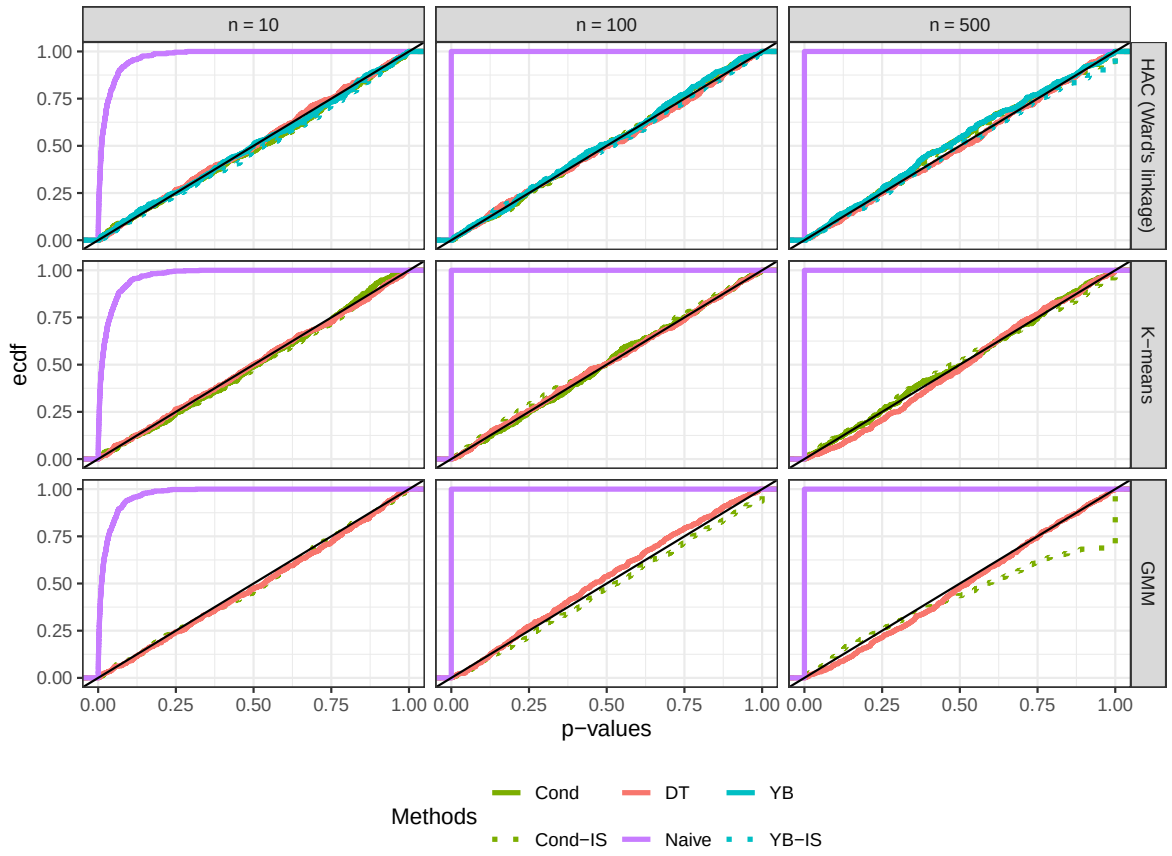

Figure B.1 – Evaluation of the type I error rate: ecdf of p-values under the null hypothesis estimating $K = 2$ clusters, using Setting 1 with $m = 2$, $\rho = 0$ and $a = 0$ (see Section 5.3.1.1). The ecdf curves are printed as a function of the value of n (in columns) and the clustering method (in rows). Only the naive method does not control the type I error rate. Note that the MC-Importance Sampling estimation of the conditional p-value of Gao et al. (2024) for the clustering obtained with the estimation of GMM is conservative.

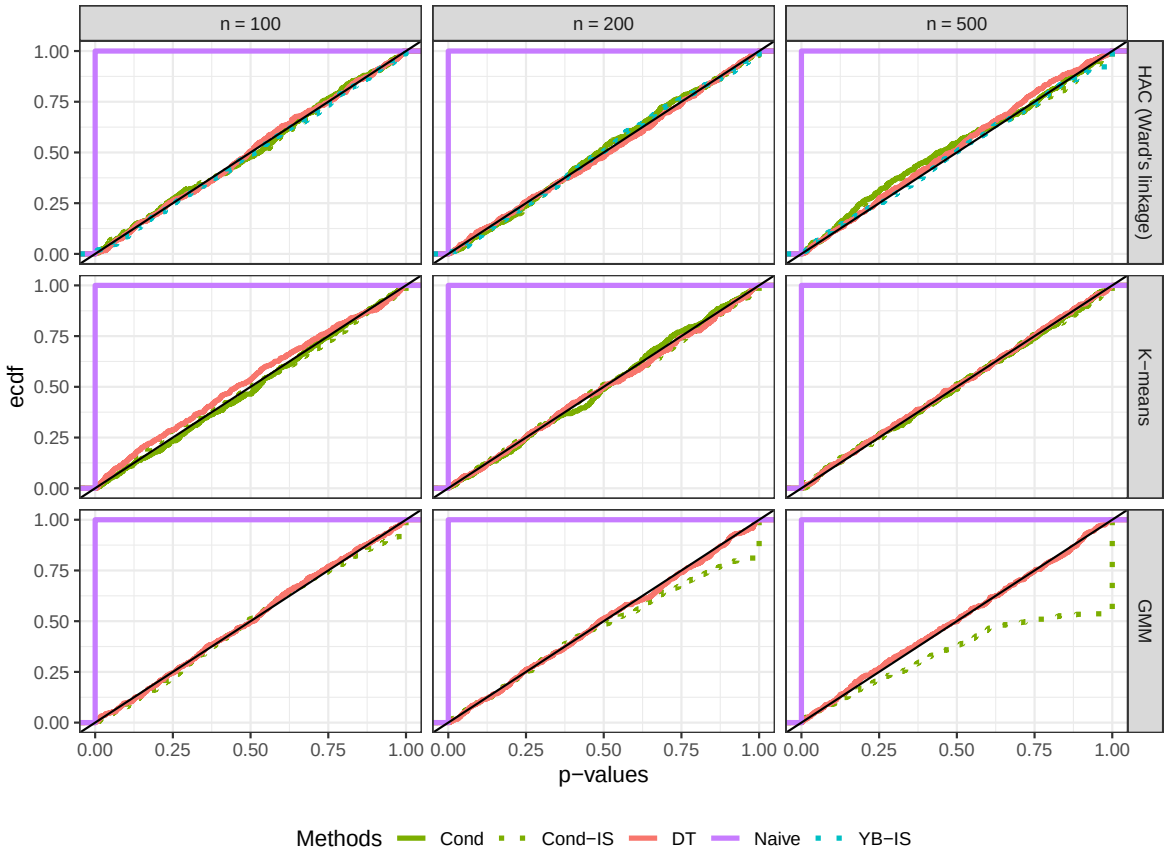

Figure B.2 – Evaluation of the type I error rate: ecdf of p-values under the null hypothesis estimating $K = 3$ clusters, using Setting 2 with $a = 0$ (see Section 5.3.1.1). The ecdf curves are printed as a function of the value of n (in columns) and the clustering method (in rows). Only the naive method does not control the type I error rate. Note that the MC-Importance Sampling estimation of the conditional p-value of Gao et al. (2024) for the clustering obtained with the estimation of GMM is conservative.

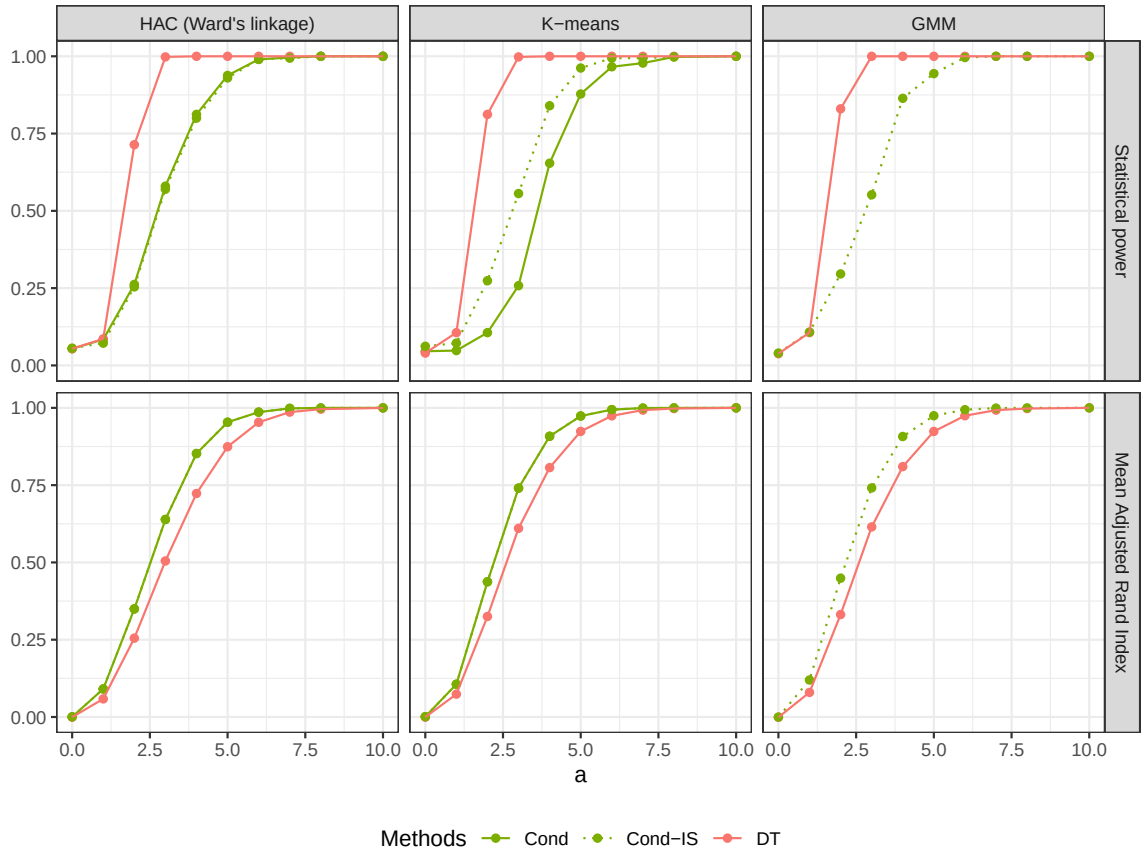

Figure B.3 – Statistical power and ARI computed for Setting 1 with $n = 100$ and $m = 2$ through the clustering methods. The power analysis is similar from one clustering method to another. While the data thinning method provides a less accurate clustering, it is more powerful method than the conditional approach. For the K-means method, the over-conditioning done to obtain an exact p-value results in a loss of statistical power. This effect is describe in Section 5.3.3

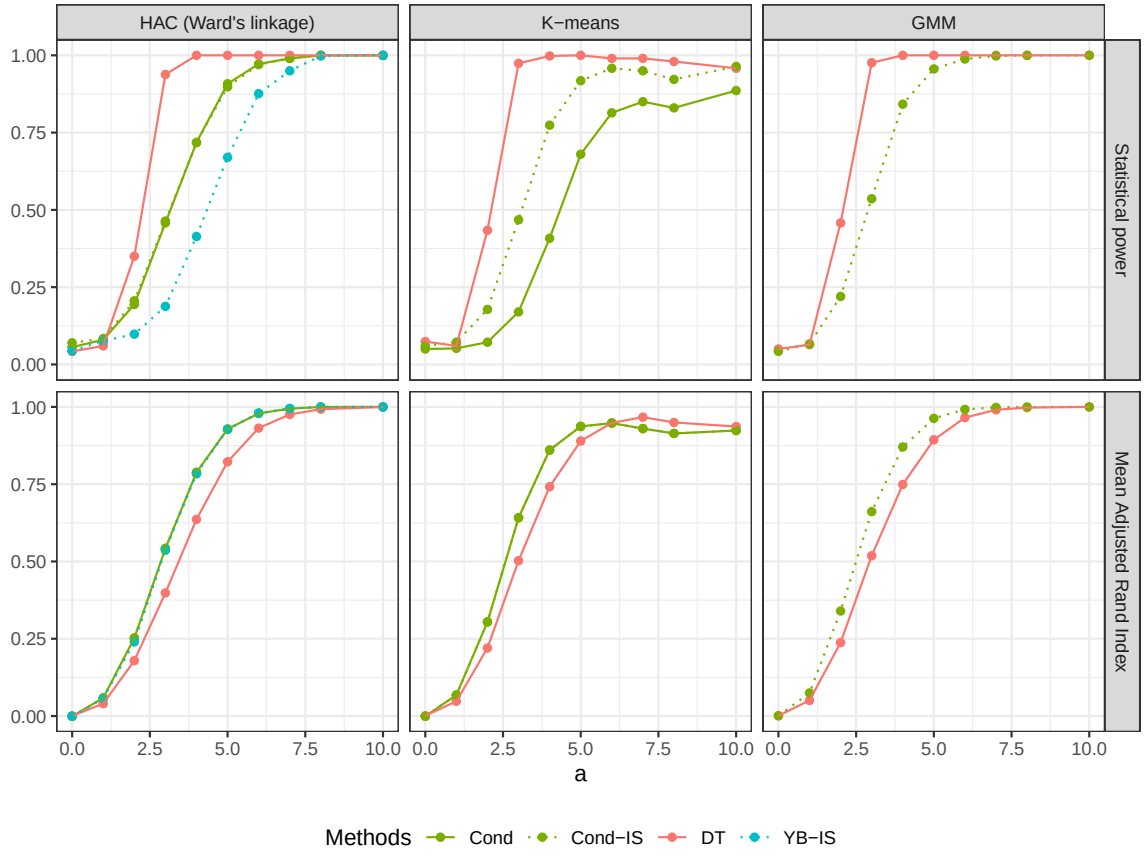

Figure B.4 – Statistical power and ARI for the multivariate test on Setting 2. The curves are printed as a function of the signal. Cases discriminate the clustering method (in rows). The power analysis is similar from one clustering method to another. While the data thinning method provides a less accurate clustering, it is more powerful method than the conditional approach. Specifically, for the K-means method, the overconditioning done to obtain an exact p-value results in a loss of statistical power. This effect is described in Section 5.3.3. Moreover, the loss of power observed for the K-means clustering method comes from the loss of recovery of the clustering.

B.3.1 Effect of the over-conditioning on the statistical power using the exact p-value for K-means clustering

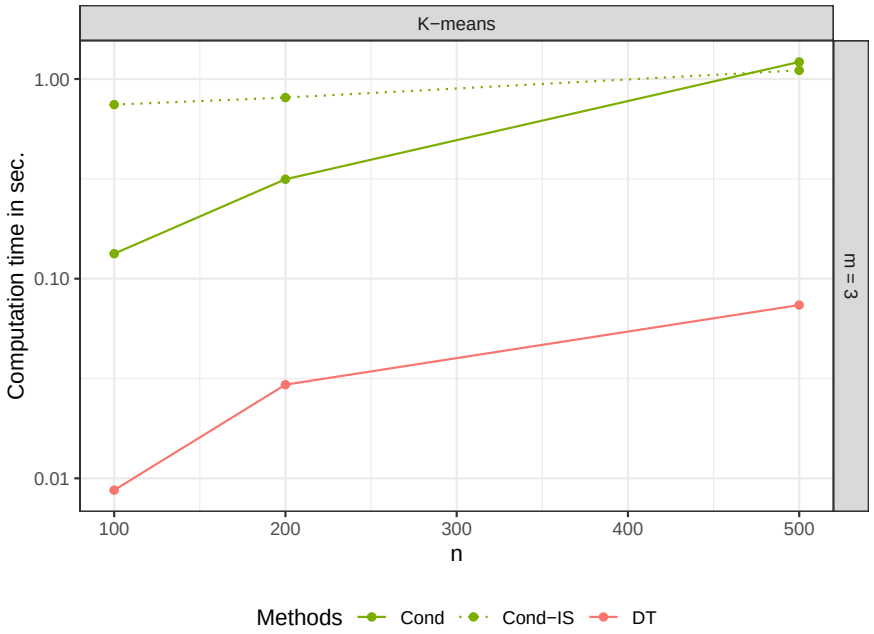

Figure B.5 – Time computation analysis for the K-means method using Setting 2. The data thinning procedure is faster than the conditional approach. For $n = 500$, the MC-Importance Sampling estimation is faster than the exact conditional p-value.

B.3.2 What thinning value ε should be used?

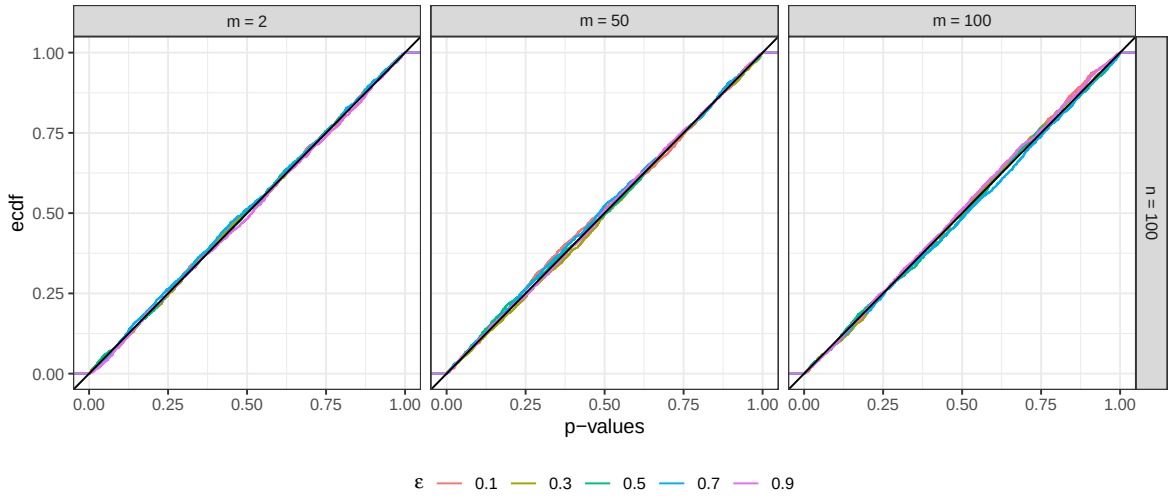

Figure B.6 – Evaluation of the type I error rate: ecdf of p-values under the null hypothesis for several values of ε in the data thinning methods for Setting 1 with $n = 100$. The method controls the type I error rate for any value of ε.

B.3.3 Explanation of the hyperparameter selection in the methods

How many Monte Carlo draws do we use for multivariate conditional approach? The estimation by MC-Importance Sampling is impacted by the number of simulations made for the estimation. Finding a minimal but sufficient number of simulations to obtain a good estimate without taking too much time is always interesting. To assess this, Setting 1 was executed with different values of Monte Carlo simulations, namely $Q \in \{100, 200, 500, 1000\}$. This simulation was performed only with the HAC clustering method for n and m, which are discussed in the main section.

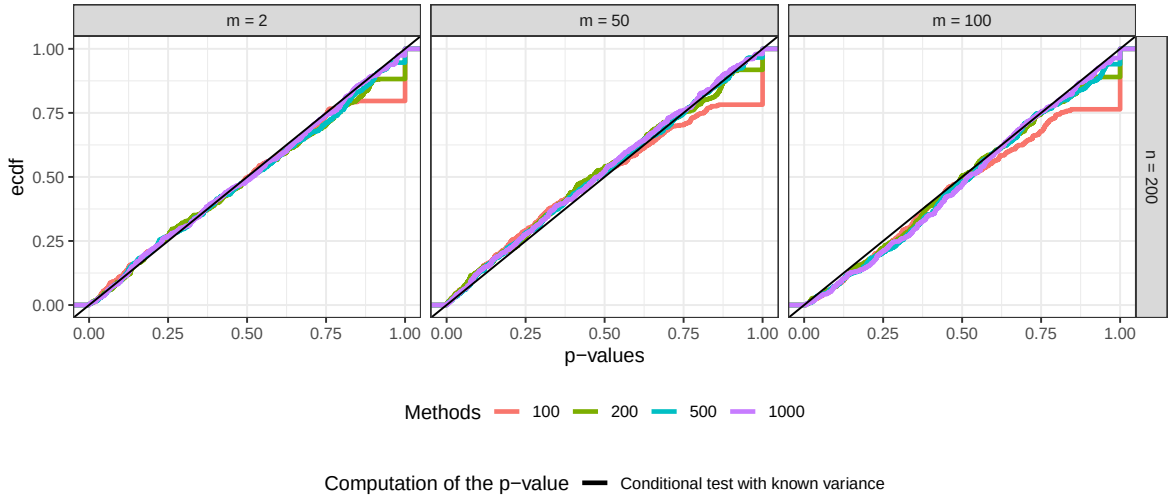

Figure B.7 – Comparison between the ecdf of p-values estimated with MC-Importance Sampling for $Q \in \{100, 200, 500, 1000\}$ draws, using Setting 1 with the HAC clustering method.

To simplify visualization, and since the results are similar for different numbers of individuals, Figure B.7 presents the results for $n = 200$. As expected, the tests are valid regardless of the number of Monte Carlo simulations performed. However, it is interesting to note that with fewer draws, there are more p-values close to 1. This effect confirms the test's conservativeness if there are not enough resources to estimate the p-value. Estimating the p-value involves examining the proportion of simulations in which the value ϕ in Equation (5.20) is greater than the observed test statistic among perturbations that preserve clustering. Thus, in the case of Gao et al. (2024), this can be understood as only perturbations that separate the clusters can maintain them.

Effect of the parameter γ in Yun and Foygel Barber (2023) method with MC-Importance Sampling Figure 5.7 shows that Yun and Foygel Barber (2023)'s method using the estimation of p-values by MC-Importance Sampling (YB-IS) gives a small excess of p-value below 0.25 for large samples $n = 500$ and $m = 10$. This effect indicates that the method is slightly anti-conservative. Indeed, the MC-Importance Sampling samples perturbation following a truncated normal distribution with a variance γ^2 defined by the user. As this value is not explicitly driven by data or method, the user has to choose this value such that approximately half of the perturbations preserve the clustering. This is equivalent to set the accuracy rate to recover the clustering to 0.5, where the accuracy rate is the proportion of MC-Importance Sampling draws which recover the true clustering partition. For our numerical simulation (Section 5.3.2), we used $\gamma = 0.05$ as proposed in numerical experiments in Yun and Foygel Barber (2023). Originally, the experiments were made with $n = 30$ and

$m = 2$, which is far from our settings. This section aims to try to find a value allowing the method to control the type I error rate.

Data are simulated from Setting 1, with $n \in \{30, 500\}$, $m \in \{2, 100\}$, $a = 0$ and $\Sigma = I_m$. The method YB-IS is used with $Q = 1000$ draws and $\gamma \in \{0.001, 0.01, 0.05, 0.25, 0.5, 0.75, 1\}$.

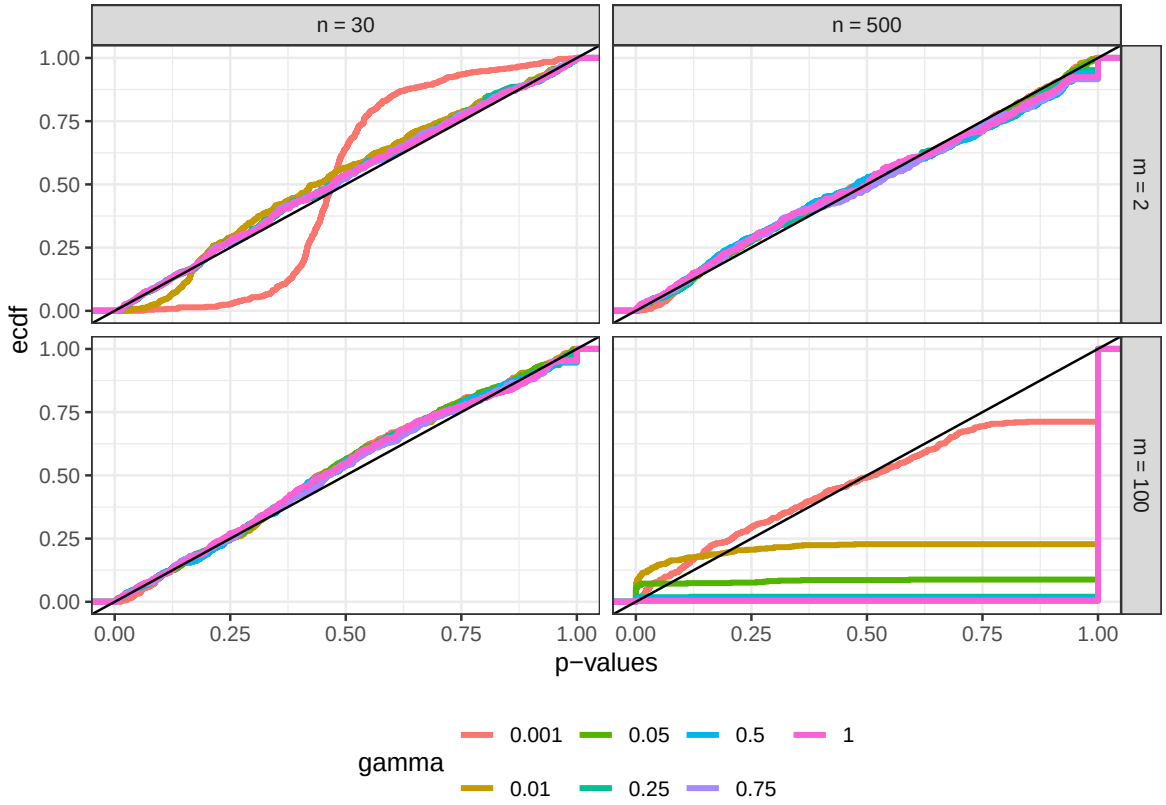

Figure B.8 – Ecdf of p-values obtained with YB-HAC-IS for different values of γ on simulated data from Setting 1.

Figure B.8 displays the ecdfs of p-values for various values of γ. Figure B.9 shows the clustering accuracy rate. The plots are based on 500 simulations of $\mathbf{X}$, using YFB-IS method with $Q = 1000$ draws. The value $\gamma = 0.05$ used in Yun and Foygel Barber (2023) yields a valid test for their case. In the case with $n = 500$ and $m = 100$, there are many p-values at 1, but the other p-values are stochastically lower than uniform, leading this method to not control the type I error rate.

Choosing a low value of γ results in sampling the perturbation ω from a very narrow distribution, approaching a Dirac delta distribution. In this case, for $n = 30$ and $m = 2$, the p-value are not uniform with a mode at 0.5 (see Figure B.8). In this case, all perturbation preserve the clustering (see Figure B.9). For the same value of γ, the p-values are uniform under $\mathcal{H}_0$ for Settings ($n = 500$, $m = 2$) and ($n = 30$, $m = 100$) while the accuracy rate is lower with the median at 0.8. An high accuracy rate implies that the MC-Importance Sampling do not explore a large range of ω. For these settings, all values of γ give uniform p-values (see Figure B.8). Nevertheless, only $\gamma = 0.05$ gives a median accuracy rate at 0.5 (as specified by Yun and Foygel Barber (2023)).

Choosing this method parameter is challenging due to its lack of interpretation and direct linkage with the data. The previous results do not aid in refining this choice, that for high

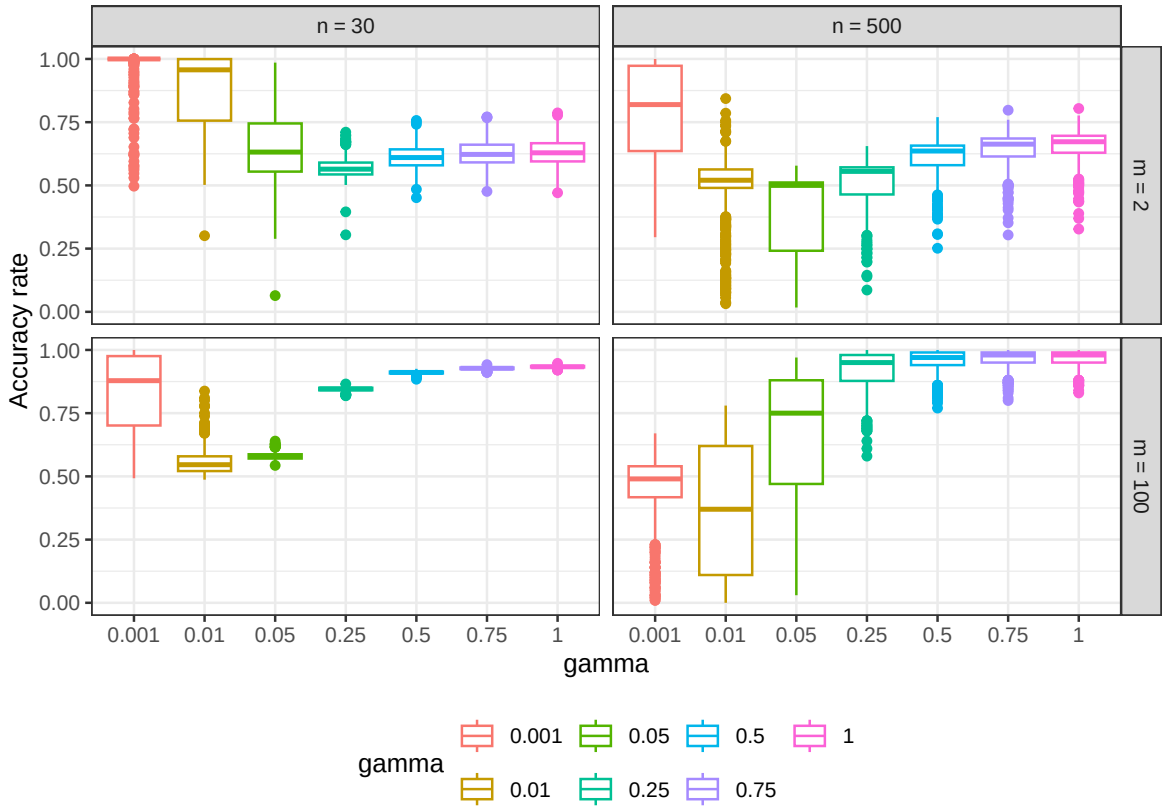

Figure B.9 – Distribution (through 500 experiments) of Accuracy rate of clusterings through 1000 Important Sampling draws. Data are simulated from Setting 1

dimensional data the test seems not to be calibrated for any value of γ. The case illustrated by Yun and Foygel Barber (2023) demonstrates that there exists a value of γ that yields a valid method. Thus, the poor results obtained in our settings seem to be due to a poor choice of γ. More investigation is needed to refine this choice.

B.3.4 Supplementary figures of Section 7.4 : Impact of the Gaussian Mixture Model estimation of the covariance matrix

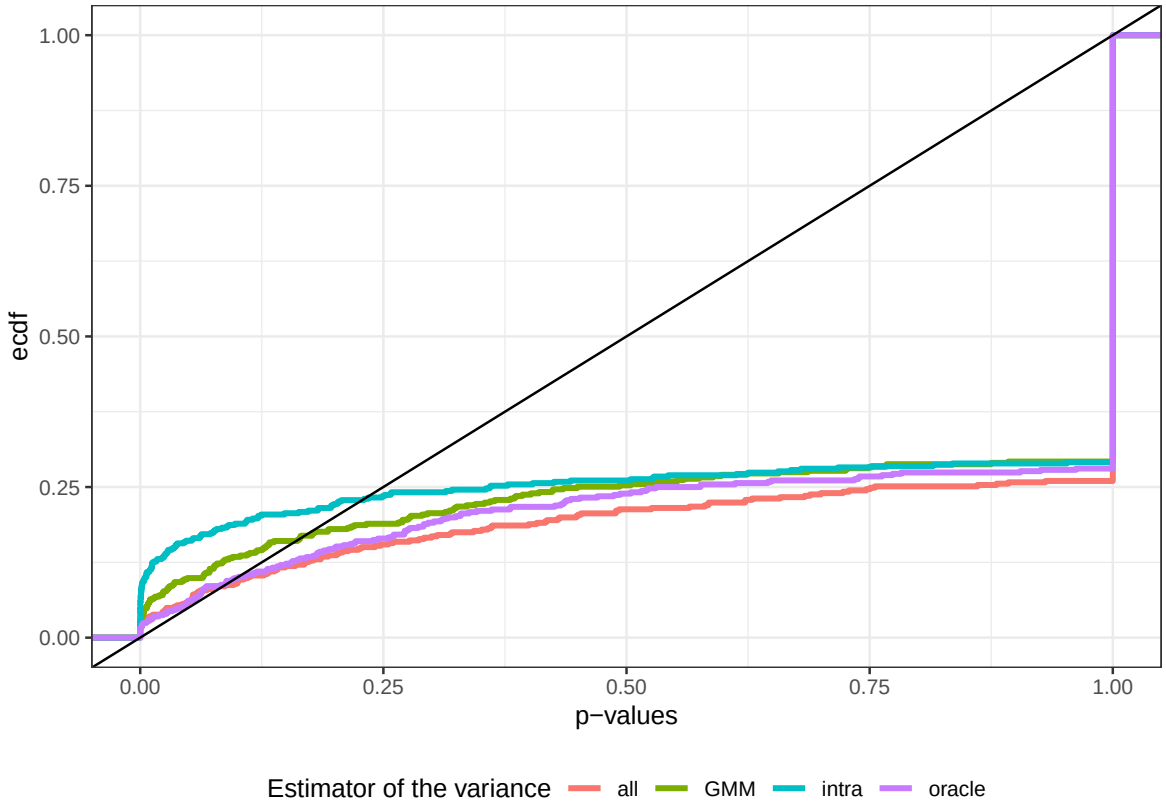

Figure B.10 – ECDF of p-values under $\mathcal{H}_0$ using Setting 1 with $m = 2$, $n = 500$, $\Sigma = I_m$ (see Section 5.3.1.1). The test is evaluated with the conditional approach of Gao et al. (2024) using the clustering obtain by the estimation of GMM (with $p_k LI$ assumption) and the p-values are estimated by MC-Importance Sampling with $Q = 1000$ draws (see Equation (5.20)). The test is computed with the known covariance matrix $\Sigma = I_m$ (oracle), and the plug-in estimators $\hat{\sigma}^2_{\text{all}}\, I_m$ (all), $\hat{\sigma}^2_{\text{intra}}\, I_m$ (intra) and $\hat{\sigma}_{\text{GMM}}\, I_m$ (GMM) (see Section 7.4 for the expression of these estimators). The MC-Importance Sampling estimation for the GMM algorithm sets many p-values to 1. With this $p_k LI$ assumption, the method of Gao et al. (2024) seems not to control the type I error rate, even with the known covariance matrix. This effect seems to come from clustering where clusters have different π_k proportions.

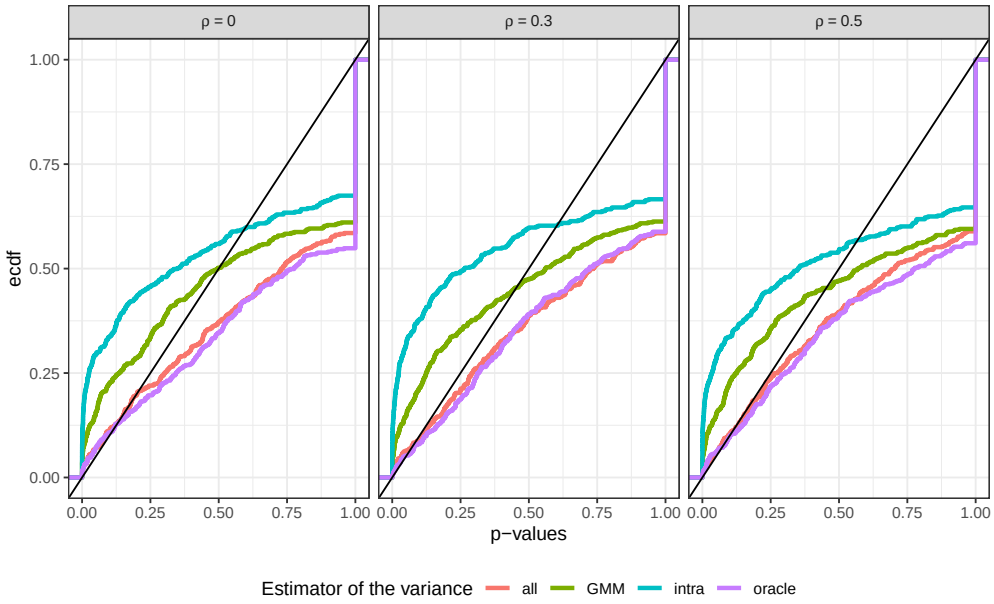

Figure B.11 – ECDF of p-values under $\mathcal{H}_0$ using Setting 1 with $m = 2$, $n = 500$, $\Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$ (see Section 5.3.1.1). The test is evaluated with the conditional approach of Gao et al. (2024) using the EM algorithm for GMM with pLC shape and the p-values are estimated by MC-Importance Sampling with $Q = 1000$ draws (see Equation (5.20)). The test is computed with the known covariance matrix Σ (oracle), and the plug-in estimators $\hat{\Sigma}_{\text{all}}$ (all), $\hat{\Sigma}_{\text{intra}}$ (intra) and $\hat{\Sigma}_{\text{GMM}}$ (GMM) (see Section 7.4 for the expression of these estimators). The MC-Importance Sampling estimation for the GMM algorithm sets many p-values to 1. As the Gao et al. (2024)'s method controls the type I error rate with the oracle, the estimator 'all' also controls it. 'intra' and 'GMM' estimator do not control the type I error rate.

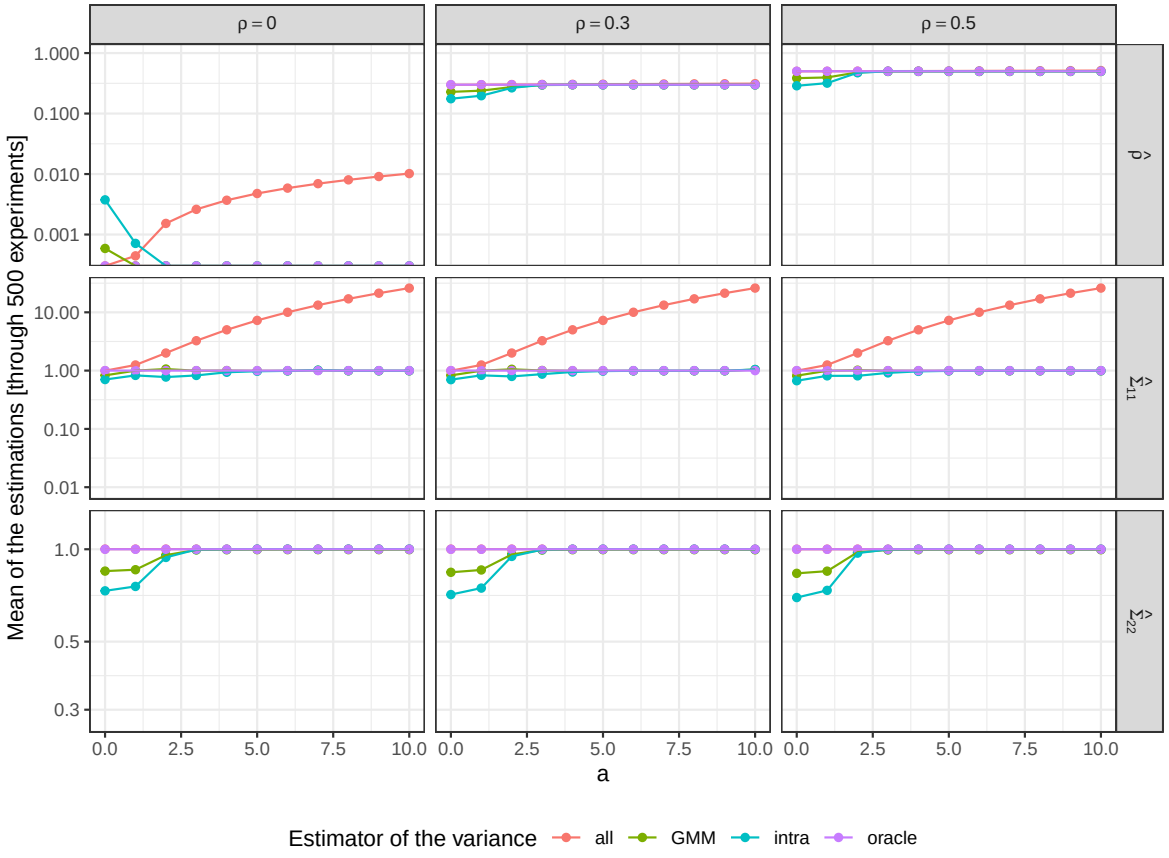

Figure B.12 – Evolution of the estimation of Σ in function of the signal. The value 'oracle' is the true value of Σ. The variance is overestimated by the estimator 'all'. The estimator 'intra' under-estimates the variance for $a \leq 4$. The 'GMM' (with a pLC assumption) estimator seems to be a good estimator since it underestimates the variance for $a \leq 2$.

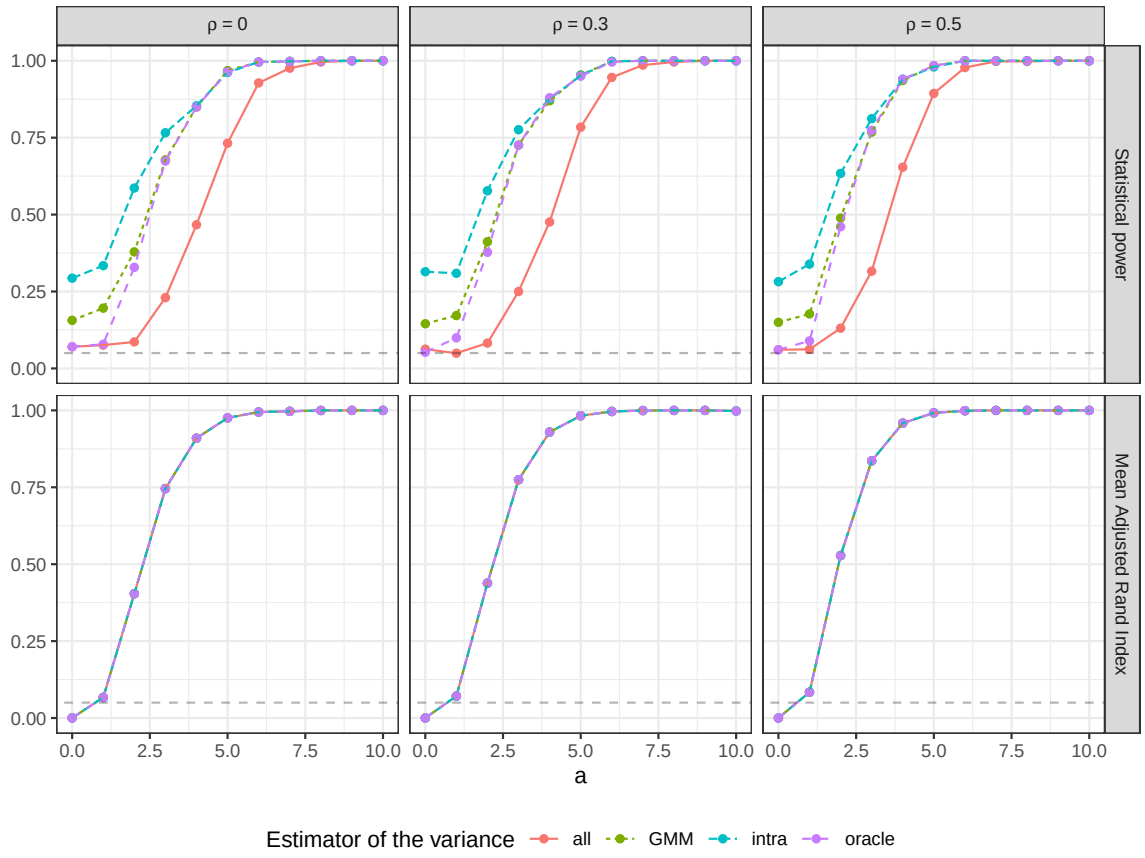

Figure B.13 – Statistical power and ARI using Setting 1 with $m = 2$, $n = 500$, $\Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$ (see Section 7.4). The test is evaluated with the conditional approach of Gao et al. (2024) using the EM algorithm for GMM with pLC model and the p-values are estimated by MC-Importance Sampling with $Q = 1000$ draws (see Equation (5.20)). The test is computed with the known covariance matrix Σ (oracle), and the plug-in estimators $\hat{\Sigma}_{\text{all}}$ (all), $\hat{\Sigma}_{\text{intra}}$ (intra) and $\hat{\Sigma}_{\text{GMM}}$ (GMM) (see Section 7.4 for expression of these estimators). The 'GMM' estimator seems to be a powerful as the oracle test. The 'intra' estimator gives an invalid test (see Figure7.2), the method cannot be interpreted as powerful with this estimator. The 'all' estimator overestimate the variance, loosing statistical power. For high signal $a \geq 8$, the method is powerful.

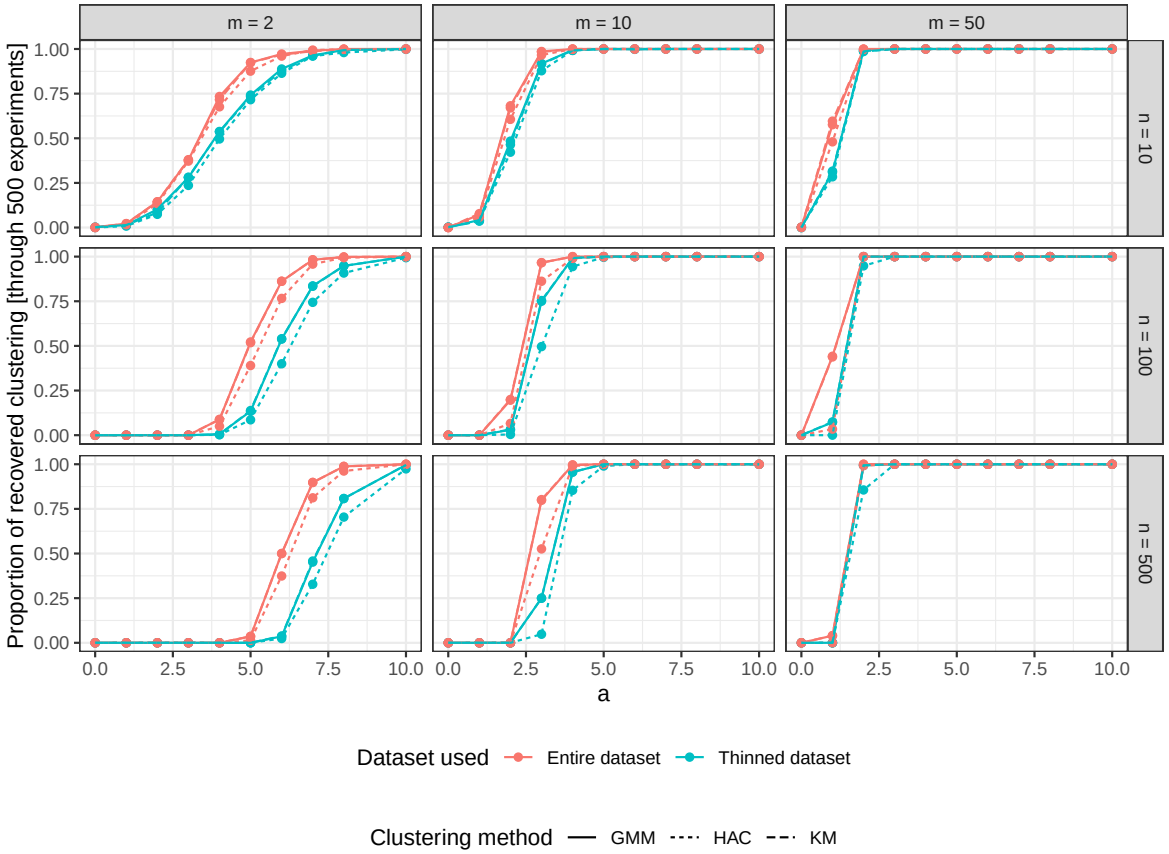

Figure B.14 – Proportion of simulations recovering the true partition of Setting 1 in function of the signal a, on the simulations made in Section 5.3. The clusterings (HAC, K-means and GMM) are estimated on the entire dataset, and the thinned dataset is obtained with the data thinning procedure using $\varepsilon = 0.7$. GMM results are superimposed on K-means results. With enough signal the clusterings can recover the true partition. The entire dataset provides more information to recover the true partition than the thinned dataset.

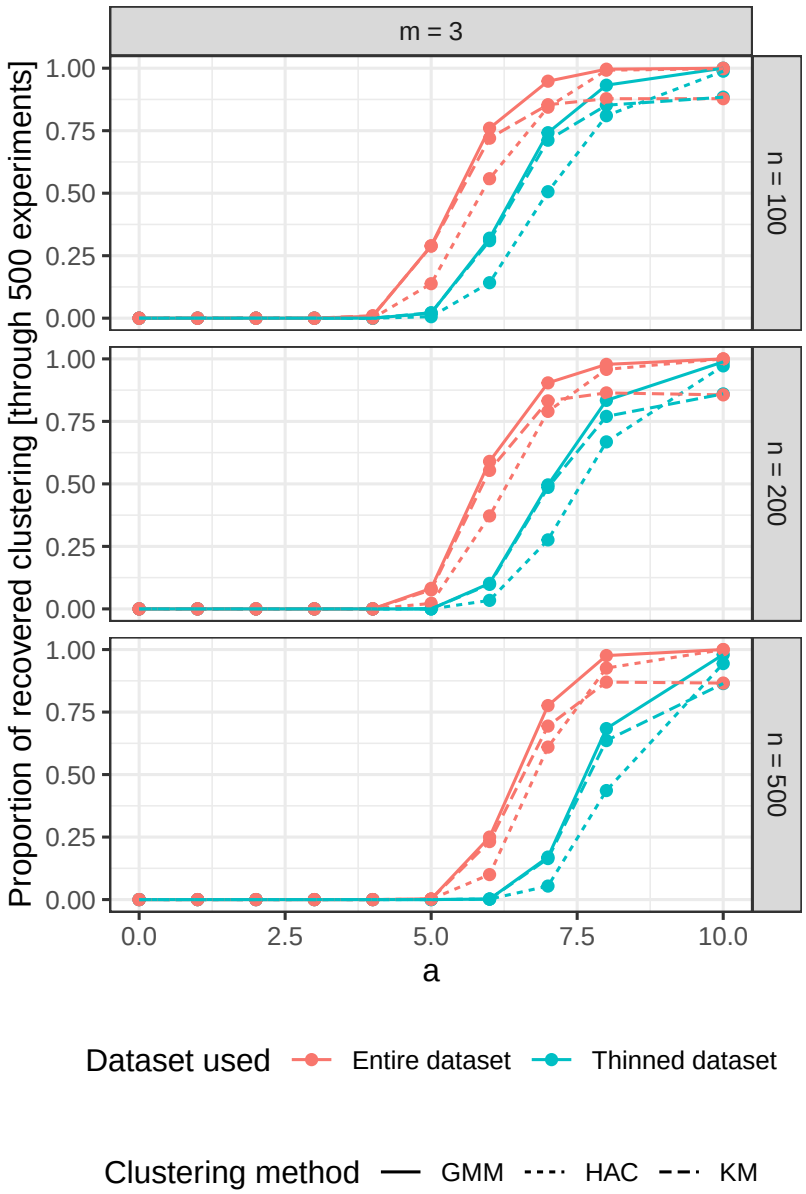

Figure B.15 – Proportion of simulation recovering the true partition of Setting 2 in function of the signal a, on the simulation made in Section 5.3. The clusterings (HAC, K-means and GMM) are estimated on the entire dataset, and the thinned dataset is obtained with the data thinning procedure with $\varepsilon = 0.7$. With enough signal ($a \geq 6$) the clusterings can recover the true partition. The entire dataset provides more information to recover the true partition than the thinned dataset.

Bibliography

Abraham, A., Pedregosa, F., Eickenberg, M., Gervais, P., Mueller, A., Kossaifi, J., Gramfort, A., Thirion, B., and Varoquaux, G. (2014). Machine learning for neuroimaging with scikit-learn. *Frontiers in Neuroinformatics*, 8:71792.

Anders, S. and Huber, W. (2010). Differential expression analysis for sequence count data. *Nature Precedings*, pages 1–1.

Andreella, A., Hemerik, J., Finos, L., Weeda, W., and Goeman, J. (2023). Permutation-based true discovery proportions for functional magnetic resonance imaging cluster analysis. *Statistics in Medicine*, 42(14):2311–2340.

Andrews, T. S. and Hemberg, M. (2017). Modelling dropouts for feature selection in scrnaseq experiments. *bioRxiv*.

Bacher, R. and Kendziorski, C. (2016). Design and computational analysis of single-cell rna-sequencing experiments. *Genome biology*, 17(1):63.

Bachoc, F., Maugis-Rabusseau, C., and Neuvial, P. (2023). Selective inference after convex clustering with l1 penalization. *arXiv preprint arXiv:2309.01492*.

Bahr, T. M., Hughes, G. J., Armstrong, M., Reisdorph, R., Coldren, C. D., Edwards, M. G., Schnell, C., Kedl, R., LaFlamme, D. J., Reisdorph, N., et al. (2013). Peripheral blood mononuclear cell gene expression in chronic obstructive pulmonary disease. *American Journal of Respiratory Cell and Molecular Biology*, 49(2):316–323.

Banfield, J. D. and Raftery, A. E. (1993). Model-based gaussian and non-gaussian clustering. *Biometrics*, pages 803–821.

Barrett, T., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., Tomashevsky, M., Marshall, K. A., Phillippy, K. H., Sherman, P. M., Holko, M., et al. (2012). Ncbi geo: archive for functional genomics data sets–update. *Nucleic Acids Research*, 41(D1):D991–D995.

Becht, E., McInnes, L., Healy, J., Dutertre, C.-A., Kwok, I. W. H., Ng, L. G., Ginhoux, F., and Newell, E. W. (2018). Dimensionality reduction for visualizing single-cell data using umap. *Nature Biotechnology*, 37:38–44.

Belloni, A., Chernozhukov, V., Fernandez-Val, I., and Hansen, C. (2017). Program evaluation and causal inference with high-dimensional data. *Econometrica*, 85(1):233–298.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: series B (Methodological)*, 57(1):289–300.

Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics*, 29(4):1165–1188.

Biernacki, C., Celeux, G., and Govaert, G. (2000). Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(7):719–725.

Biernacki, C., Celeux, G., Govaert, G., and Langrognet, F. (2006). Model-based cluster and discriminant analysis with the mixmod software. *Computational Statistics & Data Analysis*, 51(2):587–600.

Biernacki, C. and Jacques, J. (2016). Model-based clustering of multivariate ordinal data relying on a stochastic binary search algorithm. *Statistics and Computing*, 26(5):929–943.

Blain, A., Thirion, B., Grisel, O., and Neuvial, P. (2024). False discovery proportion control for aggregated knockoffs. *Advances in Neural Information Processing Systems*, 36.

Blain, A., Thirion, B., and Neuvial, P. (2022). Notip: Non-parametric true discovery proportion estimation for brain imaging. *arXiv preprint arXiv:2204.10572*.

Blanchard, G., Neuvial, P., and Roquain, E. (2020). Post hoc confidence bounds on false positives using reference families. *Annals of Statistics*, 48(3):1281–1303.

Blanchard, G., Neuvial, P., and Roquain, E. (2021). On agnostic post hoc approaches to false positive control. In Cui, X., Dickhaus, T., Ding, Y., and Hsu, J. C., editors, *Handbook of Multiple Comparisons*, Handbooks of Modern Statistical Methods. Chapman & Hall/CRC.

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008.

Bouveyron, C., Celeux, G., Murphy, T. B., and Raftery, A. E. (2019). *Model-based clustering and classification for data science: with applications in R*, volume 50. Cambridge University Press.

Butler, A., Hoffman, P., Smibert, P., Papalexi, E., and Satija, R. (2018). Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature Biotechnology*, 36(5):411.

Celeux, G. (1985). The sem algorithm: a probabilistic teacher algorithm derived from the em algorithm for the mixture problem. *Computational Statistics Quarterly*, 2:73–82.

Celeux, G. and Govaert, G. (1992). A classification em algorithm for clustering and two stochastic versions. *Computational Statistics & Data Analysis*, 14(3):315–332.

Celeux, G. and Govaert, G. (1995). Gaussian parsimonious clustering models. *Pattern Recognition*, 28(5):781–793.

Chang, W., Cheng, J., Allaire, J., Sievert, C., Schloerke, B., Xie, Y., Allen, J., McPherson, J., Dipert, A., and Borges, B. (2021). *shiny: Web Application Framework for R*. R package version 1.6.0.

Chen, J. and Andrews, I. (2023). Optimal conditional inference in adaptive experiments. *arXiv preprint arXiv:2309.12162*.

Chen, S. and Bien, J. (2020). Valid inference corrected for outlier removal. *Journal of Computational and Graphical Statistics*, 29(2):323–334.

Chen, Y., Lun, A. T., and Smyth, G. K. (2014). Differential expression analysis of complex rna-seq experiments using edger. *Statistical Analysis of Next Generation Sequencing Data*, pages 51–74.

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

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

Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., Szcześniak, M. W., Gaffney, D. J., Elo, L. L., Zhang, X., et al. (2016). A survey of best practices for rna-seq data analysis. *Genome Biology*, 17:1–19.

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

Cox, D. R. (2006). *Principles of Statistical Inference*. Cambridge university press.

Cui, X. and Churchill, G. A. (2003). Statistical tests for differential expression in cDNA microarray experiments. *Genome Biology*, 4(4):210.

Davenport, S., Thirion, B., and Neuvial, P. (2022). Fdp control in multivariate linear models using the bootstrap. *arXiv preprint arXiv:2208.13724*.

Delyon, B., Lavielle, M., and Moulines, E. (1999). Convergence of a stochastic approximation version of the em algorithm. *Annals of Statistics*, pages 94–128.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: series B (methodological)*, 39(1):1–22.

DeSantis, S. M., Houseman, E. A., Coull, B. A., Stemmer-Rachamimov, A., and Betensky, R. A. (2008). A penalized latent class model for ordinal data. *Biostatistics*, 9(2):249–262.

Dharamshi, A., Neufeld, A., Motwani, K., Gao, L. L., Witten, D., and Bien, J. (2024). Generalized data thinning using sufficient statistics. *Journal of the American Statistical Association*, pages 1–26.

Dudoit, S., Van Der Laan, M. J., and van der Laan, M. J. (2008). *Multiple Testing Procedures with Applications to Genomics*. Springer.

Dunn, O. J. (1961). Multiple comparisons among means. *Journal of the American Statistical Association*, 56(293):52–64.

Durand, G., Blanchard, G., Neuvial, P., and Roquain, E. (2020). Post hoc false positive control for structured hypotheses. *Scandinavian Journal of Statistics*.

Duy, V. N. L., Iwazaki, S., and Takeuchi, I. (2022). Quantifying statistical significance of neural network-based image segmentation by selective inference. *Advances in Neural Information Processing Systems*, 35:31627–31639.

Ebrahimpoor, M. and Goeman, J. (2021). Inflated false discovery rate due to volcano plots: Problem and solutions. *Briefings in Bioinformatics*, 22.

Ebrahimpoor, M., Spitali, P., Hettne, K., Tsonaka, R., and Goeman, J. (2020). Simultaneous enrichment analysis of all possible gene-sets: unifying self-contained and competitive methods. *Briefings in Bioinformatics*, 21(4):1302–1312.

Efron, B., Tibshirani, R., et al. (2007). On testing the significance of sets of genes. *The Annals of Applied Statistics*, 1(1):107–129.

Enjalbert-Courrech, N. (2024). *IIDEA: Interactive Inference for Differential expression analyses*. R package version 0.0.0.9000.

Enjalbert-Courrech, N. and Neuvial, P. (2022). Powerful and interpretable control of false discoveries in two-group differential expression studies. *Bioinformatics*, 38(23):5214–5221.

Estivill-Castro, V. (2002). Why so many clustering algorithms: a position paper. *ACM SIGKDD explorations newsletter*, 4(1):65–75.

Fithian, W., Sun, D., and Taylor, J. (2014). Optimal inference after model selection. *arXiv preprint arXiv:1410.2597*.

Gao, L. L., Bien, J., and Witten, D. (2024). Selective inference for hierarchical clustering. *Journal of the American Statistical Association*, 119(545):332–342.

Gauthier, M., Agniel, D., Thiébaut, R., and Hejblum, B. P. (2021). Distribution-free complex hypothesis testing for single-cell rna-seq differential expression analysis. *bioRxiv*.

Ge, Y., Dudoit, S., and Speed, T. P. (2003). Resampling-based multiple testing for microarray data analysis. *Test*, 12(1):1–77.

Geistlinger, L., Csaba, G., Santarelli, M., Ramos, M., Schiffer, L., Law, C., Turaga, N., Davis, S., Carey, V., Morgan, M., Zimmer, R., and Waldron, L. (2020). Toward a gold standard for benchmarking gene set enrichment analysis. *Briefings in Bioinformatics*.

Gene Ontology, C. (2015). Gene ontology consortium: going forward. *Nucleic Acids Research*, 43(D1):D1049–D1056.

Genovese, C. R. and Wasserman, L. (2006). Exceedance control of the false discovery proportion. *Journal of the American Statistical Association*, 101(476):1408–1417.

Goeman, J. J. and Bühlmann, P. (2007). Analyzing gene expression data in terms of gene sets: methodological issues. *Bioinformatics*, 23(8):980–987.

Goeman, J. J., Górecki, P., Monajemi, R., Chen, X., Nichols, T. E., and Weeda, W. (2023). Cluster extent inference revisited: quantification and localisation of brain activity. *Journal of the Royal Statistical Society: series B (Methodological)*, 85(4):1128–1153.

Goeman, J. J., Meijer, R. J., Krebs, T. J., and Solari, A. (2019). Simultaneous control of all false discovery proportions in large-scale multiple hypothesis testing. *Biometrika*, 106(4):841–856.

Goeman, J. J. and Solari, A. (2011). Multiple testing for exploratory research. *Statistical Science. A Review Journal of the Institute of Mathematical Statistics*, 26(4):584–597.

Goeman, J. J. and Solari, A. (2014). Multiple hypothesis testing in genomics. *Statistics in medicine*, 33(11):1946–1978.

Goeman, J. J., Van De Geer, S. A., De Kort, F., and Van Houwelingen, H. C. (2004). A global test for groups of genes: testing association with a clinical outcome. *Bioinformatics*, 20(1):93–99.

González-Delgado, J., Cortés, J., and Neuvial, P. (2023). Post-clustering inference under dependency. *arXiv preprint arXiv:2310.11822*.

Hafemeister, C. and Satija, R. (2019). Normalization and variance stabilization of single-cell rna-seq data using regularized negative binomial regression. *Genome Biology*, 20(1):296.

Haghverdi, L., Lun, A. T., Morgan, M. D., and Marioni, J. C. (2018). Batch effects in single-cell rna-sequencing data are corrected by matching mutual nearest neighbors. *Nature Biotechnology*, 36(5):421–427.

Hao, Y., Hao, S., Andersen-Nissen, E., Mauck III, W. M., Zheng, S., Butler, A., Lee, M. J., Wilk, A. J., Darby, C., Zager, M., et al. (2021). Integrated analysis of multimodal single-cell data. *Cell*.

Hao, Y., Stuart, T., Kowalski, M. H., Choudhary, S., Hoffman, P., Hartman, A., Srivastava, A., Molla, G., Madad, S., Fernandez-Granda, C., et al. (2024). Dictionary learning for integrative, multimodal and scalable single-cell analysis. *Nature Biotechnology*, 42(2):293–304.

Hartigan, J. A. and Hartigan, P. M. (1985). The dip test of unimodality. *The Annals of Statistics*, pages 70–84.

Haslbeck, J. M. and Wulff, D. U. (2020). Estimating the number of clusters via a corrected clustering instability. *Computational Statistics*, 35(4):1879–1894.

Hatfield, G. W., Hung, S.-p., and Baldi, P. (2003). Differential analysis of dna microarray gene expression data. *Molecular Microbiology*, 47(4):871–877.

Heller, M. J. (2002). Dna microarray technology: devices, systems, and applications. *Annual Review of Biomedical Engineering*, 4(1):129–153.

Hemerik, J. and Goeman, J. J. (2018). False discovery proportion estimation by permutations: confidence for significance analysis of microarrays. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(1):137–155.

Hivert, B., Agniel, D., Thiébaut, R., and Hejblum, B. P. (2024a). Post-clustering difference testing: valid inference and practical considerations with applications to ecological and biological data. *Computational Statistics & Data Analysis*, page 107916.

Hivert, B., Agniel, D., Thiébaut, R., and Hejblum, B. P. (2024b). Running in circles: is practical application feasible for data fission and data thinning in post-clustering differential analysis? *arXiv preprint arXiv:2405.13591*.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, pages 65–70.

Horn, R. A. and Johnson, C. R. (2012). *Matrix analysis*. Cambridge university press.

Hotelling, H. (1931). The generalization of student's ratio. *Annals of Mathematical Statistics*.

Hou, J., Aerts, J., Den Hamer, B., Van Ijcken, W., Den Bakker, M., Riegman, P., van der Leest, C., van der Spek, P., Foekens, J. A., Hoogsteden, H. C., et al. (2010). Gene expression-based classification of non-small cell lung carcinomas and survival prediction. *PloS one*, 5(4):e10312.

Hubert, L. and Arabie, P. (1985). Comparing partitions. *Journal of Classification*, 2(1):193–218.

Hwang, H., Dillon, W. R., and Takane, Y. (2006). An extension of multiple correspondence analysis for identifying heterogeneous subgroups of respondents. *Psychometrika*, 71(1):161–171.

Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS Medicine*, 2(8):e124.

Iodice D'Enza, A. and Palumbo, F. (2013). Iterative factor clustering of binary data. *Computational Statistics*, 28(2):789–807.

Jeffery, I. B., Higgins, D. G., and Culhane, A. C. (2006). Comparison and evaluation of methods for generating differentially expressed gene lists from microarray data. *BMC Bioinformatics*, 7:1–16.

Jollois, F.-X. and Nadif, M. (2009). Classification de données ordinales: modèles et algorithmes. In *41èmes Journées de Statistique, SFdS, Bordeaux*.

Jørgensen, B. and Song, P. X.-K. (1998). Stationary time series models with exponential dispersion model margins. *Journal of Applied Probability*, 35(1):78–92.

Kanehisa, M. and Goto, S. (2000). Kegg: kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, 28(1):27–30.

Kendall, M. G. (1938). A new measure of rank correlation. *Biometrika*, 30(1-2):81–93.

Koncevičius, K. (2023). *matrixTests: Fast Statistical Hypothesis Tests on Rows and Columns of Matrices*. R package version 0.2.3.

Korn, E. L., Troendle, J. F., McShane, L. M., and Simon, R. (2004). Controlling the number of false discoveries: application to high-dimensional genomic data. *Journal of Statistical Planning and Inference*, 124(2):379–398.

Korthauer, K. D., Chu, L.-F., Newton, M. A., Li, Y., Thomson, J., Stewart, R., and Kendziorski, C. (2016). A statistical approach for identifying differential distributions in single-cell rna-seq experiments. *Genome Biology*, 17(1):1–15.

Kriegeskorte, N., Simmons, W. K., Bellgowan, P. S., and Baker, C. I. (2009). Circular analysis in systems neuroscience: the dangers of double dipping. *Nature Neuroscience*, 12(5):535–540.

Lähnemann, D., Köster, J., Szczurek, E., McCarthy, D. J., Hicks, S. C., Robinson, M. D., Vallejos, C. A., Campbell, K. R., Beerenwinkel, N., Mahfouz, A., et al. (2020). Eleven grand challenges in single-cell data science. *Genome Biology*, 21(1):1–35.

Lance, G. N. and Williams, W. T. (1967). A general theory of classificatory sorting strategies: 1. hierarchical systems. *The Computer Journal*, 9(4):373–380.

Law, C. W., Chen, Y., Shi, W., and Smyth, G. K. (2014). voom: Precision weights unlock linear model analysis tools for rna-seq read counts. *Genome Biology*, 15:1–17.

Lee, J. D., Sun, D. L., Sun, Y., and Taylor, J. E. (2016). Exact post-selection inference, with application to the lasso. *The Annals of Statistics*, 44(3).

Leiner, J., Duan, B., Wasserman, L., and Ramdas, A. (2023). Data fission: splitting a single data point. *Journal of the American Statistical Association*, pages 1–12.

Loftus, J. R. and Taylor, J. E. (2015). Selective inference in regression models with groups of variables. *arXiv preprint arXiv:1511.01478*.

Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome Biology*, 15(12):1–21.

Lowe, R., Shirley, N., Bleackley, M., Dolan, S., and Shafee, T. (2017). Transcriptomics technologies. *PLoS Computational Biology*, 13(5):e1005457.

Lubke, G. and Neale, M. (2008). Distinguishing between latent classes and continuous factors with categorical outcomes: Class invariance of parameters of factor mixture models. *Multivariate Behavioral Research*, 43(4):592–620.

Luksa, M. (2017). *Kubernetes in Action*. Simon and Schuster.

Lun, A. T., Bach, K., and Marioni, J. C. (2016). Pooling across cells to normalize single-cell rna sequencing data with many zero counts. *Genome Biology*, 17(1):75.

Maaten, L. v. d. and Hinton, G. (2008). Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(Nov):2579–2605.

Macqueen, J. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability/University of California Press*.

Maigné, É., Sanchez, I., Carayon, D., Tran, J., Rey, J.-F., Midoux, C., and Marjou, M. (2023). Sk8: Un service institutionnel de gestion et d'hébergement d'applications shiny. In *Rencontres R 2023 et 2024*.

Mann, H. B. and Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics*, pages 50–60.

Marcus, R., Eric, P., and Gabriel, K. R. (1976). On closed testing procedures with special reference to ordered analysis of variance. *Biometrika*, 63(3):655–660.

Marden, J. I. (1996). *Analyzing and Modeling Rank Data*. CRC Press.

McLachlan, G. (2000). Finite mixture models. *A Wiley-interscience Publication*.

Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by rna-seq. *Nature Methods*, 5(7):621–628.

Nabavi, S., Schmolze, D., Maitituoheti, M., Malladi, S., and Beck, A. H. (2016). Emdomics: a robust and powerful method for the identification of genes differentially expressed between heterogeneous classes. *Bioinformatics*, 32(4):533–541.

Neufeld, A., Dharamshi, A., Gao, L. L., and Witten, D. (2024a). Data thinning for convolution-closed distributions. *Journal of Machine Learning Research*, 25(57):1–35.

Neufeld, A., Gao, L. L., Popp, J., Battle, A., and Witten, D. (2024b). Inference after latent variable estimation for single-cell rna sequencing data. *Biostatistics*, 25(1):270–287.

Neufeld, A., Popp, J., Gao, L. L., Battle, A., and Witten, D. (2023). Negative binomial count splitting for single-cell rna sequencing data. *arXiv preprint arXiv:2307.12985*.

Neuvial, P. (2008). Asymptotic properties of false discovery rate controlling procedures under independence. *Electronic Journal of Statistics*, 2:1065–1110. With corrigendum in EJS 2009(3):1083.

Neuvial, P. (2020). *Contributions to Statistical Inference from Genomic Data*. Habilitation thesis, Université Toulouse III. Available from https://tel.archives-ouvertes.fr/tel-02969229.

Neuvial, P., Blanchard, G., Durand, G., Enjalbert-Courrech, N., and Roquain, E. (2024). *sanssouci: Post Hoc Multiple Testing Inference*. R package version 0.13.0.

Neuvial, P. and Enjalbert-Courrech, N. (2024). *sansSouci.data: Companion Data Package For The sansSouci Package*. R package version 0.6.0.

Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572.

Pelckmans, K., De Brabanter, J., Suykens, J. A., and De Moor, B. (2005). Convex clustering shrinkage. *PASCAL Workshop on Statistics and Optimization of Clustering Workshop*.

Qiu, P. (2020). Embracing the dropouts in single-cell rna-seq analysis. *Nature Communications*, 11(1):1169.

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., and Smyth, G. K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, 43(7):e47.

Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2010). edger: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140.

Romano, J. P. and Wolf, M. (2005). Exact and approximate stepdown methods for multiple hypothesis testing. *Journal of the American Statistical Association*, 100(469):94–108.

Romano, J. P. and Wolf, M. (2007). Control of generalized error rates in multiple testing. *The Annals of Statistics*, 35(4):1378–1408.

Rosenblatt, J. D., Finos, L., Weeda, W. D., Solari, A., and Goeman, J. J. (2018). All-resolutions inference for brain imaging. *Neuroimage*, 181:786–796.

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65.

Sarkar, S. K. et al. (2008). On the simes inequality and its generalization. In *Beyond parametrics in interdisciplinary research: Festschrift in honor of Professor Pranab K. Sen*, pages 231–242. Institute of Mathematical Statistics.

Schwarz, G. (1978). Estimating the dimension of a model. *The annals of statistics*, pages 461–464.

Selosse, M., Jacques, J., and Biernacki, C. (2018). ordinalclust: an r package for analyzing ordinal data. *HAL*, 2018.

Sievert, C. (2020). *Interactive Web-Based Data Visualization with R, plotly, and shiny*. Chapman and Hall/CRC.

Simes, R. J. (1986). An improved Bonferroni procedure for multiple tests of significance. *Biometrika*, 73(3):751–754.

Smyth, G. K. (2004). Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology*, 3(1).

Soneson, C. and Delorenzi, M. (2013). A comparison of methods for differential expression analysis of rna-seq data. *BMC Bioinformatics*, 14:1–18.

Soneson, C. and Robinson, M. (2018). Bias, robustness and scability in single-cell differential expression analysis. *Nature Methods*, 15:255–261.

Squair, J. W., Gautier, M., Kathe, C., Anderson, M. A., James, N. D., Hutson, T. H., Hudelle, R., Qaiser, T., Matson, K. J., Barraud, Q., et al. (2021). Confronting false discoveries in single-cell differential expression. *Nature Communications*, 12(1):5692.

Steinhaus, H. et al. (1956). Sur la division des corps matériels en parties. *Bull. Acad. Polon. Sci*, 1(804):801.

Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck III, W. M., Hao, Y., Stoeckius, M., Smibert, P., and Satija, R. (2019). Comprehensive integration of single-cell data. *Cell*, 177(7):1888–1902.

Student (1908). The probable error of a mean. *Biometrika*, pages 1–25.

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., et al. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102(43):15545–15550.

TGCA, Cancer Genome Atlas Research Network, et al. (2014). Comprehensive molecular characterization of urothelial bladder carcinoma. *Nature*, 507(7492):315.

Thorndike, R. L. (1953). Who belongs in the family? *Psychometrika*, 18(4):267–276.

Thorsson, V., Gibbs, D. L., Brown, S. D., Wolf, D., Bortone, D. S., Yang, T.-H. O., Porta-Pardo, E., Gao, G. F., Plaisier, C. L., Eddy, J. A., et al. (2018). The immune landscape of cancer. *Immunity*, 48(4):812–830.

Tiberi, S., Crowell, H. L., Samartsidis, P., Weber, L. M., and Robinson, M. D. (2023). distinct: a novel approach to differential distribution analyses. *The Annals of Applied Statistics*, 17(2):1681–1700.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288.

Trapnell, C., Hendrickson, D. G., Sauvageau, M., Goff, L., Rinn, J. L., and Pachter, L. (2013). Differential analysis of gene regulation at transcript resolution with rna-seq. *Nature biotechnology*, 31(1):46–53.

Tukey, J. W. (1953). The problem of multiple comparisons. *Multiple Comparisons*.

Van Buuren, S. and Heiser, W. J. (1989). Clusteringn objects intok groups under optimal scaling of variables. *Psychometrika*, 54(4):699–706.

Vesely, A., Finos, L., and Goeman, J. J. (2023). Permutation-based true discovery guarantee by sum tests. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 85(3):664–683.

Vovk, V. and Wang, R. (2020). Combining p-values via averaging. *Biometrika*, 107(4):791–808.

Walesiak, M. (1999). Distance measure for ordinal data. *Argumenta Oeconomica*.

Wang, T. and Nabavi, S. (2018). Sigemd: A powerful method for differential gene expression analysis in single-cell rna sequencing data. *Methods*, 145:25–32.

Ward Jr, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, 58(301):236–244.

Wei, G. C. and Tanner, M. A. (1990). A monte carlo implementation of the em algorithm and the poor man's data augmentation algorithms. *Journal of the American Statistical Association*, 85(411):699–704.

Welch, B. L. (1947). The generalization of 'student's'problem when several different population varlances are involved. *Biometrika*, 34(1-2):28–35.

Westfall, P. H. and Young, S. S. (1993). *Resampling-based multiple testing: Examples and methods for p-value adjustment*, volume 279. John Wiley & Sons.

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83.

Yang, F., Foygel Barber, R., Jain, P., and Lafferty, J. (2016). Selective inference for group-sparse linear models. *Advances in Neural Information Processing Systems*, 29.

Yekutieli, D. (2008). Hierarchical false discovery rate–controlling methodology. *Journal of the American Statistical Association*, 103(481):309–316.

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

Zhang, J. M., Kamath, G. M., and David, N. T. (2019). Valid post-clustering differential analysis for single-cell rna-seq. *Cell Systems*, 9(4):383–392.

RÉPUBLIQUE FRANÇAISE
Liberté
Égalité
Fraternité

Université de Toulouse

Titre : Inférence post-selection pour l'analyse des données transcriptomiques

Mots clés : Tests d'hypothèse, Classification non supervisée, Inférence post hoc, Inférence post-clustering, Analyse de données transcriptomiques

Résumé : Dans le domaine de la transcriptomique, les avancées technologiques, telles que les puces à ADN et le séquençage à haut-débit, ont permis de quantifier l'expression génique à grande échelle. Ces progrès ont soulevé des défis statistiques, notamment pour l'analyse d'expression différentielle, visant à identifier les gènes différenciant significativement deux populations.

Cependant, les procédures classiques d'inférence perdent leurs garanties de contrôle du taux de faux positifs lorsque les biologistes sélectionnent un sous-ensemble de gènes. Les méthodes d'inférence post hoc surmontent cette limitation en garantissant un contrôle sur le nombre de faux positifs, même pour des ensembles de gènes sélectionnés de manière arbitraire. La première contribution de ce manuscrit démontre l'efficacité de ces méthodes pour les données transcriptomiques de deux conditions biologiques, notamment grâce à l'introduction d'un algorithme de calcul des bornes post hoc à complexité linéaire, adapté à la grande dimension des données. Une application interactive a également été développée, facilitant la sélection et l'évaluation simultanée des bornes post hoc pour des ensembles de gènes d'intérêt. Ces contributions sont présentées dans la première partie du manuscrit.

L'évolution technologique vers le séquençage en cellule unique a soulevé de nouvelles questions, notamment l'identification des gènes dont l'expression se distingue d'un groupe cellulaire à un (des) autre(s). Cette problématique est complexe car les groupes cellulaires doivent d'abord être estimés par une méthode de clustering, avant d'effectuer un test comparatif, menant ainsi à une analyse circulaire. Dans la seconde partie de ce manuscrit, nous présentons une revue des méthodes d'inférence post-clustering résolvant ce problème ainsi qu'une comparaison numérique des approches multivariées et marginales de comparaison de classes. Enfin, nous explorons comment l'utilisation des modèles de mélange dans l'étape de clustering peut être exploitée dans les tests post-clustering, et nous discutons de perspectives pour l'application de ces tests aux données transcriptomiques.

Title: Post-selection inference for transcriptomic data analysis

Key words: Hypothesis testing, Clustering, Post hoc inference, Post-clustering inference, Transcriptomic data analysis

Abstract: In the field of transcriptomics, technological advances, such as microarrays and high-throughput sequencing, have enabled large-scale quantification of gene expression. These advances have raised statistical challenges, particularly in differential expression analysis, which aims to identify genes that significantly differentiate between two populations.

However, traditional inference procedures lose their ability to control the false positive rate when biologists select a subset of genes. Post-hoc inference methods address this limitation by providing control over the number of false positives, even for arbitrary gene sets. The first contribution of this manuscript demonstrates the effectiveness of these methods for the differential analysis of transcriptomic data between two biological conditions, notably through the introduction of a linear-time algorithm for computing post-hoc bounds, adapted to the high dimensionality of the data. An interactive application was also developed to facilitate the selection and simultaneous evaluation of post-hoc bounds for sets of genes of interest. These contributions are presented in the first part of the manuscript.

The technological evolution towards single-cell sequencing has raised new questions, particularly regarding the identification of genes whose expression distinguishes one cellular group from another. This issue is complex because cell groups must first be estimated using clustering method before performing a comparative test, leading to a circular analysis. In the second part of this manuscript, we present a review of post-clustering inference methods addressing this problem, as well as a numerical comparison of multivariate and marginal approaches for cluster comparison. Finally, we explore how the use of mixture models in the clustering step can be exploited in post-clustering tests, and discuss perspectives for applying these tests to transcriptomic data.