Clustering et analyse différentielle de données d'expression génique

Benjamin Hivert

université de BORDEAUX

THÈSE PRESENTÉE
POUR OBTENIR LE GRADE DE

DOCTEUR DE
L'UNIVERSITÉ DE BORDEAUX

ÉCOLE DOCTORALE SOCIÉTÉS, POLITIQUE, SANTÉ PUBLIQUE
SPÉCIALITÉ SANTÉ PUBLIQUE, OPTION BIOSTATISTIQUE

Par Benjamin HIVERT

CLUSTERING ET ANALYSE DIFFÉRENTIELLE DE DONNÉES D'EXPRESSION GÉNIQUE

Sous la direction de : Rodolphe THIÉBAUT
Co-directeur : Boris HEJBLUM

Soutenue le 24 Septembre 2024

Membres du jury :

Mme Proust-Lima Cécile	Directrice de Recherche	INSERM Bordeaux	Présidente
Mme Maugis-Rabusseau Cathy	Professeure des Universités	INSA Toulouse	Rapporteure
M.Picard Franck	Directeur de Recherche	CNRS Lyon	Rapporteur
M.Neuvial Pierre	Directeur de Recherche	CNRS Toulouse	Examinateur
M.Thiébaut Rodolphe	PU-PH	Université de Bordeaux	Directeur de thèse
M.Hejblum Boris	Chargé de Recherche	INSERM Bordeaux	Co-directeur de thèse

Remerciements

Pour clore ces quatre années de thèse, qui ont été autant exceptionnelles que compliquées, je tiens à remercier ceux sans qui tout cela n'aurait pas été possible, mes deux directeurs de thèse **Rodolphe Thiébaut** et **Boris Hejblum**. **Rodolphe**, j'ai eu beaucoup de chance de pouvoir travailler et apprendre à tes côtés. Ta passion et ton enthousiasmes pour la science ont été de véritables moteurs pour le grand pessimiste que je suis. Tu as toujours su trouver du temps pour m'écouter et me conseiller quand j'étais perdu tant dans mes recherches, que sur mon avenir. Je te remercie également pour le climat de bienveillance que tu as su instaurer dans l'équipe SISTM. Grâce à toi, j'ai non seulement progressé en tant que jeune chercheur, mais aussi trouvé un environnement de travail dans lequel je me suis épanoui, et pour tout cela, je te suis grandement reconnaissant.

Boris, sans toi, je serais probablement encore en train de douter de ma capacité à réaliser une thèse. Aujourd'hui, j'écris ces quelques mots pour te remercier, car grâce à toi, non seulement j'ai découvert que j'en étais capable, mais en plus, je l'ai fait. Alors, merci pour la confiance que tu m'as accordée, depuis mon stage de M2 jusqu'à ce jour. Tu as fait preuve d'une immense patience avec moi et tu as toujours su me rassurer lors des (nombreux) moments où tout semblait aller de travers et où je pensais que tout était perdu. Malgré ces périodes de doute, ta bienveillance et ton optimisme m'ont permis de ne retenir que le meilleur de cette aventure. J'ai énormément appris grâce à toi, non seulement sur la science et le monde de la recherche académique, qui m'était inconnu, mais aussi sur moi-même. Tu m'as offert des opportunités exceptionnelles, et même si, parfois, je les voyais comme des cadeaux empoisonnés, elles m'ont forcé à me dépasser et m'ont fait grandir. Pour tout cela, je te suis infiniment reconnaissant. Cette thèse est autant le fruit de mon travail que de ton soutien, et je n'oublierai jamais l'impact que tu as eu sur mon parcours.

Je tiens également à exprimer ma profonde gratitude aux membres du jury pour leur présence et leur engagement. **Cécile**, après avoir suivi mon évolution tout au long de ma thèse lors des comités de suivi, je te remercie sincèrement d'avoir accepté de présider ce jury de thèse. Ta bienveillance et ta pédagogie constantes ont été d'une grande importance pour moi, et je suis honoré de t'avoir à mes côtés pour cette étape cruciale de mon parcours. **Franck** et **Cathy**, je vous remercie d'avoir accepté d'évaluer mon travail en tant que rapporteurs de ma thèse. Votre expertise et vos retours constructifs ont été précieux pour l'amélioration de ce travail. **Pierre**, toi qui connais aussi bien que nous les défis (et les galères) de l'inférence post-clustering, merci d'avoir accepté d'évaluer mon travail de thèse.

Denis, I would like to extend you my heartfelt thanks for your invaluable help throughout my thesis. Despite the $6,883$ kilometers ($4,273$ miles) between us, you were always available for me. I also want to thank you for being an amazing culinary guide during my trips to the United States. To **Layla**, thank you for your warm welcome in Texas and for the unforgettable stay you provided us, both scientifically and personally.

À l'ensemble des personnes qui constituent les rouages de la grande roue de l'**équipe SISTM**, merci pour votre bienveillance. C'est un réel bonheur de travailler au sein d'une telle équipe. Un immense merci tout particulier à **Sandrine** pour son efficacité, sa gentillesse et sa bonne humeur. C'est toujours un plaisir de passer te voir, ainsi que **Nadine**.

Cette thèse aurait été bien plus (administrativement) compliquée et surtout bien moins joyeuse sans toi.

Robin, j'ai eu la chance de découvrir l'enseignement à l'ISPED grâce à toi, et quel plaisir ce fut de le faire à tes côtés. J'ai appris auprès de l'un des meilleurs enseignants que j'ai rencontré. **Pierre**, merci d'être notre chef MCE : tu as toujours su faire en sorte que tout le monde soit satisfait et que cette expérience d'enseignement se déroule dans les meilleures conditions. **Anaïs**, je te remercie pour ta bonne humeur contagieuse, tes conseils précieux, et tous ces bons repas partagés au RU. **Antoine**, tu m'accompagnes depuis le M1, c'est grâce à toi que j'ai découvert le BPH lors de mon stage de M2, et je te suis infiniment reconnaissant pour tout cela.

Marine, tu as été mon enseignante, puis ma co-encadrante de stage, et nous avons fini doctorant ensemble au sein de l'équipe SISTM. Tu as été une source précieuse de conseils qui résonnent toujours dans un petit coin de ma tête, encore aujourd'hui. Je te remercie de m'avoir transmis tes clés pour affronter le doctorat.

À mes collègues du **Bureau N311 (ex John Snow)**, à qui j'ai souvent imposé mes karaokés-gourdes, je tiens à tous vous remercier pour la bonne ambiance qui règne dans notre bureau. Merci à **Quentin** de ramener un peu de Paris à Bordeaux (et de Bordeaux à Paris) ; à **Marie** d'avoir été un si bel exemple ; à **Hélène** (ma co-thésarde préférée) pour tous ces moments de partages sur nos aventures de thèse, tu m'as beaucoup appris et ce n'est que le début ; à **Clément** ; à **Mélanie** d'avoir été là après le Covid quand tout était encore compliqué et d'être toujours là pour moi trois ans plus tard ; à **Mathieu** pour ses visites surprises depuis Boston ; à **Kalidou** ; à **Anto** pour ses talents de dessinateur ; à **Arthur** pour notre séjour texan inoubliable ; à **Auriane** d'avoir apporté un grain de folie dans le bureau ; à **Sara** d'avoir toujours su m'écouter et me rebooster quand j'étais "blop" ; à **Quentin (Q2)** de supporter avec moi l'UBB ; et à tous ceux qui sont arrivés plus récemment **Aurore**, **Nico** et **Annesh**. Je souhaite aussi remercier tous les stagiaires qui ont travaillé avec nous et en particulier **Maud** et **Mathéo** avec lesquels j'ai eu la chance de beaucoup plus échanger. Un immense merci à tous les collègues du troisième étage de manière générale, et tout particulièrement à **Ilana** et **Blandine**, toujours souriantes et de bonne humeur au détour d'un couloir. Enfin, un grand merci aux collègues du **Bureau N101**, dans lequel je me suis souvent réfugié. Merci à **Coco** pour partager sa passion pour le pain, le chou et le sel à chaque repas, à **Justine** pour avoir toujours la réponse aux questions que tout le monde se pose, et à **Marius** pour avoir été toujours présent, aussi bien dans les bons moments que dans les mauvais.

Au cours de ces quatre années, j'ai eu la chance de rencontrer des personnes avec lesquelles j'ai pu forger de vraies amitiés. **Mélanie**, **Sara**, **Riri**, **Marius**, plus que des "amigollègues", je suis vraiment fier de pouvoir vous compter parmi mes amis aujourd'hui. À la grande famille biostat élargie : **Capu**, **Hugo**, **Matthias**, **Arnaud**, **Adrien**, **Marie**, **Caro** et **Momo**, je suis vraiment heureux que vous soyez entrés dans ma vie (merci au cours de R).

Je tenais également à remercier mes amis de longue date sans qui rien de tout cela n'aurait été possible : **Charlotte**, mon binôme de master et de tous les jours maintenant, qui connaît cette thèse aussi bien que moi à force de me relire et de me faire répéter ; **Flav**, mon *bad-boy* préféré ; **Baptiste** de la célèbre team des "Superquantileurs" et **Laura** mon petit rayon de soleil de Cavaillon. Un grand merci également à mes amis qui sont là depuis le lycée. **Val**, qui a partagé cette incroyable aventure qu'est la thèse avec moi, même depuis

un labo différent. **Manon**, merci d'être la belle personne que tu es et d'avoir toujours été présente pour moi, peu importe la distance qui nous a séparé. **Nana**, je ne saurais jamais assez te remercier pour tout ce que tu as fait pour moi, tu le sais déjà mais je suis fier de pouvoir le graver officiellement dans ce manuscrit.

Je conclurai ces remerciements par ceux grâce à qui tout est possible : ma famille. À **mes parents**, je suis profondément reconnaissant pour les valeurs que vous m'avez inculquées, valeurs qui ont guidé chacune de mes étapes dans cette thèse et qui continueront à éclairer ma vie. Merci d'avoir toujours cru en moi et de m'avoir donné la liberté d'être pleinement qui je veux être. À **mon frère**, tu es mon héros au quotidien, et à ma **belle-sœur**, devenue ma sœur de cœur, merci d'être des piliers solides dans ma vie. Enfin, à **Simon** et **Louise**, vous êtes ma source de bonheur quotidienne, irradiant ma vie de votre amour et de votre joie de vivre. J'espère qu'un jour, vous serez aussi fiers de moi que je le suis déjà de vous.

Production scientifique & Enseignement

Publications scientifiques

► Lévy, Y., Wiedemann, A., Hejblum, B.P, Durand, M., Lefebvre, C., Surénaud, M., Lacabaratz, C., Perreau, M., Foucat, E. Déchenaud, M., Tisserand, P, Blengio, F., Hivert, B., Gauthier, M., Cervantes-Gonzalez, M., Bachelet, D., Laouénan, C.,Bouadma, L., Timsit, JF., Yazdanpana, Y., Pantaleo, G., Hocini, H., Thiébaut.R. (2021). CD177, a specific marker of neutrophil activation, is associated with coronavirus disease 2019 severity and death. Iscience, 24(7).
DOI: 10.1016/ j.isci.2021.102711

► Hivert, B., Agniel, D., Thiébaut, R., & Hejblum, B. P. (2024). Post-clustering difference testing : Valid inference and practical considerations with applications to ecological and biological data. Computational Statistics & Data Analysis, 193, 107916
DOI : 0.1016/j.csda.2023.107916

► Hivert, B., Agniel, D., Thiébaut, R., & Hejblum, B. P. (2024). Running in circles : practical limitations for real-life application of data fission and data thinning in post-clustering differential analysis. arXiv preprint, arXiv:2405.13591

Communications

Communications orales en conférences

► Hivert, B., Agniel, D., Thiébaut, R., & Hejblum, B. P. Inférence post-clustering pour l'identification des variables responsables de la séparation de paires de clusters, *Journées nationales « Statistiques & Santé »*, Octobre 2021, France (À distance)

► Hivert, B., Agniel, D., Thiébaut, R., & Hejblum, B. P. Méthodes pour l'inférence post-clustering appliquées à l'expression génique, *Journées de la Statistiques de la SFDS (JDS 2022)*, Juin 2022, Lyon, France

► Hivert, B., Agniel, D., Thiébaut, R., & Hejblum, B. P. Post-clustering differential expression analysis : valid inference and practical considerations, *31st International Biometric Conference (IBC2022)*, Juillet 2022, Riga, Lettonie.

► Hivert, B., Agniel, D., Thiébaut, R., & Hejblum, B. P. Post-clustering Differential Testing : valid inference and practical considerations, *Statistical Methods for Post Genomic Data (SMPGD 2023)*, Février 2023, Gand, Belgique.

▶ Hivert, B., Agniel, D., Thiébaut, R., & Hejblum, B. P. Apports et Challenges de la Fission de Données pour les problèmes d'inférences post-clustering, *Journées de la Statistiques de la SFDS (JDS 2023)*, Juillet 2023, Bruxelles, Belgique

▶ Hivert, B., Agniel, D., Thiébaut, R., & Hejblum, B. P. Data fission for post-clustering differential analysis using dearseq, *EuroBioc2023*, Septembre 2023, Gand, Belgique.

▶ Hivert, B., Agniel, D., Thiébaut, R., & Hejblum, B. P. Fission de données pour l'inférence post-classification : de la théorie à la pratique, *Journées de la Statistiques de la SFDS (JDS 2024)*, Mai 2024, Bordeaux, France

Communications écrites (posters) en conférences

▶ Hivert, B., Agniel, D., Thiébaut, R., & Hejblum, B. P. Contributions and challenges of data fission for post-clustering differential analysis, *44th Annual Conference of the International Society for Clinical Biostatistics (ISCB 2023)*, Aout 2023, Milan, Italie

Paquet 

▶ `VALIDICLUST` : VALID Inference for Clusters Separation Testing. Disponible sur le CRAN

Enseignements

▶ Master 1 Santé Publique :
  ▷ 2021-2022 : STAT103 : Outils de simulation en biostatistique, TD, 10h
  ▷ 2022-2023 : EPISTA101 : Approche quantitative en santé publique, TD, 2h

▶ Master 2 Biostatistique :
  ▷ 2021-2022 : STAT104 : Bases en biostatistique et plan d'expérience, TD, 17h
  ▷ 2022-2023 : STAT104 : Bases en biostatistique et plan d'expérience, TD, 18h

▶ 2ème Année de Médecine :
  ▷ 2021-2022 : UE Informatique, TD, 4h
  ▷ 2022-2023 : UE Informatique, TD, 12h

▶ DU Méthodes Statistiques en Santé - via internet :
  ▷ 2020-2021 : Module introduction à R, 18h
  ▷ 2021-2022 : Module introduction à R, 18h

Table des matières

Production scientifique & Enseignement **7**

1 Introduction **13**
1.1 Données d'expression génique . . . 13
1.1.1 L'expression génique et sa mesure . . . 14
1.1.2 Les données d'expression génique . . . 16
1.1.3 L'exemple de la COVID-19 . . . 18
1.1.4 Statistiques en grande dimension . . . 20
1.2 Méthodes pour le clustering & l'analyse différentielle . . . 23
1.2.1 Outils pour le clustering . . . 23
1.2.2 L'analyse différentielle de l'expression génique . . . 32
1.2.3 Clustering & analyse différentielle dans l'analyse de données scRNA-seq . . . 37
1.2.4 Le défi de l'inférence post-clustering . . . 38
1.3 Organisation de la thèse . . . 42

2 Centrer-réduire les données RNA-seq ou non en amont du clustering ? **43**
2.1 Introduction . . . 43
2.2 Standardiser les données pour le clustering ? . . . 46
2.2.1 Cadre général & Définition . . . 46
2.2.2 La standardisation dans le cadre des données RNA-seq . . . 48
2.3 Méthodes . . . 50
2.3.1 Étude de simulations . . . 50
2.3.2 Analyse de données réelles . . . 51
2.3.3 Critère d'évaluation . . . 52
2.4 Résultats . . . 53
2.4.1 Étude de simulations . . . 53
2.4.2 Analyse de données réelles . . . 55
2.5 Discussion . . . 56

3 Etat de l'art des méthodes d'inférence post-clustering **59**
3.1 Introduction . . . 59
3.2 Méthodes basées sur un conditionnement . . . 60
3.2.1 Lemme polyédrique . . . 60
3.2.2 Méthodes d'inférence post-clustering . . . 61
3.3 Méthodes basées sur une décomposition explicite de l'information . . . 65
3.3.1 Découpage de l'échantillon . . . 65
3.3.2 Découpage d'une observation . . . 66

4 Tests de différences post-clustering : inférence valide et considérations pratiques avec des applications aux données écologiques et biologiques 71
4.1 Introduction . . . 71
4.2 Méthodes . . . 73
4.2.1 Test d'inférence sélective . . . 74
4.2.2 Un test plus puissant en présence de clusters intercalés . . . 77
4.2.3 Test de multimodalité . . . 79
4.3 Résultats . . . 80
4.3.1 Étude de simulations . . . 80
4.3.2 Différences morphologiques chez les manchots de l'archipel de Palmer 87
4.3.3 Clustering de cellules immunitaires issues de mesures de cytométrie en flux . . . 91
4.4 Discussion . . . 92

5 Limites pratiques des approches basées sur une décomposition de l'information pour répondre aux problèmes d'inférence post-clustering 95
5.1 Introduction . . . 95
5.2 Méthodes . . . 97
5.2.1 Rappels : fission de données et dilution de données . . . 97
5.2.2 Limites de l'application pratique de la fission et de la dilution de données . . . 99
5.2.3 Connaissance a priori et estimation du paramètre d'échelle . . . 101
5.2.4 Solutions pratiques . . . 103
5.3 Résultats . . . 105
5.3.1 Définition de l'erreur de type I comme une fonction du biais dans l'estimation de la variance . . . 105
5.3.2 Performances de l'estimateur de variance locale . . . 111
5.3.3 Application à l'analyse de données scRNA-seq . . . 113
5.4 Discussion . . . 115

6 Conclusion et perspectives 119

Bibliographie 125

A Matériels Supplémentaires associés au Chapitre 4 141
A.1 Détails sur le caclul de la p-valeur du test d'inférence sélective . . . 141
A.2 Figure Supplémentaire 1 . . . 143
A.3 Figure Supplémentaire 2 . . . 144
A.4 Figure Supplémentaire 3 . . . 145
A.5 Figure Supplémentaire 4 . . . 146
A.6 Figure Supplémentaire 5 . . . 147
A.7 Figure Supplémentaire 6 . . . 149
A.8 Tableau Supplémentaire 1 . . . 150
A.9 Tableau Supplémentaire 2 . . . 151
A.10 Figure Supplémentaire 7 . . . 152
A.11 Tableau Supplémentaire 3 . . . 153

A.12 Figure Supplémentaire 8 154
A.13 Comportement des tests proposés dans le cadre multivarié 155
A.13.1 Les corrélations entre les variables peuvent empêcher un bon contrôle de l'erreur de type I 155
A.13.2 Le nombre de variables a un impact relativement faible sur les performances des tests 156

Chapitre 1

Introduction

Contenu

1.1 Données d'expression génique 13
1.1.1 L'expression génique et sa mesure 14
1.1.2 Les données d'expression génique 16
1.1.3 L'exemple de la COVID-19 18
1.1.4 Statistiques en grande dimension 20
1.2 Méthodes pour le clustering & l'analyse différentielle 23
1.2.1 Outils pour le clustering 23
1.2.2 L'analyse différentielle de l'expression génique 32
1.2.3 Clustering & analyse différentielle dans l'analyse de données scRNA-seq 37
1.2.4 Le défi de l'inférence post-clustering 38
1.3 Organisation de la thèse 42

1.1 Données d'expression génique

L'expression génique, processus fondamental de la biologie moléculaire, révèle comment l'information génétique héritée, stockée dans l'ADN (acide désoxyribonucléique), orchestre la synthèse des protéines, acteurs clés de la vie cellulaire et des organismes. Étant maintenant mesurable, ces nouvelles données permettent alors de caractériser les mécanismes sous-jacents à de nombreuses maladies (Sokal et al. 2021, Kotliar et al. 2020) et constituent un outil précieux dans la conception et l'évaluation des vaccins (Rechtien et al. 2017, Liu et al. 2021, O'Connor & Pollard 2013). Longtemps coûteuse et complexe à mesurer, l'expression génique est maintenant mesurable à partir de prélèvements capillaires, comparables à la mesure de la glycémie pour le diabète (Obermoser et al. 2013, Rinchai et al. 2022). Il devient possible d'étudier les réponses biologiques à diverses infections ou vaccins tout en garantissant une fiabilité des mesures par rapport aux méthodes traditionnelles plus invasives, basées sur des prises de sang (Reust et al. 2018). L'utilisation des données d'expression génique offre alors des perspectives prometteuses

d'une médecine plus personnalisée (Cotugno et al. 2019), et c'est pour cette raison que ces données et leur analyse font l'objet de recherches actives.

1.1.1 L'expression génique et sa mesure

Le génome d'un organisme, contenu à l'interieur du noyau des cellules, correspond à l'ensemble de ses gènes. Celui de l'Homme, dont le séquençage a été initié à la fin de années 1980 serait composé d'environ 30 000 gènes (Lander et al. 2001, Venter et al. 2001). Ces unités biologiques héréditaires sont des segments d'ADN, une molécule en double hélice contenant l'information génétique. Cette information est stockée sous forme de séquences de nucléotidiques représentées par les lettres A (adénine), T (thymine), C (cytosine) et G (guanine). Ces séquences déterminent les instructions nécessaires à la synthèse des protéines.

L'expression génique désigne alors l'ensemble des processus qui conduisent à la synthèse d'une protéine à partir d'un gène. Elle repose sur deux étapes fondamentales : la transcription et la traduction, qui sont souvent simplifiées au travers du dogme central de la biologie moléculaire. La transcription consiste en la synthèse d'une molécule d'ARN (acide ribonucléique) messager (ARNm) à partir d'un brin d'ADN dans le noyau cellulaire. Cette étape donne son nom au transcriptome, qui, à la manière du génome pour les gènes, désigne la totalité des ARNm produits dans une cellule, un tissu ou un organisme à un instant précis. L'ARNm, contrairement à l'ADN, est une molécule en simple brin, complémentaire à l'ADN, où la thymine (T) est remplacée par l'uracile (U). Après sa formation, l'ARNm quitte le noyau et migre vers le cytoplasme de la cellule, où se déroule la traduction. Au cours de cette étape, les ribosomes lisent l'ARNm et utilisent cette information pour agencer les acides aminés dans le bon ordre, formant ainsi une protéine spécifique. Les protéines ainsi produites remplissent une multitude de fonctions dans l'organisme : elles jouent un rôle crucial dans la structure, la fonction et la régulation des cellules et des tissus. Elles ont une place centrale dans le système immunitaire, où certaines d'entre elles, appelées anticorps, agissent comme des défenseurs contre les infections et les maladies en reconnaissant, neutralisant et détruisant les agents pathogènes. On dit qu'un gène s'exprime lorsqu'il est transcrit pour produire une protéine spécifique et cette expression est régulée dans les cellules de manière à répondre aux besoins de l'organisme.

Mesurer l'expression génique est possible grâce au séquençage de l'ARN messager (RNA-seq) développé dans les années 2000 (Nagalakshmi et al. 2008, Weber 2015). Cette technologie permet de quantifier l'abondance des transcrits d'ARNm, c'est-à-dire de déterminer le nombre de copies de chaque ARNm spécifique dans un échantillon biologique. Contrairement au séquençage du génome, qui ne fournit que la cartographie des gènes sans indication de leur activité, le RNA-seq offre une image dynamique de l'expression génique en mesurant l'abondance des d'ARNm, directement liée à l'activité des gènes. En

effet, plus un gène est activement transcrit, plus la quantité d'ARNm correspondant sera élevée dans l'échantillon. L'abondance de ces ARNm donne donc une indication directe sur le niveau d'expression d'un gène, reflétant les besoins en protéines spécifiques codées par ces mêmes gènes, qui varient selon les processus biologiques en cours dans l'organisme.

La mesure de l'expression génique avec le RNA-seq repose sur plusieurs étapes successives. Pour commencer, les ARNm sont extraits de l'échantillon biologique, puis ils sont rétro-transcrits en ADN complémentaire (ADNc) à l'aide d'une enzyme appelée transcriptase inverse. Ce processus permet de transformer les ARNm en ADN, qui peuvent ensuite être amplifiés et séquencés. Les fragments d'ADNc sont alors amplifiés pour préparer la librairie, qui correspond à l'ensemble des fragments d'ADNc destinés au séquençage. Cette librairie est ensuite analysée par des techniques de séquençage de nouvelle génération, aussi appelées *Next generation high-throughput sequencing technologies* en anglais (Reuter et al. 2015), qui permettent de lire les séquences des nucléotides (A, T, C, G) dans les fragments d'ADNc. Ces données sont ensuite alignées selon un génome de référence établi pour l'organisme étudié. L'alignement consiste à rechercher dans le génome de référence les sous-séquences identiques à celles séquencées dans l'échantillon. Plusieurs outils, tels que Salmon (Patro et al. 2017), peuvent être utilisés pour effectuer cet alignement de manière efficace et précise. Mais il s'agit d'une tâche complexe qui constitue un domaine de recherche actif à part entière. Une fois l'alignement effectué, l'expression de chaque gène est évaluée en comptant le nombre de fois où son transcrit, donc sa séquence de nucléotides, a été identifiée durant l'alignement. Ce comptage, représentant le nombre de lectures de chaque gène, permet alors d'estimer son niveau relatif d'expression dans l'échantillon.

Au départ, le RNA-seq était utilisé pour mesurer l'expression génique dans des échantillons ou des tissus, qui contiennent un mélange de cellules. Ces mesures génèrent ce que l'on appelle des données RNA-seq en masse (ou *bulk* RNA-seq en anglais), qui témoignent de l'expression moyenne de chaque gène au sein de l'échantillon, ignorant ainsi son hétérogénéité cellulaire. Cependant, avec les progrès technologiques récents, il est maintenant possible d'aller plus loin dans la résolution et de mesurer l'expression génique au niveau individuel des cellules qui composent ces échantillons (Nawy 2014). Cette technique, connue sous le nom de séquençage de l'ARN en cellule unique (ou scRNA-seq pour *single-cell* RNA-seq en anglais), permet d'isoler les cellules une par une, souvent grâce à des techniques de microfluidique ou de tri cellulaire. L'ARN contenu dans chaque cellule est ensuite converti en ADNc, auxquels des identifiants moléculaires uniques, des UMI (pour *Unique Molecular Identifiers*) (Islam et al. 2014), sont attachés pour identifier les molécules d'ARN provenant de différentes cellules. L'ajout de ces UMI est nécessaire en raison de la faible quantité de matériel biologique disponible pour la mesure. En effet, pour être séquencé, l'ADNc est amplifié à l'aide d'une PCR (*Polymerase Chain Reaction*) qui prend la molécule d'origine et la duplique plusieurs fois. Cela a l'effet d'amplifier le

signal mais si une molécule est amplifiée un nombre de fois trop important, il devient possible de la compter plusieurs fois. Ainsi, la PCR distord les données à cause de cette potentielle multiplication. Les UMI permettent de corriger cette distorsion en attribuant un identifiant unique à chaque molécule d'ADNc, ce qui permet de compter avec précision le nombre de molécules originales présentes dans l'échantillon.

1.1.2 Les données d'expression génique

Les données d'expression génique obtenues à partir du RNA-seq en masse ou du scRNA-seq se présentent sous forme de comptage, reflétant le nombre de fois où un transcrit d'ARN spécifique est identifié dans un échantillon biologique. Les données de comptage peuvent être modélisées par une loi de Poisson (Townes 2020), une distribution qui décrit la probabilité d'observer un certain nombre d'événements, tels que la détection d'un transcrit spécifique d'ARNm, dans un intervalle de temps ou un espace donné. La loi de Poisson est caractérisée par un seul paramètre de moyenne λ, égal à sa variance, qui permet de quantifier l'espérance des événements observés.

Initialement proposée par Marioni et al. (2008) pour modéliser les données RNA-seq en masse, la distribution de Poisson ne peut être utilisée que pour tenir compte de la variabilité technique dans les données. Cette dernière, causée par des facteurs expérimentaux externes à la biologie de l'échantillon, est généralement considérée comme faible dans les données RNA-seq en masse (Marioni et al. 2008). La distribution de Poisson semble alors adaptée en présence de réplicats techniques, c'est-à-dire des échantillons multiples provenant d'une même source biologique et traités indépendamment dans une expérience pour évaluer la variabilité introduite par les méthodes expérimentales. Cependant, Anders & Huber (2010) ont démontré que la variabilité biologique, résultant de facteurs biologiques, dépasse souvent celle prédite par la loi de Poisson, remettant ainsi en question son adéquation pour modéliser correctement les données de RNA-seq. En effet, en présence de réplicats biologiques, c'est-à-dire différents échantillons biologiques provenant de différents individus mais soumis aux mêmes conditions expérimentales, l'égalité entre la moyenne et la variance imposée par la loi de Poisson n'est plus vérifiée. On observe au contraire une variance des comptes supérieure à leur moyenne : c'est ce qu'on appelle la surdispersion. Il devient donc nécessaire d'adapter la modélisation pour tenir compte de cette surdispersion non prise en compte par la distribution de Poisson.

De manière générale, les modèles de compositions sont souvent utilisés pour prendre en compte la surdispersion (Agresti 2015). C'est pourquoi, la distribution binomiale négative, en tant que composition Poisson-Gamma, est couramment employée pour modéliser les données RNA-seq. Considérons Y_{ig} comme le compte du gène g dans l'échantillon i. Si $Y_{ig}|\lambda_{ig} \sim \mathcal{P}(\lambda_{ig})$ et que l'on suppose que $\lambda_{ig} \sim \Gamma(k_{ig}, \mu_{ig})$ (composition Poisson-Gamma), alors marginalement, $Y_{ig} \sim \text{NegBin}(\mu_{ig}, \theta_{ig})$ (Bonafede et al. 2016, Townes 2020), où

μ_{ig} représente le paramètre de moyenne et $\theta_{ig} = 1/k_{ig}$ est le paramètre de surdispersion reflétant les variations biologiques. Lorsque θ_{ig} tend vers zéro (donc en l'absence de surdispersion), la distribution binomiale négative se réduit à une distribution de Poisson. Ainsi, cette approche de modélisation permet de prendre en compte la variabilité technique grâce à la distribution de Poisson, tout en incorporant la surdispersion due à la variabilité biologique (Robinson et al. 2010). Les variances des comptes ne sont donc plus égales à leurs moyennes comme pour la distribution de Poisson, mais bien supérieures, puisque $\text{Var}(Y_{ig}) = \mu_{ig} + \mu_{ig}^2\theta_{ig}$.

La nature des données RNA-seq (des données de comptages surdispersées) implique également une relation entre la moyenne et la variance des comptes, ce qui correspond à une forme d'hétéroscédasticité. Elle se traduit par une variance plus grande pour les gènes exprimés à des niveaux plus élevés par rapport à ceux exprimés à des niveaux plus faibles. Law et al. (2014) ont alors proposé d'estimer cette relation moyenne-variance, associée à une normalisation appropriée des données, pour contourner l'utilisation de la distribution binomiale négative au profit de la distribution gaussienne. Cette dernière permet alors d'utiliser une palette d'outils statistiques plus large, tels que les modèles linéaires pondérés, tout en tenant compte des spécificités des données RNA-seq.

Les données scRNA-seq se distinguent des données RNA-seq en masse par une caractéristique notable : la prévalence élevée de comptes nuls (c'est-à-dire de zéros), atteignant jusqu'à 90% selon les études (Townes et al. 2019). On parle alors d'une inflation en zéros. Elle peut être expliquée par divers facteurs, comme l'ont souligné Hicks et al. (2018). Certains zéros peuvent être dus à des facteurs biologiques : au moment de la mesure, le gène ne s'exprime tout simplement pas. En revanche, une grande proportion de ces zéros est souvent causée par des facteurs techniques. En raison de la faible quantité d'ARN extraite de chaque cellule individuelle, il devient plus difficile de mesurer avec précision les transcrits. Ce défi technique conduit au phénomène de *dropout*, où certains transcrits effectivement présents dans la cellule ne sont pas détectés dans les données scRNA-seq (Qiu 2020). Pour tenir compte de cette forte inflation en zéros, des modèles binomiaux négatifs enflés en zéros ont été proposés pour modéliser ces données (Risso et al. 2018). Ce sont des modèles de mélanges à deux composantes : une masse en zéro et une composante binomiale négative permettant de modéliser les comptes non-nuls. Cependant, en raison de la forte hétérogénéité biologique, une simple modélisation bimodale des données comme celle obtenue avec les modèles binomiaux négatifs enflés en zéros, peut ne pas être suffisante pour caractériser tous les modes présents dans l'expression génique (Korthauer et al. 2016).

1.1.3 L'exemple de la COVID-19

Les données d'expression génique font l'objet d'une recherche active, notamment en termes de développements méthodologiques, en raison de la place qu'elles ont maintenant au sein de la recherche médicale contemporaine. L'analyse de ces données est devenue un bon outil pour comprendre les mécanismes biologiques sous-jacents aux maladies, identifier des biomarqueurs diagnostiques et thérapeutiques, et développer des thérapies personnalisées. Pour illustrer l'importance de ces analyses dans un contexte concret, cette section se penchera sur l'exemple de la pandémie de la COVID-19. Causée par le virus SARS-CoV-2, elle a débuté à Wuhan, en Chine, en 2019, et s'est rapidement propagée dans le monde entier. Dans ce contexte de crise sanitaire, la communauté scientifique a déployé des efforts considérables pour faire face à l'épidémie. Les analyses des données RNA-seq ont alors permis d'améliorer la compréhension de ce virus, ses mécanismes et les réponses immunitaires des patients, comptabilisant plus de 800 publications sur PubMed.

Les données RNA-seq sont souvent combinées avec d'autres types de données provenant des mêmes échantillons, telles que des données cliniques ou d'autres données moléculaires, dans ce qu'on appelle des analyses multiomiques ou intégratives. Ces approches permettent d'identifier des signatures biologiques caractéristiques associées à une maladie. Par exemple, dans leur étude, Bernardes et al. (2020) ont analysé des données RNA-seq en masse, des données scRNA-seq et d'autres types de données moléculaires à différents moments de l'évolution de la maladie chez 14 patients infectés par la COVID-19. Cette analyse intégrative longitudinale a permis d'identifier plusieurs signatures moléculaires associées à la trajectoire de la sévérité de la COVID-19. Dans le même esprit, nous avons, dans l'étude de Lévy et al. (2021), intégré des données RNA-seq en masse avec les phénotypes cellulaires (c'est-à-dire les proportions de différentes populations cellulaires) ainsi que la concentration de cytokines (des protéines permettant aux cellules du système immunitaire de communiquer entre elles) chez 54 patients, dont 44 étaient infectés par la COVID-19 et 10 étaient des patients sains. Grâce à un clustering basé uniquement sur les données d'expression génique (29 302 gènes), nous avons pu mettre en avant trois groupes de patients infectés par la COVID-19 ayant des profils d'expression génique différents. Nous avons ensuite conduit une analyse intégrative des trois différentes sources de données (incluant le RNA-seq) en utilisant MOFA (Argelaguet et al. 2020), une méthode de réduction de dimension adaptée aux données multiomiques. Il a donc été possible de constuire des composantes qui sont des combinaisons linéaires des variables d'origines et qui expliquent la variabilité biologique inter-patient qui est partagée par l'ensemble des sources de données. La première composante, qui est celle expliquant la plus grande partie de cette variabilité, permettait de séparer deux des premiers clusters entre eux, et de manière générale les trois clusters de patients infectés des donneurs sains comme le montre la Figure 1.1A. Cette composante permet donc d'interpréter les trois clusters construits

en termes de sévérité de l'infection. En examinant la contribution des variables à cette première composante (Figure 1.1B), et grâce à des analyses supplémentaires, nous avons ainsi pu mettre en évidence l'activation des neutrophiles, un type de globules blancs essentiels du système immunitaire, comme une signature de la sévérité et identifier le gène CD177 comme un marqueur de celle-ci. Cette signature liée aux neutrophiles a également été retrouvée dans d'autres études transcriptomiques (Aschenbrenner et al. 2021, Jackson et al. 2022).

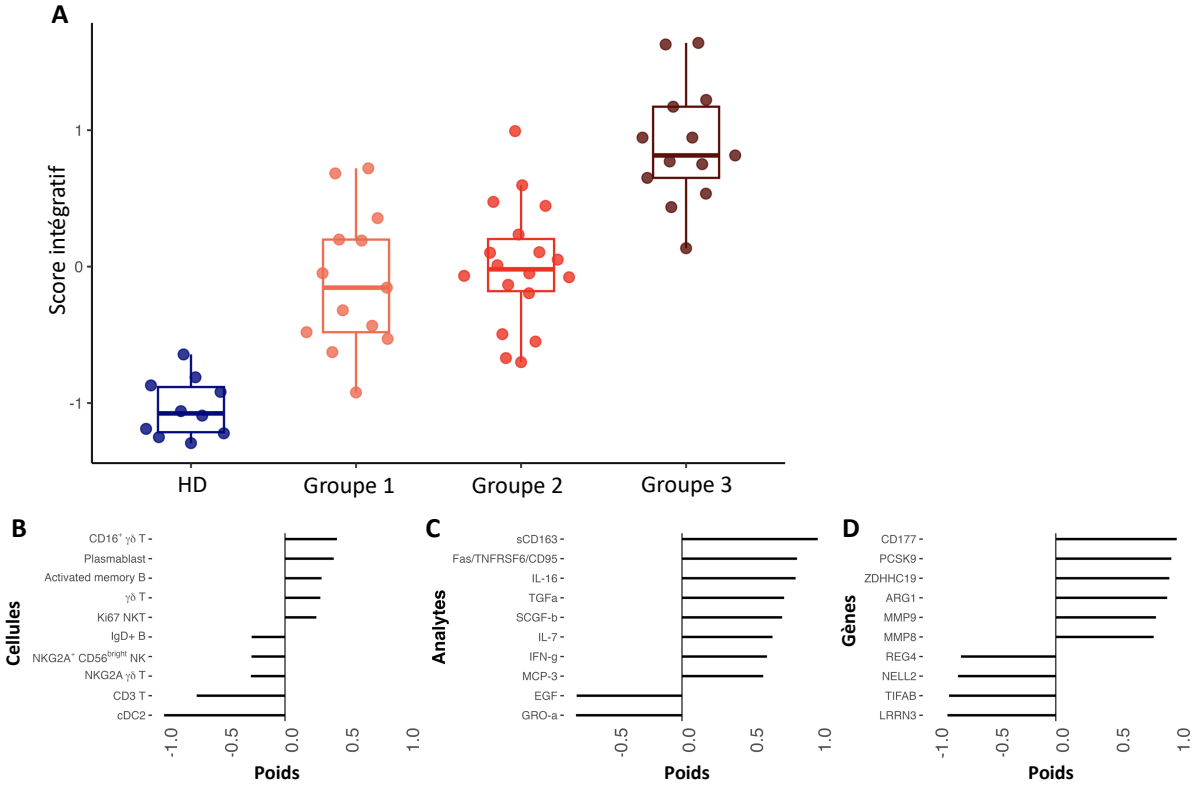


FIGURE 1.1 – **Analyse intégrative des données FrenchCOVID (Lévy et al. 2021).** Le panneau A montre la projection des 54 patients sur la première composante de MOFA (score intégratif) en fonction des résultats du clustering en trois clusters basé sur les données RNA-seq. Les panneaux B-D montrent la contribution des différentes variables (populations cellulaires, cytokines, gènes) à la construction de la première composante.

Les données scRNA-seq sont un outil puissant pour caractériser l'hétérogénéité cellulaire, offrant une image précise des populations cellulaires présentes dans l'organisme d'un individu à un instant donné. Elles contribuent ainsi à la compréhension des mécanismes biologiques sous-jacents aux infections. Par exemple, Blanco-Melo et al. (2020) ont étudié la réponse de l'organisme face à l'infection à la COVID-19 et l'ont comparée à celle d'autres virus respiratoires. Ils ont ainsi pu mettre en avant des signatures biologiques spécifiques à la COVID-19. À partir de l'analyse transcriptomique des cellules du sang courant chez 7 patients COVID-19 (dont 4 avec une infection sévère) et 6 donneurs sains, Wilk et al. (2020) ont pu mettre en avant les différences en termes d'abondance des po-

pulations cellulaires liées à la réponse immunitaire entre les donneurs sains et les patients infectés. Ils ont également identifié une nouvelle population cellulaire dont la proportion était plus élevée chez les patients dans un état sévère. Zhang et al. (2020) ont mené une étude similaire à partir de données scRNA-seq issues du sang courant de 13 patients COVID-19 avec des niveaux de sévérité différents et ont ainsi illustré la dynamique de la réponse immunitaire durant la progression de la maladie. Sokal et al. (2021) se sont eux concentrés sur les cellules B, composantes clés de la mémoire immunitaire, à travers une étude scRNA-seq longitudinale chez des patients COVID-19 sévères jusqu'à 6 mois après l'infection pour étudier la protection à long terme induite par l'infection.

Ces exemples mettent en évidence l'importance qu'ont prises les analyses de données RNA-seq dans la compréhension des maladies, comme l'illustre la pandémie de COVID-19. Les approches multiomiques, intégrant les données RNA-seq avec d'autres sources de données, ont permis d'identifier des signatures moléculaires associées à la maladie et d'en explorer les mécanismes biologiques sous-jacents. Les données scRNA-seq offrent une cartographie précise de la réponse immunitaire face à l'infection à la COVID-19. Ces analyses ont non seulement aidé à une meilleure compréhension de la maladie, mais ont également ouvert la voie à la découverte de cibles thérapeutiques potentielles et permis l'accélération du développement de vaccins, contribuant ainsi à faire progresser la médecine et à relever les défis de santé publique.

1.1.4 Statistiques en grande dimension

Au cours des dernières décennies, un changement de paradigme majeur s'est opéré dans le domaine de la statistique. La statistique s'est longtemps concentrée sur l'analyse d'un petit nombre de variables mesurées chez un grand nombre d'individus. Les méthodes classiques telles que la régression linéaire étaient alors développées en accord avec ce paradigme. Cependant, avec les progrès technologiques récents, il est désormais possible de mesurer un nombre croissant de caractéristiques chez un même individu. Cette évolution se traduit par une augmentation significative du nombre de variables, comme en témoigne l'essor des données RNA-seq. En effet, grâce au séquençage de l'ARNm, il est désormais possible de mesurer l'expression des 30 000 gènes qui composent le génome humain, mais sur un nombre limité de patients (pour des exemples supplémentaires de données de grande dimension, voir Donoho et al. (2000)). Ce nouveau paradigme, où le nombre de variables est nettement plus important que le nombre d'échantillons, correspond au concept de grande dimension. Cette capacité croissante à mesurer davantage de caractéristiques semble être une avancée prometteuse pour une meilleure description et compréhension des objets d'études. Cependant, l'ajout de variables supplémentaires, pourtant source d'information, se traduit également par une augmentation du bruit dans les données. Dissocier ce bruit du signal devient alors un enjeu majeur de la grande dimen-

sion. D'autres phénonèmes plus inattendus émergent également de la grande dimension des données : c'est ce que Bellman et al. (1957) nomment le fléau de la dimension (*the curse of dimensionality* en anglais). Les méthodes statistiques traditionnelles deviennent difficiles voire impossibles à appliquer en grande dimension. Un exemple concret de cette limitation se retrouve dans le domaine de la régression linéaire, où il devient impossible d'obtenir des estimations des coefficients de régression en grande dimension, en raison de la non-inversibilité de la matrice des covariables.

Dans son livre, Giraud (2021) aborde les phénomènes qui apparaissent dans les espaces de grande dimension en statistique. Il met d'abord en avant l'effet de l'expansion des espaces de haute dimension sur la distribution des observations. Dans de tels espaces, les observations tendent à être isolées les unes des autres. Plus précisément, alors que la dimension p des données augmente, les distances entre les observations dans l'espace augmentent également. Cependant, cette augmentation des distances est relativement homogène, ce qui signifie que les distances entre les observations sont similaires entre elles, malgré leur éloignement croissant. Ce phénomène est illustré par la Figure 1.2, qui montre la distribution des distances euclidiennes entre $n = 100$ observations distribuées uniformément dans l'hypercube de dimension p, $[0, 1]^p$. À mesure que la dimension p croît, les distances euclidiennes entre les observations augmentent également, mais leur répartition reste concentrée autour d'une valeur moyenne. Autrement dit, bien que chaque observation semble éloignée des autres, les différences entre ces distances restent petites, ce qui peut sembler contre-intuitif. Ainsi, dans les espaces de grande dimension, bien que les

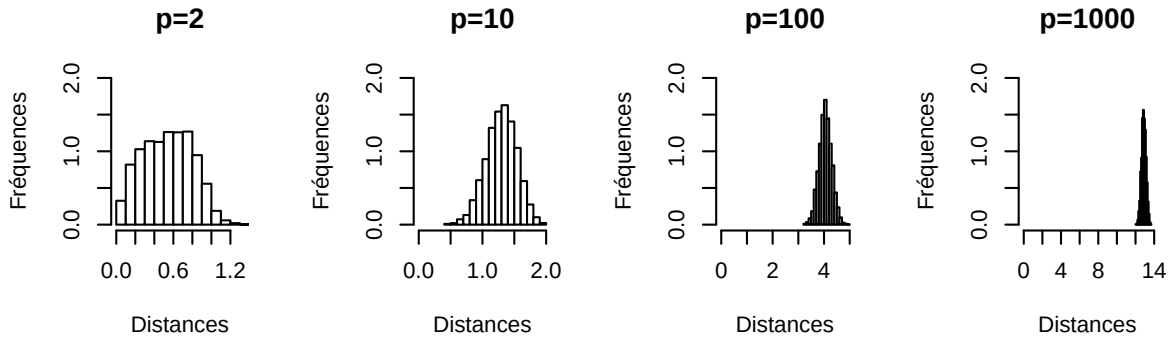


FIGURE 1.2 – **Distance deux à deux entre $n = 100$ observations dans l'hypercube $[0, 1]^p$.**

distances entre les observations augmentent, elles restent étonnamment similaires les unes par rapport aux autres. Cela signifie que, malgré une dispersion apparente où les points semblent isolés, ils restent à des distances relativement comparables les uns des autres. En conséquence, la notion de voisinage disparait en grande dimension, rendant inefficaces les méthodes statistiques traditionnelles qui reposent sur des mesures de distance ou des moyennes locales (Hinneburg et al. 2000).

La grande dimension est également propice à l'apparition de phénomènes rares. Giraud (2021) prend l'exemple des fausses découvertes pour illustrer ce point. En effet, dans le cadre de tests d'hypothèses, tels que des tests d'association entre des variables et un résultat dans un ensemble de données de grande dimension, la probabilité de trouver des résultats significatifs par pur hasard augmente considérablement avec le nombre de tests effectués, entraînant une explosion du nombre de fausses découvertes. Considérons la probabilité de déclarer à tort une variable comme étant significativement associée à une réponse. Habituellement, un seuil de significativité, tel que 5%, est utilisé pour conclure : nous avons alors 5% de chances de considérer à tort la variable comme significativement associée à la réponse, en faisant un évènement rare. Cependant, lorsque ce même test d'association est effectué sur un grand nombre de variables, par exemple $p = 1000$ variables, en utilisant le même seuil de significativité, cette probabilité de 5% se répercute sur l'ensemble des variables testées. Par conséquent, au total, nous risquons de conclure à tort que $0.05 \times 1000 = 50$ variables sont significativement associées à la réponse, conduisant ainsi à 50 fausses découvertes, simplement en raison du hasard. Ce phénomène est un exemple classique des problèmes liés à la multiplicité des tests, où la chance de trouver des résultats apparemment significatifs augmente avec le nombre de tests effectués. Une méthode intuitive pour corriger ce problème est la méthode de Bonferroni (Holm 1979), qui consiste à diviser le niveau de significativité par le nombre de variables testées. En reprenant notre exemple, en considérant un seuil de 5×10^{-5} au lieu de 0.05, seulement $5 \times 10^{-5} \times 1000 = 0.05$ variables auront une association à la réponse due au hasard. Bien que largement utilisée, cette méthode de correction est critiquée pour son caractère trop conservateur.

Enfin, de manière générale, la compréhension de l'origine de l'information en grande dimension est complexe. Un petit signal présent sur un grand nombre de variables peut générer un signal global conséquent, ce qui rend difficile la distinction entre signal et bruit. L'omniprésence de corrélations entre les variables complexifie cette dissociation entre signal et bruit. Ces défis, auxquels s'ajoutent une grande complexité computationnelle, rendent donc l'analyse des données en grande dimension particulièrement délicate et la statistique en grande dimension est alors devenue un pan actif de la recherche en statistiques.

Fort heureusement, dans les données de grande dimension, il existe généralement des structures de dimensions plus petites contenant les sources principales d'information. La recherche de ces structures cachées devient ainsi un enjeu majeur en statistique de grande dimension, car une fois identifiées, elles permettent de revenir à un paradigme plus simple de petite dimension. Deux types d'approches sont alors à distinguer : les méthodes de réduction de dimension et les méthodes de sélections de variables. Les methodes de réduction de dimension s'intéressent à résumer l'information pertinente pour ensuite faciliter l'application d'autres méthodes. La méthode la plus connue est l'Analyse en Composante

Principale (ACP). Il s'agit d'une technique de réduction de dimension linéaire (au sens où elle projette les observations dans un espace de faible dimension défini par des combinaisons linéaires des variables d'origine) qui a pour objectif de conserver le plus possible la structure globale de variation des données. Il s'agit de la méthode de réduction de dimension la plus couramment utilisée bien que d'un point de vu purement probabiliste, elle ne soit adaptée qu'aux données distribuées selon une loi normale (Tipping & Bishop 1999). D'autres méthodes de réduction de dimension ont été proposées pour généraliser l'ACP à d'autres distributions comme la Poisson (PLN-PCA, (Chiquet et al. 2018)) ou la binomiale négative (GLM-PCA, (Townes et al. 2019)) qui sont donc plus adaptées à la nature des données d'expression génique. Les méthodes de sélection de variables ont quant à elles pour ambition de sélectionner un petit sous-ensemble de variables (de caractéristiques) en éliminant celles qui sont non-informatives (Sammut & Webb 2011). Contrairement aux méthodes de réduction de dimension, leur objectif n'est pas de résumer l'information, mais plutôt de la filtrer en éliminant les variables non-informatives. Il existe une grande variété d'algorithmes de sélection de variables (Liu & Yu 2005) mais pour l'analyse de données RNA-seq la sélection de variables (dans ce cas de gènes) se fait traditionnellement en ne conservant que les gènes avec l'expression la plus variable ou la plus élevée (Townes et al. 2019).

1.2 Méthodes pour le clustering & l'analyse différentielle

1.2.1 Outils pour le clustering

Le clustering, ou regroupement, est un ensemble de méthodes statistiques visant à regrouper des observations, des variables, ou les deux, en sous-groupes appelés des clusters ou des classes (nous nous concentrerons ici principalement sur le clustering d'observations). Depuis des siècles, les scientifiques ont eu recours au regroupement d'objets ou d'entités sur la base de caractéristiques observables. Un exemple célèbre remonte à Aristote, qui a regroupé les espèces vivantes selon des caractéristiques morphologiques telles que la couleur de leur sang (Everitt & Hothorn 2006). En regroupant des observations entre-elles, il devient alors possible de les organiser voire de les hiérarchiser et donc de mieux les comprendre. Avec l'avènement de la grande dimension, le clustering permet d'identifier la structure sous-jacente des données et de l'extraire sous forme de cluster, offrant ainsi un résumé compact des relations entre les observations au sein de l'espace de haute dimension.

Bien que le clustering soit largement utilisé depuis des années, Hennig et al. (2015) soulignent qu'il n'existe toujours pas de définition formelle de ce que sont les clusters. En

effet, plusieurs critères peuvent être utilisés pour les définir. Certains se basent sur des mesures de distance : les observations au sein d'un même cluster doivent être proches les unes des autres, tandis que celles de clusters différents doivent être éloignées. D'autres critères considèrent que les observations d'un même cluster appartiennent au même mode de la distribution sous-jacente des données ou encore qu'elles sont issues de la même distribution de probabilité. Pour englober ces différents critères, nous adoptons ici la définition de Everitt & Hothorn (2006), qui se concentre sur deux aspects : i) l'homogénéité des observations au sein d'un cluster, et ii) la séparabilité des observations entre deux clusters différents.

Considérons un ensemble de n observations $\boldsymbol{x}_i \in \mathbb{R}^p$, décrites par p variables. L'objectif du clustering, tel que défini précédemment, est de construire une partition ou une hiérarchie des n observations $\boldsymbol{x}_i$ en K clusters $C_1, \ldots, C_K$, qui sont à la fois homogènes et séparés les uns des autres. Une partition en K clusters des n observations conduit à des clusters disjoints ($C_k \cap C_l = \emptyset$, $k, l \in \{1, \ldots, K\}$ pour $k \neq l$), où chaque observation ne peut appartenir qu'à un seul et unique cluster. En revanche, une hiérarchie sépare les observations en une série de clusters qui sont imbriqués les uns dans les autres. Deux clusters C_k et C_l, $k, l \in \{1, \ldots, K\}$ sont soit disjoints, soit l'un est inclu dans l'autre. À la base de la hiérarchie se trouvent les clusters formés par chaque observation individuelle et en son sommet se trouve le cluster formé par l'entièreté des n observations.

Parmi les méthodes de clustering les plus couramment utilisées, trois approches se distinguent par les critères qu'elles emploient pour définir l'homogénéité au sein des clusters et leur séparation. Tout d'abord, les méthodes basées sur la distance utilisent des mesures géométriques pour évaluer la similarité entre les observations, regroupant celles qui sont proches selon un critère de distance donné. Ensuite, les approches probabilistes modélisent les données en supposant que les observations de chaque cluster proviennent d'une distribution de probabilité spécifique. Enfin, les méthodes basées sur les graphes exploitent la structure de connexion des données, représentant les observations comme des nœuds reliés par des arêtes pondérées, ce qui permet de détecter des communautés ou des groupes en s'appuyant sur la théorie des graphes.

Classification ascendante hiérarchique et *k*-means

La classification ascendante hiérarchique (CAH) et les k-means sont deux des méthodes de clustering les plus connues. Elles s'appuient toutes les deux sur la notion de distance entre observations pour décrire l'homogénéité au sein d'un cluster ainsi que la séparation entre les clusters. L'homogénéité d'un cluster C_k est alors décrite par son inertie définie par :

$$W_k = \sum_{i \in C_k} d\left(\boldsymbol{x}_i, \overline{\boldsymbol{x}}_{C_k}\right)^2$$

où d est une distance (*e.g* la distance euclidienne) et $\overline{\boldsymbol{x}}_{C_k} = \frac{1}{|C_k|} \sum_{i \in C_k} \boldsymbol{x}_i$ est le barycentre du cluster C_k. Ainsi, l'inertie d'un cluster donne une indication de l'éloignement des observations contenues dans ce cluster par rapport à son centre. Un cluster homogène doit donc avoir une faible inertie (Figure 1.3, gauche). L'inertie intra-classe W associée à la partition est donc définie comme étant la somme des inerties de chaque cluster W_k : $W = \sum_{k=1}^{K} W_k$. Plus les clusters sont homogènes entre eux et plus W sera basse. La séparabilité entre les clusters C_k et C_l pour $k, l \in \{1, \ldots, K\}$ est quand à elle décrite par l'inertie inter-classe. Il s'agit de la somme des distances des barycentres de chaque cluster par rapport au barycentre global des données. Elle est donnée par :

$$B = \sum_{k=1}^{K} d\left(\overline{\boldsymbol{x}}_{C_k}, \overline{\boldsymbol{x}}\right)^2$$

où $\overline{\boldsymbol{x}} = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{x}_i$. Lorsque les clusters sont séparés les uns des autres, alors leurs barycentres $\overline{\boldsymbol{x}}_{C_k}$ sont éloignés du barycentre global des observations et nécessairement B est grande (Figure 1.3, milieu). Un bon clustering des observations doit donc conduire à une faible inertie intra-classe W et à une grande inertie inter-classe B. Fort heureusement, d'après le théorème d'Huygens, l'inertie totale des données $T = \sum_{i=1}^{n} d\left(\boldsymbol{x}_i, \overline{\boldsymbol{x}}\right)^2$ (Figure 1.3, droite) est égale à la somme entre l'inertie intra et inter-classe : $T = W + B$. Il suffit donc de maximiser l'inertie inter-classe pour minimiser l'inertie intra-classe.

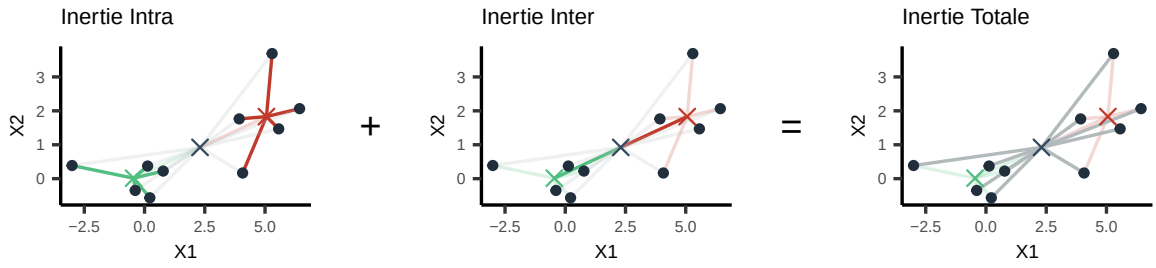


FIGURE 1.3 – **Les inerties comme mesures d'homogénéité et de séparation.** Sur la figure, les points corrrespondent aux observations, les croix représentent les barycentres et la coloration se fait en fonction des deux clusters d'observations présents dans les données.

Historiquement, la méthode de classification ascendante hiérarchique a été la première à être introduite par Ward Jr (1963). Cette méthode construit une hiérarchie des observations de manière ascendante, d'où son nom. En partant du clustering en n clusters des observations, où chaque observation forme donc son propre cluster, la CAH agrège itérativement les clusters jusqu'à parvenir à la partition en un cluster contenant toutes les observations. Les deux clusters agrégés à chaque itération sont ceux qui sont définis comme étant les plus similaires selon une mesure d'agrégation. L'algorithme de la CAH

est présenté dans l'Algorithme 1.

Algorithme 1 Classification Ascendante Hiérarchique (CAH)

Initialisation : $K = n$
Partition en n clusters $(C_1, \ldots, C_n)$ où $C_k = \{\boldsymbol{x}_k\}, k = 1, \ldots, n$
tant que $K \neq 1$ **faire**
  Calculer les distances $D_{kl}, k \neq l, k, l = 1, \ldots, K$ entre chaque paire de clusters C_k et C_l de la partition en K clusters $(C_1, \ldots, C_K)$
  Agréger les deux clusters C_k et C_l qui minimisent D_{kl} pour former le cluster $C_{k'}$: $C_{k'} = C_k \cup C_l$
  Mettre à jour la partition en $K = K - 1$ clusters

Au cœur de cette méthode se trouve donc le calcul de la mesure d'agrégation D_{kl} entre les deux clusters C_k et C_l qui reflète leur similarité. Bien que pour l'initialisation, comme chaque cluster se compose d'une unique observation, il soit possible d'utiliser la distance euclidienne, il est nécessaire de généraliser cette notion de distance dès la première étape de l'algorithme pour pouvoir la définir entre des groupes d'observations. A partir d'une distance d, il est alors possible de définir plusieurs mesures d'agrégation. Parmi les plus connues illustrées sur la Figure 1.4, on trouve :

— La distance minimale (*single link* (Sneath 1957)), définie par

$$\mathrm{Dmin}_{kl} = \min_{i \in C_k, j \in C_l} d(\boldsymbol{x}_i, \boldsymbol{x}_j)$$

qui s'interprète comme la plus petite distance entre deux observations de deux clusters disjoints.

— La distance maximale (*complete link* (Defays 1977)), définie par

$$\mathrm{Dmax}_{kl} = \max_{i \in C_k, j \in C_l} d(\boldsymbol{x}_i, \boldsymbol{x}_j)$$

qui s'interprète comme la plus grande distance entre deux observations de deux clusters disjoints.

— La distance de Ward (Ward Jr 1963), définie par

$$\mathrm{DWard}_{kl} = \frac{|C_k| \times |C_l|}{|C_k| + |C_l|} d(\overline{\boldsymbol{x}}_{C_k}, \overline{\boldsymbol{x}}_{C_l})^2$$

Il est important de noter que seule la classification ascendante hiérarchique de Ward, c'est-à-dire celle utilisant la mesure de Ward pour agréger les clusters, permet de conduire à une partition des observations qui minimise l'inertie intra-classe (et donc qui maximise l'inertie inter-classe). La distance utilisée pour définir les agrégations a donc un fort impact sur la hiérarchie ainsi que sur ses propriétés.

Les résultats de la classification ascendante hiérarchique peuvent être représentés par

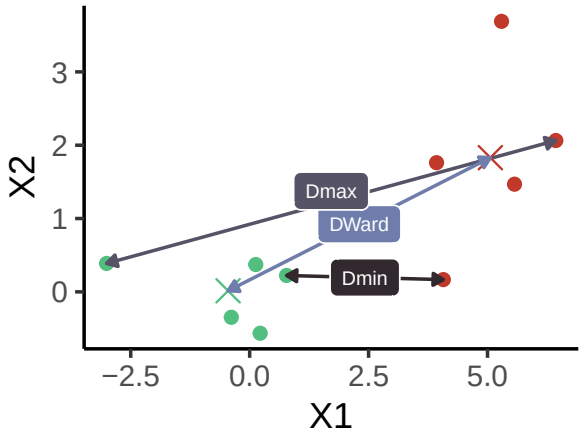


FIGURE 1.4 – **3 mesures d'agrégation usuelles en classification ascendante hiérarchique.** Sur la figure, les points corrrespondent aux observations, les croix représentent les barycentres et la coloration se fait en fonction des deux clusters d'observations présents dans les données.

un dendrogramme, tel que celui présenté à la Figure 1.5. Pour obtenir une partition à partir de ce dendrogramme, il suffit de le découper horizontalement. La hauteur h à laquelle le découpage se fait donne alors le nombre K de clusters qui seront construits. Elle représente la mesure d'agrégation utilisée pour passer de la partition en $K-1$ clusters à celle en K clusters.

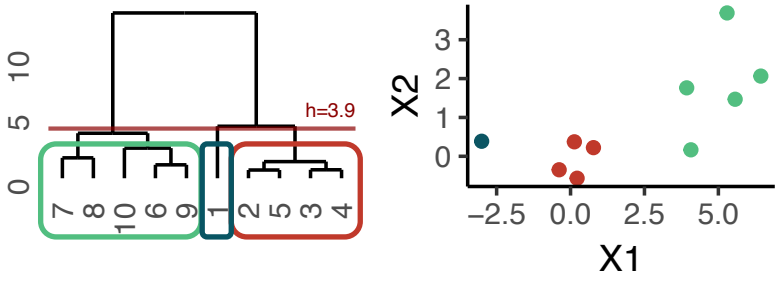


FIGURE 1.5 – **La classification ascendante hiérarchique**. Sur la figure, les points correspondent aux observations et leur coloration se fait en fonction des $K = 3$ clusters d'observations construits en découpant le dendrogramme à la hauteur $h = 3.9$.

Les k-means (ou k-moyennes), introduits par MacQueen et al. (1967), constituent également une approche largement utilisée en clustering basé sur des distances. Son objectif est de partitionner un ensemble de n observations en K clusters de manière à minimiser l'inertie intra-classe W. L'algorithme commence par une initialisation aléatoire, où K observations sont tirées aléatoirement parmi les n et servent de barycentres initiaux $\overline{\boldsymbol{x}}_{C_1}, \ldots, \overline{\boldsymbol{x}}_{C_K}$ pour les K clusters. La première étape de l'algorithme, l'étape d'affectation, consiste à calculer la distance de chaque observation par rapport aux barycentres

des K clusters. Chaque observation est alors affectée à la classe C_k telle que $i \in C_k$ si $k = \text{argmin}_{k=1,\dots,K} d(\boldsymbol{x}_i, \overline{\boldsymbol{x}}_{C_k})$, c'est-à-dire celle dont elle est la plus proche (au sens d'une distance). Les barycentres $\overline{\boldsymbol{x}}_{C_k}$ des clusters sont ensuite mis à jour en tenant compte de ces nouvelles affectations : c'est l'étape de représentation. Les k-means répètent alors ces étapes d'affectation et de représentation jusqu'à convergence vers une partition stable ou jusqu'à atteindre un nombre maximum d'itérations spécifié par l'utilisateur. Les k-means sont illustrés par la Figure 1.6

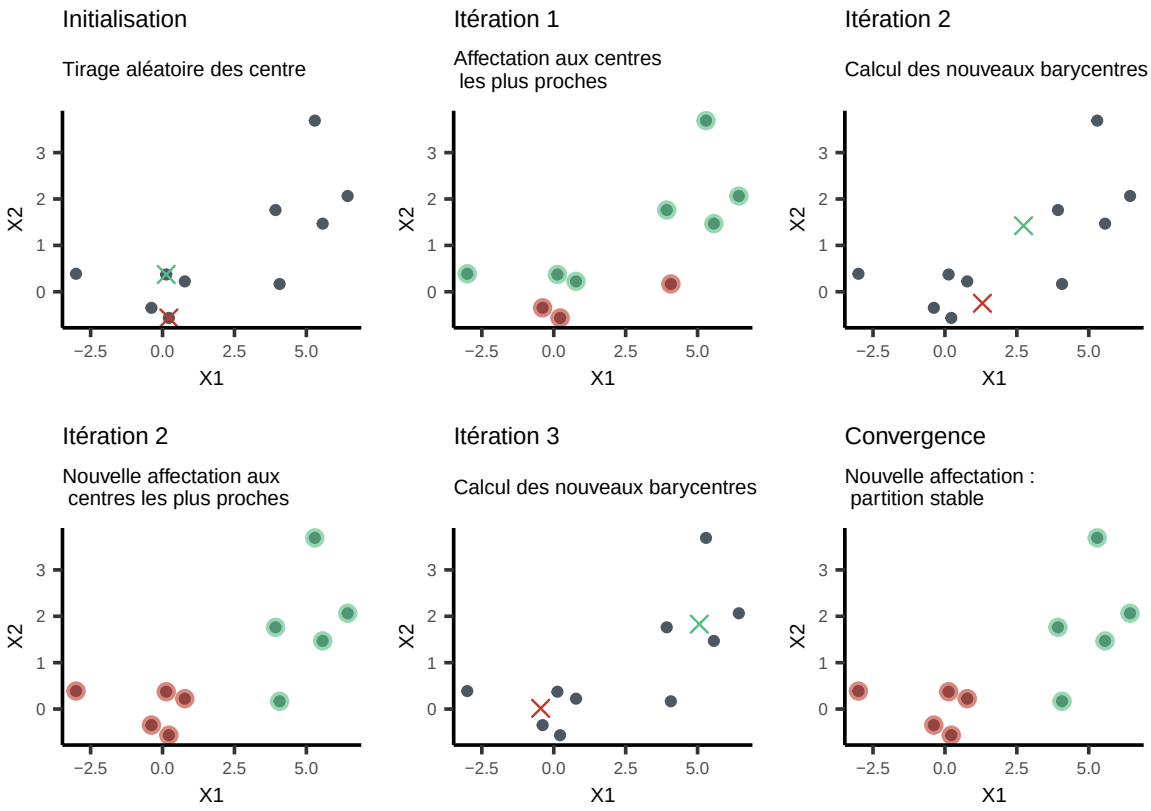


FIGURE 1.6 – **Illustration de l'algorithme des k-means.** Sur la figure, les points correspondent aux observations, les croix représentent les barycentres et la coloration se fait en fonction des $K = 2$ clusters d'observations obtenus à chaque itération de l'algorithme des k-means.

Il a été démontré que l'algorithme des k-means converge vers un minimum local de l'inertie intra-classe (MacQueen et al. 1967). L'initialisation aléatoire des k-means peut ainsi influencer la partition finale obtenue, car différentes initialisations peuvent conduire à des solutions de clustering différentes, c'est-à-dire converger vers des minimums locaux différents. Par conséquent, en pratique, les k-means sont souvent appliqués plusieurs fois avec des initialisations aléatoires différentes, et la partition finale retenue est celle présentant la plus petite inertie intra-classe.

1.2.1.1 Clustering basé sur un modèle probabiliste

Les méthodes de clustering basées sur des modèles probabilistes adoptent une approche différente de la CAH et des k-means pour décrire ce qu'est un cluster. Elles supposent en effet que les observations d'un même cluster (ou d'une même classe) sont générées à partir d'une même distribution de probabilité. Contrairement aux approches basées sur des distances présentées précédemment, qui évaluent la similarité entre les observations en utilisant des distances, ces méthodes cherchent à modéliser directement les données en termes de distributions probabilistes. Selon cette approche, nos n observations $\boldsymbol{x}_i$ sont considérées comme des réalisations de n variables aléatoires $\boldsymbol{X}_i, i = 1, \ldots, n$, indépendantes et identiquement distribuées, qui suivent un modèle de mélange. Ainsi, chaque cluster $C_k, k = 1, \ldots, K$ est représenté par une composante du mélange. La densité de probabilité de ce type de distribution est définie par :

$$f(\boldsymbol{x}_i) = \sum_{k=1}^{K} \pi_k f(\boldsymbol{x}_i|\boldsymbol{\theta}_k) \tag{1.1}$$

où π_k est la probabilité d'appartenir à la $k^{\text{ème}}$ composante (donc au cluster C_k), $f(.|\boldsymbol{\theta}_k)$ est sa densité et $\boldsymbol{\theta}_k$ son vecteur de paramètres associé. Dans ce genre de modèle, la densité des observations issues du cluster C_k est donc donnée par $f(\boldsymbol{x}_i|\boldsymbol{x}_i \in C_k) = f(\boldsymbol{x}_i|\boldsymbol{\theta}_k)$ et la probabilité qu'une observation appartienne à ce cluster C_k est décrite par $\mathbb{P}(\boldsymbol{x}_i \in C_k) = \pi_k$.

Parmi les méthodes de clustering basées sur un modèle probabiliste les plus connues, on trouve le mélange de gaussiennes initialement introduit par Wolfe (1963). Ce modèle suppose que les observations au sein de chaque cluster C_k sont distribuées selon une loi normale, c'est-à-dire que $f(\boldsymbol{x}_i|\boldsymbol{\theta}_k) = \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}_k|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{x}_i - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu}_k)\right)$ où $\boldsymbol{\theta}_k = (\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ avec $\boldsymbol{\mu}_k \in \mathbb{R}^p$ le vecteur de moyennes associé à C_k et $\boldsymbol{\Sigma}_{k\times k} \in \mathbb{R}^{p\times p}$ sa matrice de variance-covariance. Cependant, d'autres types de modèles de mélange peuvent être utilisés, comme les mélanges à processus de Drichlet (Dahl 2006, Hejblum et al. 2019, Rouanet et al. 2023) ou encore, dans le cadre des données RNA-seq, les mélanges de distributions binomiales négatives (Li et al. 2018, 2023).

Les paramètres du modèle contenus dans le vecteur $\boldsymbol{\Theta} = (\pi_1, \ldots, \pi_K, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K)$, sont alors estimés à partir des données par maximum de vraisemblance. Une fois estimés, on peut attribuer chaque observation à la composante (cluster) la plus probable a posteriori, c'est-à-dire le cluster C_k tel que $\mathbb{P}(\boldsymbol{x}_i \in C_k|\boldsymbol{x}_i)$ soit maximale. Cette probabilité est facilement calculable à partir du vecteur des paramètres estimés $\hat{\boldsymbol{\Theta}}$, car en utilisant le théorème de Bayes, on a :

$$\mathbb{P}(\boldsymbol{x}_i \in C_k|\boldsymbol{x}_i) = \frac{f(\boldsymbol{x}_i|\boldsymbol{x}_i \in C_k) \times \mathbb{P}(\boldsymbol{x}_i \in C_k)}{f(\boldsymbol{x}_i)} = \frac{f_k(\boldsymbol{x}_i|\hat{\boldsymbol{\theta}}_k) \times \hat{\pi}_k}{\sum_{l=1}^{K} \hat{\pi}_l f_l(\boldsymbol{x}_i|\hat{\boldsymbol{\theta}}_l)} \tag{1.2}$$

Ainsi la tâche la plus ardue du clustering basé sur un modèle probabiliste réside dans la maximisation de la vraisemblance des données pour estimer le vecteur de paramètre $\boldsymbol{\Theta}$. Cette vraisemblance est donnée par :

$$\begin{aligned} \mathcal{L}(\boldsymbol{x}|\boldsymbol{\Theta}) &= \prod_{i=1}^{n} f(\boldsymbol{x}_i|\boldsymbol{\theta}) \\ &= \prod_{i=1}^{n} \sum_{k=1}^{K} \pi_k f_k(\boldsymbol{x}_i|\boldsymbol{\theta}_k) \end{aligned} \tag{1.3}$$

Cette vraisemblance est malheureusement difficile à maximiser sans savoir à quelle composante appartient chaque observation. Cette information est donnée par une variable latente (non-observée) $\boldsymbol{z}_i = (z_{i1}, \ldots, z_{ik})$ telle que $z_{ik} = 1$ si $\boldsymbol{x}_{ik} \in C_k$ et 0 sinon. On dit alors que la vraisemblance définie en (1.3) est incomplète. La vraisemblance complète des données est alors donnée par : $\mathcal{L}_c(\boldsymbol{x}, \boldsymbol{z}|\boldsymbol{\Theta})$. La vraisemblance incomplète des données se retrouve en intégrant la vraisemblance complète par rapport à $\boldsymbol{z}$. Cette caractéristique a motivé l'utilisation de l'algorithme EM (pour *Expectation-Maximisation*) proposé par Dempster et al. (1977) et détaillé notamment dans Bishop (2006) et Bouveyron et al. (2019). Il s'agit d'un algorithme itératif en deux étapes permettant la maximisation de la vraisemblance lorsque cette dernière est incomplète. L'idée principale derrière cette algorithme est de se dire que puisque la connaissance de $\boldsymbol{z}_i$ est nécessaire pour maximiser la vraisemblance, alors on peut commencer par estimer les valeurs de $\boldsymbol{z}_i$ en utilisant les paramètres actuels du modèle $\hat{\boldsymbol{\Theta}}^{(t)}$ à l'itération t et les données observées. Une fois ces valeurs estimées, on peut utiliser les données complètes (à la fois les observations et les estimations des variables cachées) pour mettre à jour les paramètres du modèle dans le but de maximiser la vraisemblance des données. Il est d'ailleurs important de remarquer que la méthode des k-means détaillée en section 1.2.1 peut être vue sous le prisme des méthodes de clustering basées sur un mélange gaussien, où $\boldsymbol{\Sigma}_k$ a une certaine structure (Bouveyron et al. 2019). En effet, les deux étapes des k-means peuvent être assimilées aux deux étapes de l'algorithme EM. Le clustering basé sur un modèle gaussien avec l'algorithme EM est illustré sur la Figure 1.7

1.2.1.2 Méthodes de clustering basées sur des graphes

Les méthodes de clustering basées sur des graphes se distinguent des approches traditionnelles telles que les méthodes basées sur des distances et les modèles probabilistes en exploitant les principes de la théorie des graphes. La théorie des graphes propose un ensemble de méthodes analogues au clustering, appelées méthodes de détection de communautés (Raghavan et al. 2007). Ces méthodes visent à identifier des sous-ensembles de nœuds dans un graphe qui sont plus densément connectés entre eux qu'avec le reste du graphe. Parmi ces méthodes, la plus connue, celle de Louvain, introduite par

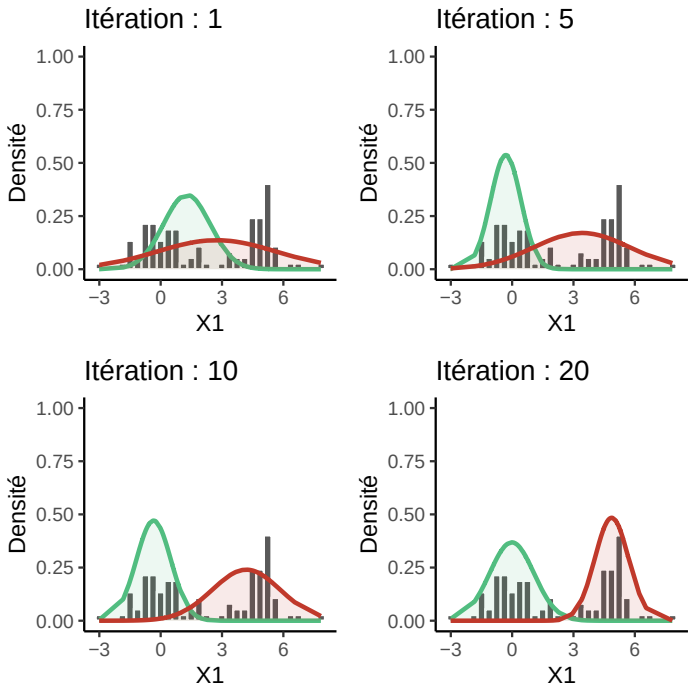


FIGURE 1.7 – **De l'initialisation à la convergence de l'algorithme EM pour le clustering basé sur un modèle de mélange gaussien.** Sur la figure, l'histogramme donne la distribution des observations. La convergence de l'algorithme EM vers les $K = 2$ clusters est illustrée à l'aide des densités intra-classe (dont la coloration se fait en fonction des deux clusters) obtenues lors des itérations de l'algorithme.

Blondel et al. (2008), repose sur la maximisation de la modularité. Il s'agit d'une mesure permettant de quantifier à quel point une partition du graphe en communautés est plus dense en termes de connexions internes (analogue à l'inertie intra-classe) qu'en termes de connexions externes (analogue à l'inertie inter-classe).

L'application des méthodes de détection de communautés pour partitionner les n observations en K clusters repose donc sur la construction, à partir de ces observations, d'un graphe. L'un des moyens les plus intuitifs est de construire un graphe des k-plus proches voisins (Eppstein et al. 1997). Cette méthode consiste à calculer pour chaque observation $\boldsymbol{x}_i$ sa distance euclidienne $d(\boldsymbol{x}_i, \boldsymbol{x}_j), i, j = 1, \ldots, n, i \neq j$ avec chacune des observations. A partir de ces distances, il est possible pour chaque observation $\boldsymbol{x}_i$ de définir son ensemble des k-plus proches voisins noté $N_k(\boldsymbol{x}_i)$ défini comme étant l'ensemble contenant les k observations $\boldsymbol{x}_j$ pour lesquelles la valeur $d(\boldsymbol{x}_i, \boldsymbol{x}_j)$ est la plus faible (Figure 1.8 haut). Dans le graphe des k-plus proches voisins associé, chaque observation $\boldsymbol{x}_i$ est donc reliée par une arête aux k observations appartenant à $N_k(\boldsymbol{x}_i)$ (Figure 1.8 en bas à gauche). Il est alors possible d'appliquer n'importe quel algorithme de détection de communauté sur ce graphe pour estimer K communautés représentant les K clusters (Figure 1.8 en bas à droite).

Ces méthodes de clustering basées sur des graphes, implémentées dans le paquet `R Seurat`

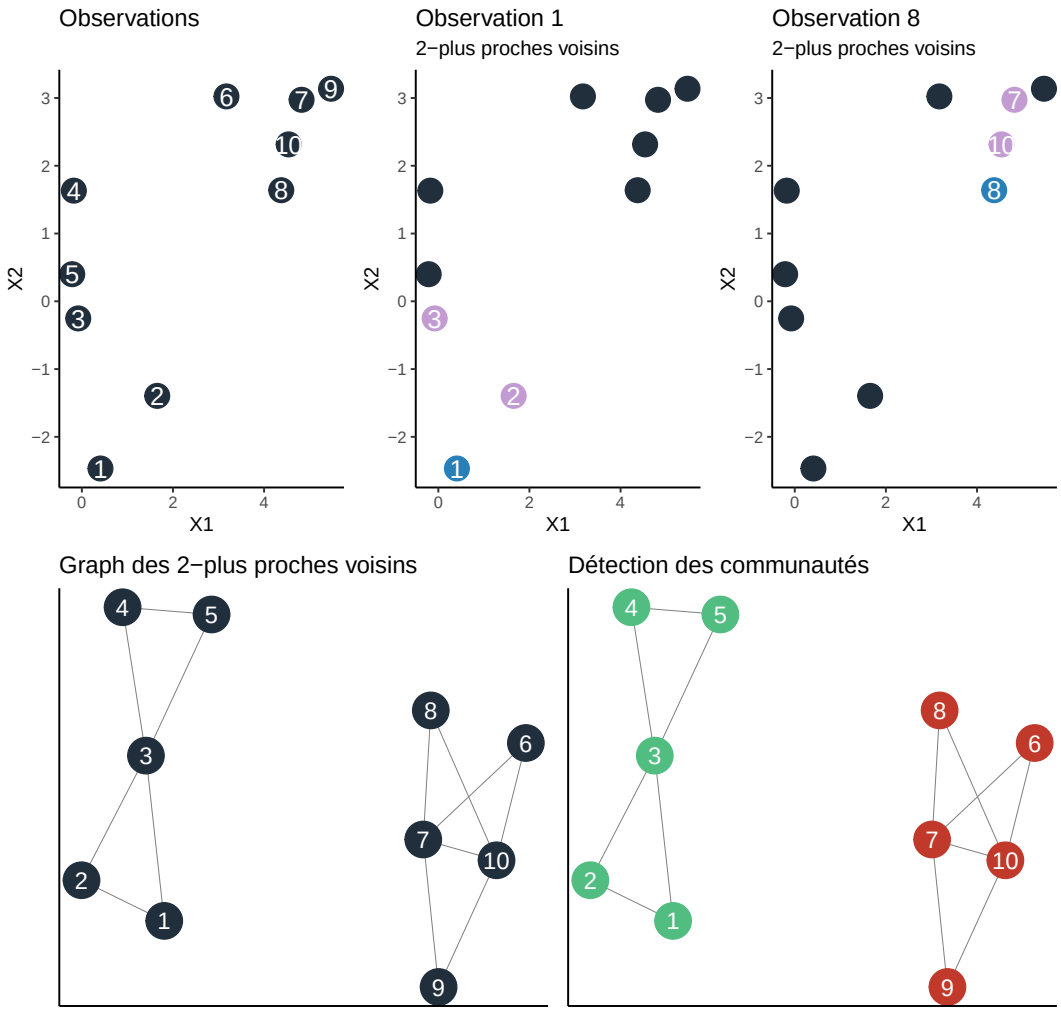


FIGURE 1.8 – **De la création d'un graphe des $k = 2$-plus proches voisins à la détection de communautés.** Sur la figure, les points correspondent aux observations. Les 2-plus proches voisins des observations 1 et 8 en rose sont représentés en bleu. Une fois le graphe des 2-plus proches voisins construit, les $K = 2$ communautés détectées sont données en rouge et en vert.

(Hao et al. 2023) très utilisées en bioinformatique, sont celles traditionnellement utilisées dans le cadre des analyses de données scRNA-seq pour le clustering de cellule.

1.2.2 L'analyse différentielle de l'expression génique

L'analyse différentielle constitue une étape courante des analyses de données RNA-seq. Son objectif principal est d'identifier les gènes dont l'expression varie significativement entre différents groupes d'observations, typiquement associés à des conditions expérimentales distinctes. On dit alors que ces gènes sont différentiellement exprimés entre les conditions expérimentales. Cette différence d'expression génique permet ensuite une meilleure compréhension des processus biologiques impliqués en réponse à différents traitements ou vaccins. Les méthodes d'analyse différentielle s'inscrivent dans le cadre des

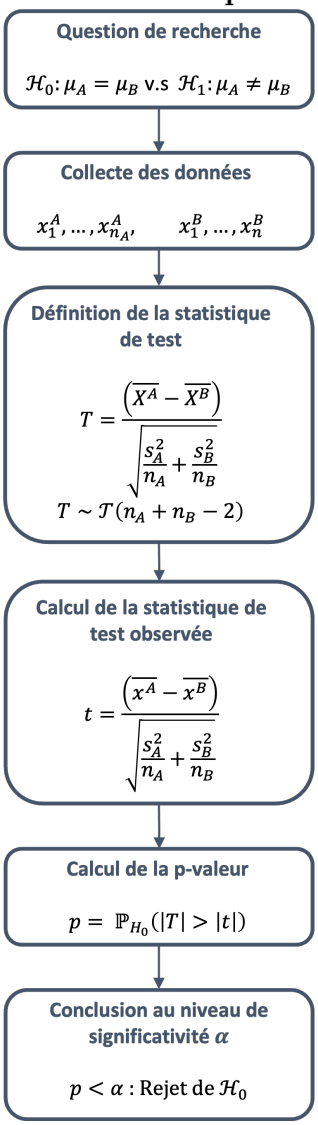
tests d'hypothèses statistiques.

1.2.2.1 Rappels sur les tests statistiques

Cadre théorique des tests d'hypothèses

Les tests d'hypothèses (ou tests statistiques) sont des outils statistiques permettant d'inférer, à partir d'un ou plusieurs échantillons, la validité d'une hypothèse. Ils aident alors à répondre à une question de recherche donnée, par exemple, cela peut être d'évaluer l'effet d'un traitement sur une maladie. Cette question de recherche doit alors être traduite et formalisée en une hypothèse à tester. Un test statistique nécessite alors deux types d'hypothèses : $\mathcal{H}_0$ (l'hypothèse nulle) et $\mathcal{H}_1$ (l'hypothèse alternative). Elles sont le fondement de toute analyse statistique faisant appel à des tests statistiques, car elles énoncent les affirmations à évaluer et à tester. $\mathcal{H}_0$ traduit couramment l'absence d'effet, de différence et est donc formulée comme une égalité à une certaine valeur ou une égalité entre deux groupes (« le traitement n'a pas d'effet sur la maladie »). $\mathcal{H}_1$ correspond elle à l'affirmation contraire, traduisant généralement un effet ou une différence. Une fois les hypothèses du test posées, il faut collecter des données pour y répondre. Pour tester l'efficacité d'un traitement, il est courant de mesurer certains biomarqueurs dans deux groupes de patients : un groupe de patients traités et un groupe de patients contrôles qui ne recoivent pas le traitement. On pose alors $\mathcal{H}_0 : \mu_{\text{traité}} = \mu_{\text{contrôle}}$ où $\mu_{\text{traité}}$ et $\mu_{\text{contrôle}}$ dénotent respectivement les moyennes du biomarqueur dans le groupe traité et le groupe contrôle, alors que l'hypothèse alternative peut se traduire par $\mathcal{H}_1 : \mu_{\text{traité}} \neq \mu_{\text{contrôle}}$.

FIGURE 1.9 – **L'exemple du test *t***



En fonction du type de données, du type d'étude et des hypothèses à tester, il faut choisir le test statistique le plus adapté. Il donne alors le cadre théorique et permet de définir

les propriétés de la statistique de test sous $\mathcal{H}_0$, c'est-à-dire dans le cadre hypothétique où $\mathcal{H}_0$ est vérifiée. La statistique de test T est une fonction de variables aléatoires dont la distribution de probabilité est connue sous $\mathcal{H}_0$. Il est ensuite nécessaire de calculer la valeur observée de la statistique de test observée t, c'est-à-dire, calculer la statistique de test en se basant sur les données mesurées. La conclusion du test se fait alors en calculant la p-valeur définie comme étant la probabilité sous $\mathcal{H}_0$ que la statistique de test observée soit au moins aussi extrême que la statistique de test théoriquement observée si $\mathcal{H}_0$ était vérifiée, c'est-à-dire $\mathbb{P}_{\mathcal{H}_0}(|T| > |t|)$. Si cette p-valeur est très petite, inférieure à un niveau de significativité α, cela signifie qu'il n'y a qu'une très faible probabilité que ce que l'on observe soit réellement observable si $\mathcal{H}_0$ était vérifiée. On rejette alors l'hypothèse nulle $\mathcal{H}_0$ en faveur de l'hypothèse alternative $\mathcal{H}_1$. Par exemple, pour tester l'égalité des moyennes du biomarqueur entre le groupe traité et le groupe contrôle de l'exemple, un test adapté est le test t de Student (Welch 1947) dont le cadre est résumé en Figure 1.9.

Formellement, le niveau de significativité α utilisé pour tirer une conclusion lors d'un test d'hypothèse représente la probabilité de rejeter l'hypothèse nulle alors qu'elle est en réalité vraie. Cette probabilité, notée $\mathbb{P}_{\mathcal{H}_0}(\text{rejeter } \mathcal{H}_0)$, est ce que l'on appelle le risque de première espèce. L'erreur de type I correspond à l'événement consistant à rejeter $\mathcal{H}_0$ alors qu'elle est vraie. Autrement dit, c'est une conclusion incorrecte qui survient lorsque l'on déclare une différence ou un effet statistiquement significatif alors qu'il n'existe pas réellement. Fixer une valeur de α à un certain seuil pour conclure permet de définir la probabilité maximale acceptable de commettre une erreur de type I. En d'autres termes, α est le seuil de risque que l'on est prêt à accepter pour rejeter $\mathcal{H}_0$ à tort.

Le risque de seconde espèce, noté β, correspond à la probabilité de ne pas rejeter $\mathcal{H}_0$ alors qu'elle est fausse. L'erreur de type II associée correspond alors à l'évènement consistant à ne pas rejeter $\mathcal{H}_0$ alors qu'elle est fausse, c'est-à-dire, manquer la détection d'un effet ou d'une différence réelle dans les données. Le risque de seconde espèce est inversement lié à la puissance statistique du test. Cette dernière mesure la capacité du test à détecter un effet réel lorsque celui-ci existe. Elle se définit alors comme $\mathbb{P}_{\mathcal{H}_1}(\text{rejeter } \mathcal{H}_0) = 1 - \beta$. Une puissance élevée signifie donc une probabilité faible d'erreur de type II. La puissance statistique dépend de plusieurs facteurs, notamment la taille de l'effet, la taille de l'échantillon et le niveau de significativité α choisi pour le test. Le choix du niveau de significativité est donc important car en plus de définir la probabilité à laquelle on accepte de rejeter à tort l'hypothèse nulle, il va également impacter la puissance du test. Fixer α trop bas permet de réduire le risque d'erreur de type I, mais également la puissance du test et inversement. En pratique, le niveau de significativité est souvent choisi en fonction du domaine d'application : en biostatistique il est courant de prendre $\alpha = 5\%$.

Multiplicité des tests

Comme expliqué en Section 1.1.4, la grande dimension inhérente aux données d'expression génique pose des problèmes liés à la multiplicité des tests. En effet, lorsque les tests d'hypothèses sont appliqués marginalement sur plusieurs variables, les erreurs de type I de chacun des tests se cumulent entre elles. Il devient alors nécessaire d'utiliser des méthodes de correction qui permettent de tenir compte de cette multiplicité des tests pour ainsi contrôler le risque global d'erreur de type I. Plusieurs méthodes de correction existent, s'appuyant chacune sur une généralisation de l'erreur de type I.

L'une des plus connue est la méthode de Bonferonni introduite en Section 1.1.4. Formellement, il s'agit d'une méthode permettant de contrôler le *Family Wise Error Rate* (FWER), défini comme la probabilité que, dans un ensemble de m tests effectués où toutes les hypothèses nulles sont vraies, au moins un test conduise à un rejet à tort de l'hypothèse nulle. Un moyen de controler le FWER au niveau α consiste à réduire le seuil de significativité de chaque test marginal à $\alpha' = \frac{\alpha}{m}$. En effet, si chacun des tests contrôle bien l'erreur de type I au niveau α', alors $\mathbb{P}_{\mathcal{H}_0}(\text{ne pas rejeter } \mathcal{H}_0) = 1 - \alpha'$ pour chaque test. Si l'ensemble des m tests sont appliqués de manière indépendante, alors la probabilité que les m tests n'aboutissent pas à un rejet de $\mathcal{H}_0$ sachant que $\mathcal{H}_0$ est vraie vaut simplement le produit de ces probabilités individuelles, soit : $(1-\alpha')^m$. Ainsi, la probabilité inverse qu'au moins un test conclu à tort au rejet de $\mathcal{H}_0$ vaut $1-(1-\alpha')^m$. Donc :

$$\begin{aligned} \text{FWER} &< 1 - (1-\alpha')^m \\ &= 1 - \left(1 - \frac{\alpha}{m}\right)^m \xrightarrow[m\to\infty]{} \alpha. \end{aligned}$$

L'une des principales limites des méthodes comme celle de Bonferroni, qui cherchent à contrôler le FWER, est le fait qu'elles sont très conservatives, c'est-à-dire qu'elles ont tendance à rejeter $\mathcal{H}_0$ de manière très prudente, avec une puissance statistique limitée (Nakagawa 2004).

Une autre généralisation de l'erreur de type I est le Taux de Fausses Découvertes (FDR pour *False Discovery Rate*). Formellement, le FDR est défini comme l'espérance du rapport entre le nombre de fausses découvertes (d'hypothèses nulles rejetées à tort) et le nombre total de découvertes significatives (d'hypothèses nulles rejetées). L'une des méthodes les plus connues garantissant le contrôle du FDR est la méthode de Benjamini-Hochberg (Benjamini & Hochberg 1995). Il s'agit de la méthode la plus utilisée dans le cadre de l'analyse différentielle de données d'expression génique. Elle se décompose en plusieurs étapes :

1. Ordonner les m p-valeurs par ordre croissant $p_{[1]}, \ldots, p_{[m]}$ et soit $\mathcal{H}_{0,[1]}, \ldots, \mathcal{H}_{0,[m]}$ les m hypothèses nulles associées
2. Identifier l'indice i^* du plus grand rang tel que $p_{[i^*]} \leq \frac{i^*}{m}\alpha$

3. Rejeter toutes les hypothèses nulles $\mathcal{H}_{0,[1]}, \ldots, \mathcal{H}_{0,[i^*]}$

Les méthodes de correction basées sur le FDR sont moins conservatives que celles basées sur le FWER, et en particulier, si le FWER est contrôlé au niveau α alors le FDR l'est aussi (Shaffer 1995).

1.2.2.2 Tester l'expression différentielle

Les groupes d'observations, déterminés par les conditions expérimentales dont elles sont issues, servent à formuler les hypothèses nulles à tester, qui consistent généralement en l'absence de différence d'expression entre les groupes comparés. Cependant, en raison de leurs caractéristiques intrinsèques, la définition de l'expression différentielle se traduit de manière différente selon que l'on considère des données RNA-seq en masse ou des données scRNA-seq. Dans le contexte des données RNA-seq, l'expression différentielle d'un gène entre deux conditions expérimentales se traduit principalement par une différence de moyennes d'expression du gène entre ces conditions. En revanche, les données scRNA-seq présentent une plus forte hétérogénéité en raison de leur résolution très fine, ce qui rend la caractérisation de l'expression différentielle plus délicate. Dans ce contexte, l'expression différentielle est plutôt définie par une différence de distribution de l'expression du gène entre les deux conditions (Korthauer et al. 2016). Cette définition englobe ainsi une large palette de possibilités, allant d'une simple différence de moyenne entre les conditions, comme c'est déjà le cas pour les données RNA-seq en masse (Figure 1.10 gauche), à des profils d'expression plus complexes (Figure 1.10 droite). Ainsi chacun des deux types de données possèdent sa propre panoplie d'outils d'analyse différentielle.

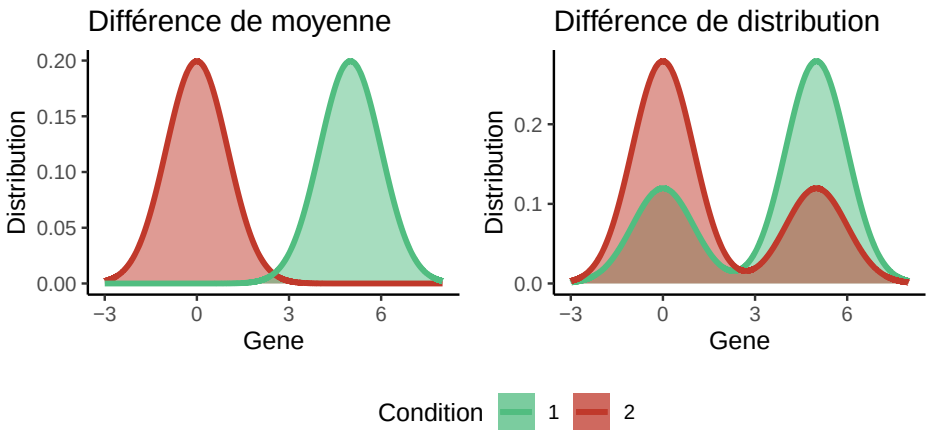


FIGURE 1.10 – **Différents types d'expression différentielle entre deux conditions** (inspirée de Korthauer et al. (2016)).

Trois méthodes d'analyse différentielle pour données RNA-seq en masse ressortent particulièrement de la littérature. Il s'agit des méthodes `edgeR` (Robinson et al. 2010)

(36 594 citations sur Google Scholar en juin 2024) , `DESeq2` (Love et al. 2014) (6 6104 citations sur Google Scholar en juin 2024) et `limma-voom` (Law et al. 2014) (5 580 citations sur Google Scholar en juin 2024). `edgeR` et `DESeq2` utilisent toutes les deux une distribution binomiale négative pour modéliser la distribution de chaque gène alors que `limma-voom`, s'appuie sur une hypothèses gaussienne couplée à une modélisation de l'hétéroscédasticité. Des méthodes sans hypothèses distributionnelles, telles que `dearseq` (Gauthier et al. 2020) ou le test de Wilcoxon (Li et al. 2022), ont également été proposées et ont démontré une meilleure performance que les méthodes traditionnelles en termes de contrôle du taux de fausses découvertes. Concernant l'analyse des données scRNA-seq, il n'existe pas encore de consensus sur des méthodes spécifiques d'analyse différentielle. Parmi les méthodes les plus couramment utilisées, on retrouve celles implémentées dans `Seurat`, notamment le test de Wilcoxon, ainsi que `MAST` (Finak et al. 2015). Cette dernière méthode repose sur un modèle linéaire généralisé en deux parties : une première partie dichotomise l'expression de chaque gène en une variable binaire (exprimée ou non), tandis que la seconde partie modélise l'expression des gènes exprimés à l'aide d'une distribution gaussienne. D'autres méthodes existent indépendemment de `Seurat` telles que `DEsingle` (Miao et al. 2018) qui se base sur un modèle binomiale négatif enflé en 0 ou encore la méthode SCDD (Korthauer et al. 2016) qui teste l'analyse différentielle d'un gène entre deux conditions en testant si sa distribution est significativement liée à cette condition.

1.2.3 Clustering & analyse différentielle dans l'analyse de données scRNA-seq

L'analyse des données scRNA-seq s'organise généralement autour de plusieurs étapes clés qui font consensus dans la communauté bioinformatique (Hwang et al. 2018, Amezquita et al. 2020, Villani et al. 2017, Pasquini et al. 2020, Sokal et al. 2021). Une première étape cruciale est le contrôle qualité (QC), qui cherche à évaluer la qualité des données afin d'éliminer les cellules de qualité médiocre qui pourraient perturber les analyses. Ensuite vient l'étape de la normalisation des données (Oshlack et al. 2010, Wagner et al. 2012), indispensable pour rendre les échantillons comparables entre eux tout en préservant les différences biologiques, en éliminant les variations techniques potentielles entre les cellules. Les méthodes de normalisation varient en fonction du contexte et des outils qui sont utilisés pour les analyses. La sélection de caractéristiques ou de gènes pertinents constitue également une étape importante. L'objectif est de réduire en amont des analyses la dimensionnalité des données pour conserver une plus grande puissance statistique. Là encore, plusieurs méthodes existent, mais les plus connues visent à conserver les gènes les plus variables ou les plus éloignés d'une hypothèse d'expression constante entre toutes les cellules (Townes et al. 2019). Étant donné la grande dimension des données, une réduction de dimension comme détaillée en Section 1.1.4 est couramment appliquée sur les données

afin de surmonter le fléau de la dimension en résumant l'information contenue dans les données en un nombre restreint de dimension. Cela permet de faciliter la suite des analyses. À noter qu'il est également courant d'utiliser des méthodes de visualisation pour représenter les données en deux dimensions. Bien que très ressemblantes, la réduction de dimension et la visualisation ont des finalités différentes. En effet, la visualisation consiste à projeter les données originales de dimension élevée dans un sous-espace de dimension fortement réduite (2 ou 3) afin de pouvoir les représenter par un nuage de points. Il s'agit donc d'une réduction de dimension pour laquelle l'information est très fortement condensée. Parmi les méthodes de visualisation les plus connues, on trouve t-SNE (Van der Maaten & Hinton 2008) et UMAP (McInnes et al. 2018) qui ont toutes deux pour ambition de préserver la structure locale des données, c'est-à-dire de placer deux cellules qui se ressemblent proches l'une de l'autre sur le plan. Cependant, il a récemment été démontré par Chari & Pachter (2023) que ces méthodes de visualisation, qui réduisent très fortement la dimension des données pour ne construire que deux composantes, pouvaient conduire à des distortions de l'information.

La suite des analyses dépend des questions biologiques auxquelles on souhaite répondre. Un clustering des cellules est souvent appliqué sur les données pour construire des sous-groupes de cellules ayant des profils d'expression génique similaires. Ces clusters sont alors généralement interprétés comme représentant différentes sous-populations cellulaires. L'objectif principal devient alors de labelliser chaque cluster avec une sous-population cellulaire spécifique (d'annoter chaque cluster). Cela nécessite l'identification des gènes différentiellement exprimés entre les sous-groupes de cellules. Ces gènes, souvent appelés gènes marqueurs, jouent un rôle essentiel dans l'annotation de chaque cluster, car ils sont caractéristiques de la sous-population cellulaire correspondante. En effet, chaque sous-population cellulaire possède son propre ensemble distinct de gènes exprimés exclusivement au sein de cette population, pouvant donc servir à les annoter (Pasquini et al. 2021). Ainsi, le clustering et l'analyse différentielle sont deux étapes majeures de l'analyse des données scRNA-seq. Les résultats du clustering servent de condition à tester lors de l'analyse différentielle rendant alors ces deux étapes interdépendantes.

1.2.4 Le défi de l'inférence post-clustering

Le pipeline d'analyse des données scRNA-seq contredit la méthodologie traditionnelle des tests d'hypothèses définie en Section 1.2.2.1 puisque les résultats du clustering servent à formuler les hypothèses de test de l'analyse différentielle. La question de recherche n'est donc plus formulée en amont de la collecte de données comme il se devrait, mais à partir même des données collectées. On parle alors d'inférence post-clustering. Comme l'illustre la Figure 1.11, les données sont donc utilisées deux fois. Grâce au clustering, elles servent une première fois à définir les conditions expérimentales à tester en construisant K

clusters $C_1, \ldots, C_K$. Ces mêmes données sont ensuite ré-utilisées durant l'étape d'analyse différentielle pour tester l'expression différentielle de chaque gène entre chacun des K clusters construits. De cette double utilisation des données, aussi appelée *double dipping* (Kriegeskorte et al. 2009), en contradiction avec le cadre traditionnel des tests statistiques, découlent plusieurs problématiques qui compromettent les propriétés attendues des tests d'hypothèses et donc des méthodes d'analyse différentielle.

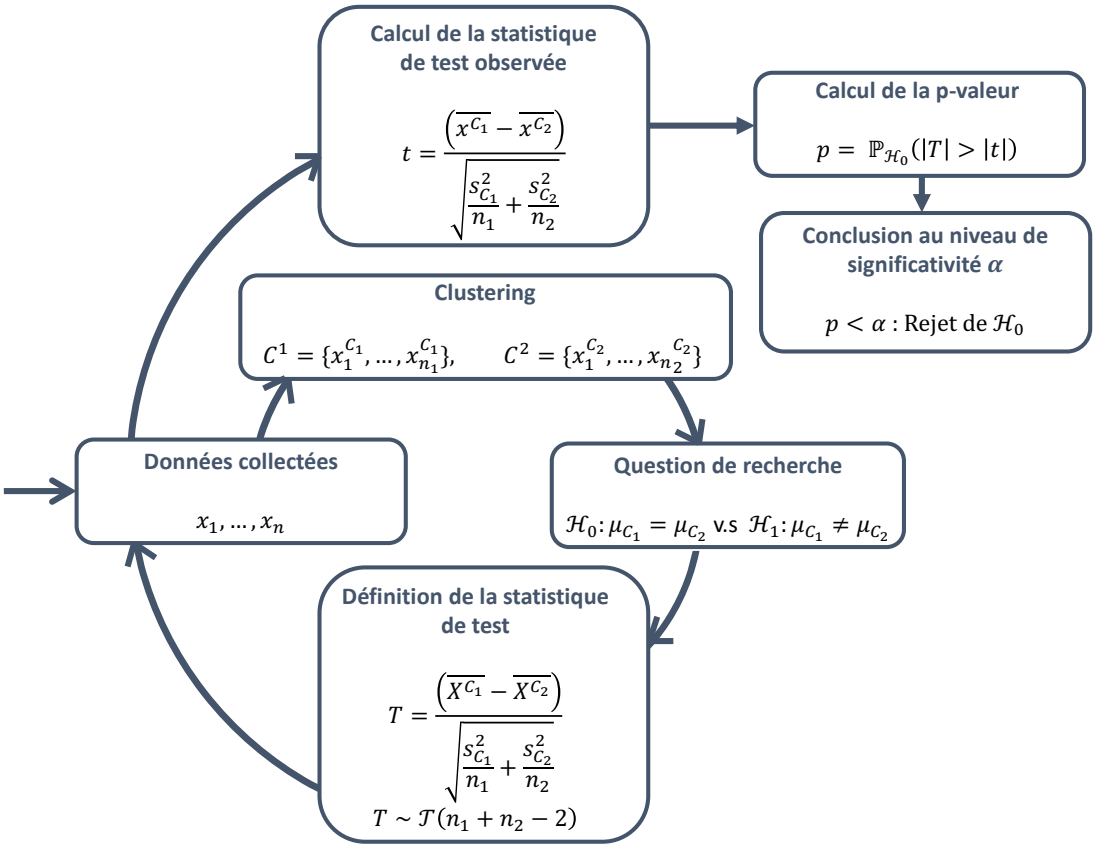


FIGURE 1.11 – **La double utilisation des données dans le cadre de l'inférence post-clustering.**

1.2.4.1 Incertitudes sur le clustering et différences artificielles

Le clustering regroupe un ensemble de méthodes non supervisées qui permettent de partitionner les observations en sous-groupes homogènes et séparés. Cependant, il est impossible de connaître, s'il en existe une, la véritable partition des observations, ce qui engendre plusieurs sources d'incertitudes dans les résultats du clustering. Tout d'abord, il y a une incertitude concernant l'assignation d'une observation à un cluster spécifique. Effectivement, il est impossible d'affirmer avec certitude qu'une observation est correctement labelisée. Certaines méthodes de clustering permettent de quantifier directement cette incertitude ; par exemple, les méthodes de clustering basées sur un modèle probabiliste grâce aux probabilités a posteriori d'appartenance à un cluster C_k définies en

(1.2). Des approches statistiques ont été développées pour prendre en compte ces incertitudes dans les analyses dépendantes des résultats du clustering (Vermunt 2010, Bakk et al. 2014). Mais, l'incertitude concernant l'assignation des observations n'est pas la seule source d'incertitude découlant des résultats du clustering ; il existe également une incertitude concernant le nombre de clusters à construire à partir des données (Suzuki & Shimodaira 2006).

Effectivement, il est toujours possible de partitionner les observations en K clusters, même en l'absence d'un processus latent justifiant ce partitionnement. Dans de tels cas, l'algorithme de clustering exagère les différences entre les observations pour former les clusters, comme illustré dans la Figure 1.12. Dans cet exemple jouet, nous avons simulé $n = 200$ observations provenant d'une même distribution gaussienne centrée-réduite. Il est évident qu'aucun cluster n'est présent dans les données au sens de la définition énoncée en Section 1.2.1 (panneau A). Si l'on aborde la question en termes d'expression différentielle, en considérant que nos observations reflètent l'expression d'un gène, il est évident que la distribution des données ne révèle pas d'expression différentielle entre deux conditions comme c'était le cas sur la Figure 1.10. Néanmoins, il est possible de partitionner ces observations en $K = 2$ clusters, comme le montre le panneau B. Ces clusters ont été construits en utilisant le clustering hiérarchique avec la mesure de Ward, garantissant ainsi une homogénéité maximale (minimisant l'inertie intra-classe) et une séparation optimale (maximisant l'inertie inter-classe). Cependant, l'homogénéité et la séparabilité observées ici résultent uniquement du partitionnement des observations en 2 clusters, et non d'un processus latent, d'une réelle condition, expliquant la présence de ces 2 clusters. En d'autres termes, il n'y a rien d'autre que la méthode de clustering elle-même qui permet d'expliquer les ressemblances et les différences entre ces clusters. Il est donc important de garder en tête que l'incertitude sur le nombre de classes peut aboutir à des différences artificielles entre les clusters estimés.

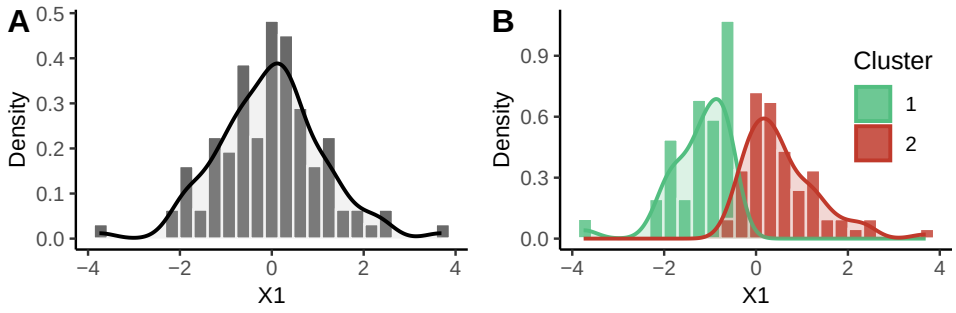


FIGURE 1.12 – **Distribution des données avant le clustering (panneau A) et après le clustering (panneau B).**

1.2.4.2 Inférence post-clustering et inflation de l'erreur de type I

L'importance des incertitudes associées aux résultats du clustering se manifeste pleinement dans le contexte de l'inférence post-clustering. Étant donné que les hypothèses de test sont formulées à partir des clusters, eux-mêmes dérivés des données, et que les tests sont effectués à partir des mêmes données, il devient impossible de distinguer les effets relevant d'un réel processus latent des potentiels artefacts du clustering lui-même. Reprenons l'exemple de la Figure 1.12. Comme on sait qu'il n'existe pas de réel processus latent permettant de partitionner les observations en deux clusters distincts, alors les p-valeurs devraient être uniformément distribuées (car sous $\mathcal{H}_0$) lorsqu'on effectue un test de différence de moyenne tel que le test t de Student entre les deux clusters estimés. Néanmoins, comme le présente la Figure 1.13, la fonction de répartition empirique des p-valeurs calculées pour 1 000 simulations des données est très éloignée de la fonction de répartition de la distribution Uniforme sur $[0, 1]$. Il est évident que dans ce cas-là, les différences identifiées par le test t de Student sont celles artificiellement créées par le clustering pour construire les deux clusters. Dans ce cas-là, bien que significatifs, ces résultats ne sont pas pertinents : ils ne sont que le produit d'un mauvais clustering, mais ils ne reflètent en rien le processus sous-jacent à la génération des données.

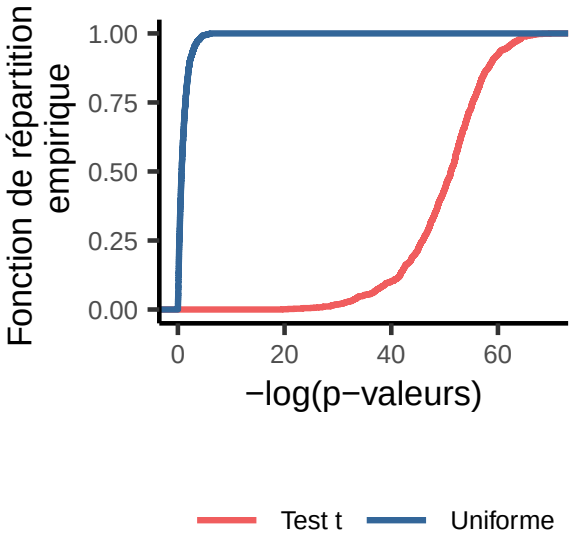


FIGURE 1.13 – **Fonction de répartition empirique des p-valeurs du test t de Student par rapport à la fonction de répartition d'une loi Uniforme sur $[0, 1]$.**

En réalité, comme les hypothèses de tests reposent sur les résultats du clustering et qu'elles sont testées sur les mêmes données, ce ne sont plus des hypothèses, mais déjà des affirmations. Si deux clusters sont construits à partir des données, c'est qu'ils peuvent être différenciés l'un de l'autre, ne serait-ce que parce qu'ils sont homogènes et séparés par définition. Ces différences sont alors ensuite facilement identifiables par

les tests d'hypothèses. Ce n'est donc plus tant la différence entre les clusters qui nous intéresse, mais plutôt de connaître l'origine de cette différence : est-ce un réel processus latent ou le simple fait qu'un utilisateur ait partitionné les observations en clusters ? C'est dans ce dernier cas que l'on peut dire que l'inférence post-clustering conduit à une inflation de l'erreur de type I, car les différences identifiées ne sont que des artefacts du mauvais clustering.

1.3 Organisation de la thèse

Dans ces pipelines où les données sont utilisées deux fois, comme celles couramment mises en oeuvre pour analyser les données scRNA-seq, il est donc impossible de conclure qu'une différence significative entre deux clusters est bien le résultat d'un processus latent, tout simplement car il est impossible de savoir si la partition identifiée par le clustering reflète bien le réel processus sous-jacent aux données. Conjointement à ces fortes incertitudes quant aux résultats du clustering vient s'ajouter la réutilisation des données pour effectuer les tests d'hypothèses entre clusters soulevant les problématiques de *double-dipping* et d'inférence post-clustering. Les tests dépendant du clustering peuvent alors identifer les différences artificielles émanant d'un clustering en un trop grand nombre de sous-groupes, conduisant à une inflation de leur erreur de type I. Motivé par l'analyse de données scRNA-seq, ce travail de thèse s'efforce donc de relever les défis liés à l'inférence post-clustering dans un contexte où la littérature sur cette problématique reste limitée.

Étant donné le lien étroit entre mauvais clustering des observations et création de différences artificielles, il est crucial de commencer à examiner cette question sous l'angle de la performance du clustering. Bien que la double utilisation des données reste discutable, un clustering approprié des observations permet de réduire les incertitudes, limitant ainsi le risque de créer des différences artificielles entre les clusters. Dans le Chapitre 2, nous présentons alors les principaux défis liés au clustering des données RNA-seq, en mettant en évidence l'impact de la standardisation des données sur la qualité du clustering. Dans un second temps, nous nous penchons sur la problématique de la double utilisation des données de manière rigoureuse. Dans le Chapitre 3, nous dressons un état de l'art des méthodes d'inférence post-clustering. Nous exposons ensuite, dans le Chapitre 4, nos méthodes d'inférence post-clustering implémentées dans le paquet R `VALIDICLUST`, qui ont démontré de bonnes performances en petite dimension. Enfin, dans le Chapitre 5, nous explorons la complexité du problème d'inférence post-clustering, soulignant la nécessité de connaître les vraies classes en amont du clustering, pour appliquer efficacement les méthodes existantes.

Chapitre 2

Centrer-réduire les données RNA-seq ou non en amont du clustering ?

Contenu

2.1 Introduction 43
2.2 Standardiser les données pour le clustering ? 46
2.2.1 Cadre général & Définition 46
2.2.2 La standardisation dans le cadre des données RNA-seq 48
2.3 Méthodes 50
2.3.1 Étude de simulations 50
2.3.2 Analyse de données réelles 51
2.3.3 Critère d'évaluation 52
2.4 Résultats 53
2.4.1 Étude de simulations 53
2.4.2 Analyse de données réelles 55
2.5 Discussion 56

2.1 Introduction

Les données RNA-seq subissent un ensemble complexe d'étapes de prétraitement, comme expliqué en détail dans la Section 1.2.3. Toutes ces étapes sont essentielles pour assurer la pertinence des résultats, en rendant les échantillons comparables entre eux, ou permettent simplement de faciliter les analyses grâce à la réduction de dimension et à la sélection de caractéristiques. Néanmoins, chacune d'entre elles exerce un impact significatif sur les résultats, notamment en ce qui concerne le clustering (Wang et al. 2020). De plus, pour une même étape du pipeline, il existe parfois un grand nombre de méthodes applicables qui sont, à leur tour, susceptibles d'influencer les résultats du clustering. La méthode de clustering elle-même, qu'elle soit spécifique ou non à l'analyse des données d'expression génique, engendre une variabilité potentielle dans les résultats (Duò et al. 2018). Ainsi, chaque choix effectué durant les analyses peut impacter les résultats du clustering, et plus largement, toutes analyses qui reposent sur ses résultats. Comprendre

l'impact de chaque étape de prétraitement et déterminer la combinaison optimale de méthodes permettant de garantir un clustering représentatif du véritable processus sous-jacent aux données devient donc un enjeu important, et pourtant peu étudié, de l'analyse des données RNA-seq.

Le contrôle qualité et la normalisation sont deux étapes indispensables dans les analyses de données RNA-seq et particulièrement dans le cadre du clustering. En effet, elles permettent toutes deux d'éviter de potentiels biais dans les résultats du clustering. Premièrement, le contrôle qualité s'efforce de repérer et d'éliminer les observations de faibles qualités pour éviter un partitionnement des données qui serait expliqué uniquement par la qualité des observations (Luecken & Theis 2019). L'influence du contrôle de qualité sur les résultats du clustering demeure relativement peu étudiée. Ainsi, Luecken & Theis (2019) recommandent de ré-analyser les données en ajustant les critères de contrôle qualité en cas de résultats peu satisfaisants, car il peut s'avérer que ce dernier se soit avéré trop permissif. La normalisation des données, en plus de rendre comparables les observations entre elles, permet de supprimer des sources de variations non désirées (typiquement de la variabilité technique). Cela a pour effet de faciliter un partitionnement expliqué par des processus biologiques (Molania et al. 2023). Cependant, il n'y a pas une unique méthode de normalisation garantissant des résultats optimaux en termes de clustering, et le choix de cette méthode de normalisation dépend des données. Cole et al. (2019) ont introduit un outil, `scone` permettant d'évaluer plusieurs méthodes de normalisation afin de conserver celles engendrant les meilleurs résultats.

Les données normalisées sont généralement soumises à une transformation, souvent par une fonction de type $\log(\cdot + 1)$. Cette étape permet d'atténuer la forte variabilité observée dans les données, en particulier, en stabilisant la relation moyenne-variance. De plus, elle contribue à corriger l'asymétrie fréquemment constatée dans la distribution de l'expression génique, ce qui rend ces données compatibles avec les hypothèses distributionnelles de nombreux outils d'analyse qui reposent sur une distribution gaussienne. Enfin, cette transformation permet d'interpréter les différences d'expression entre deux conditions en termes de fold-changes (Luecken & Theis 2019). Malgré sa grande utilisation, l'impact de cette transformation, notamment sur les performances du clustering, demeure peu étudié (Wang et al. 2020, Risso & Pagnotta 2021). Dans une étude, Jaskowiak et al. (2018) ont démontré que la transformation $\log 2$ était bénéfique sur les performances du clustering par rapport à une absence de transformation des données. De même, Wang et al. (2020) ont confirmé la supériorité de cette transformation logarithmique à la fois par rapport à l'absence de transformation, mais également par rapport à d'autres types de transformations.

Une étape du pipeline connue pour son impact significatif sur les performances du clustering est la sélection des caractéristiques, *i.e* des gènes dans le cas des données RNA-seq. Initialement utilisée pour réduire la dimensionnalité des données, cette étape permet non

seulement de diminuer le bruit contenu dans les données, mais également d'accélérer les temps de calcul des méthodes de clustering (Kiselev et al. 2019). Les travaux de Källberg et al. (2021) ont souligné l'importance de cette étape en démontrant que conserver tous les gènes pour le clustering entraînait une dégradation des performances de celui-ci. Un large éventail de méthodes de sélection de caractéristiques existe, et l'impact de ce choix sur le clustering a déjà été mis en évidence par Duò et al. (2018). Outre la méthode elle-même, le nombre de caractéristiques à conserver est également un facteur déterminant dans le clustering (Luecken & Theis 2019). Bien que Jaskowiak et al. (2018) aient observé une tendance favorable à conserver un nombre restreint de gènes pour le clustering (de l'ordre du millier), les études de Vidman et al. (2019) et Källberg et al. (2021) ont toutes deux révélé des performances variables des méthodes de sélection de gènes, rendant difficile la préconisation d'une méthode spécifique. Leur conclusion générale est que le choix de la méthode et du nombre de gènes doit être adapté aux caractéristiques propres des données analysées. De manière générale, Luecken & Theis (2019) recommandent de ne conserver qu'entre 1 000 et 5 000 gènes pour le clustering selon la complexité du jeu de données.

Les performances des méthodes de clustering pour l'analyse des données RNA-seq ont déjà été largement étudiées. Une comparaison équitable de ces méthodes nécessite des étapes de prétraitement similaires afin d'isoler l'impact propre à chaque méthode de clustering. Jaskowiak et al. (2018) ont étudié les performances de différents algorithmes de clustering basés sur des distances, en explorant 12 distances, incluant la distance euclidienne classique, et recommandent alors l'utilisation de distances basées sur des mesures de corrélations. Leur analyse a également mis en évidence de faibles performances de la classification ascendante hiérarchique utilisant la mesure d'agrégation minimale (*single link*) sur les données RNA-seq. D'autre part, Duò et al. (2018) ont comparé 14 méthodes de clustering dédiées ou non à l'analyse de données scRNA-seq, révélant une grande variabilité dans leurs performances pour retrouver les partitions connues des données, performances largement influencées par le degré de séparation entre les groupes. En général, la méthode Louvain implémentée dans `Seurat` et présentée en Section 1.2.1.2, se distingue parmi les méthodes de clustering, conduisant généralement à de meilleures performances. Elle est alors celle recommandée par Luecken & Theis (2019).

Parmi toutes les étapes de l'analyse des données RNA-seq, une étape a été largement moins étudiée que les autres : la standardisation. Elle implique généralement la transformation de l'expression génique pour qu'elle présente une distribution homogène, souvent en la centrant et en la mettant à l'échelle. Cette pratique, courante pour toutes analyses de clustering, n'est pas spécifique aux données RNA-seq. Elle est utilisée en amont du clustering pour assurer que toutes les variables contribuent équitablement à la formation des groupes et éliminer les artefacts dus aux écarts d'échelle entre les variables. Cependant, il n'y a toujours pas de consensus sur la nécessité de standardiser les données RNA-seq en amont du clustering (Luecken & Theis 2019). Nous nous efforcerons alors dans ce Chapire

d'expliquer pourquoi, bien que fréquemment effectuée, la standardisation est discutable dans le cadre des analyses de données RNA-seq. À travers une combinaison d'études de simulations et d'analyses de données d'expression génique réelles, nous chercherons à éclairer l'impact de cette étape sur les résultats du clustering.

2.2 Standardiser les données pour le clustering ?

2.2.1 Cadre général & Définition

La standardisation des variables en amont du clustering cherche principalement à attribuer le même poids à toutes les variables. En effet, les variables peuvent être mesurées sur des échelles différentes, ce qui les rend incomparables entre elles à cause de leur magnitude ou de leurs échelles de mesure différentes (Anderberg 1973). Si les données ne sont pas standardisées avant le clustering, les algorithmes de clustering, et en particulier ceux basés sur les distances, peuvent être biaisés et peuvent privilégier un partitionnement des observations en fonction uniquement de la variance et de la magnitude des variables. Ce phénomène est illustré sur la Figure 2.1. Les données présentées dans les sous-Figures A et B ont été générées de manière identique, mais seules les données représentées dans la sous-Figure B ont été standardisées avant l'application de l'algorithme des k-means. Il est notable que les deux partitions en deux clusters diffèrent considérablement. Dans la sous-Figure A, la partition semble principalement influencée par la seconde variable, bien que cette dernière ne contribue en rien à la séparation des clusters, étant simplement une variable gaussienne. En revanche, dans la sous-Figure B, la partition se fait selon la première variable, qui a été délibérément générée selon un mélange gaussien à deux composantes. Sans standardisation, les distances entre les observations sur la première variable peuvent sembler moindre par rapport à celles sur la seconde variable, car l'échelle de mesure sur cette dernière ainsi que sa variance sont nettement supérieures. Cela a pour effet de fausser la perception de la contribution de chaque variable à la structure des données. La standardisation corrige ce problème en accordant le même poids à toutes les variables lors du calcul des distances et du clustering.

Maintenant que nous avons illustré l'importance de la standardisation dans le cadre du clustering, comment s'opère-t-elle ? Soit $x_1, \ldots, x_n$, n observations d'une variable continue x. La standardisation la plus connue consiste à centrer et réduire les observations de x pour qu'elles aient une moyenne nulle et une variance unitaire. Cette standardisation, aussi appelée z-score, s'écrit alors :

$$z_i = \frac{x_i - \overline{x}}{\widehat{s}}, i = 1, \ldots, n$$

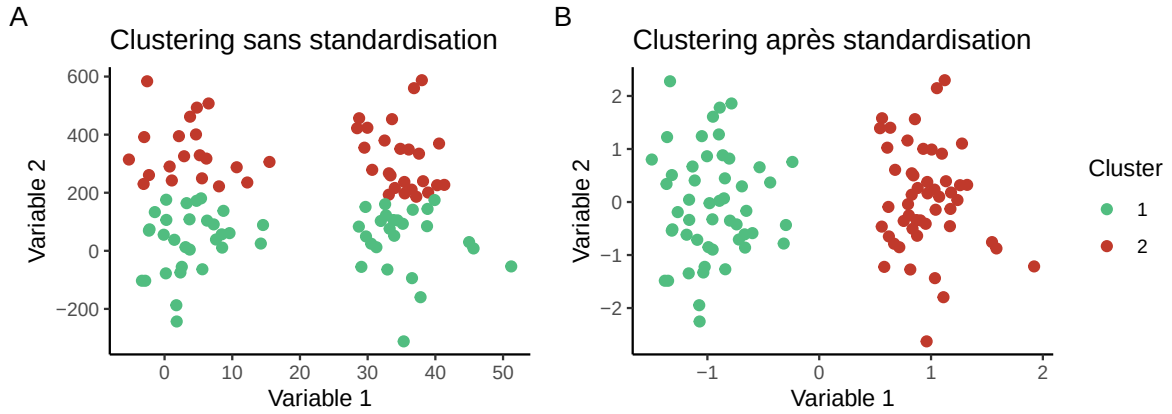


FIGURE 2.1 – **Impact de la standardisation sur les résultats des *k*-means.**

où $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ est la moyenne empirique des n observations et $\widehat{s} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2}$ est leur écart-type estimé. Ainsi, grâce au z-score, chaque observation est transformée en une mesure de déviation par rapport à la moyenne, exprimée en termes d'écart-type. Cette transformation rend les variables comparables entre elles, car elles sont toutes exprimées dans des unités d'écart-type, éliminant ainsi les différences d'échelle et de magnitude initiales. Risso & Pagnotta (2021) se sont inspirés de ces z-scores, traditionnellement utilisés dans le cadre de l'analyse de données supposées gaussiennes, pour proposer une méthode de standardisation pour les données RNA-seq qui respecte davantage leur nature. Ils suggèrent de remplacer les estimations traditionnelles de la moyenne et de l'écart-type en exploitant les propriétés de la distribution log-normale, observable dans la queue droite de la distribution de l'expression génique.

D'autres types de standardisation existent, se distinguant principalement par le diviseur utilisé pour ramener les variables à une échelle commune. En effet, la soustraction au numérateur des z-scores ne fait que translater les observations vers zéro, ce qui ne résout pas les problèmes de magnitude entre les variables. Une option pour le diviseur est la déviation absolue médiane (ou MAD, pour *Median Absolute Deviation* en anglais) (Everitt & Hothorn 2006), une mesure de dispersion robuste souvent utilisée en remplacement de l'écart-type. Cette standardisation est définie comme suit :

$$z_i^{\text{MAD}} = \frac{x_i - \text{median}(x)}{\text{median}\left(|x_i - \text{median}(x)|\right)}$$

Une autre forme de standardisation consiste à diviser chaque observation par la plage de valeurs prises par les observations de x (Milligan & Cooper 1988), ce qui conduit à une standardisation de la forme :

$$z_i^{\text{RANGE}} = \frac{x_i - \min(x)}{\max(x) - \min(x)}$$

Dans la littérature sur le clustering, deux études sont souvent citées concernant l'évaluation de l'impact de la standardisation sur les résultats du clustering. La première étude, menée par Milligan & Cooper (1988), a évalué les performances de sept méthodes de standardisation, incluant les z-scores et z^{RANGE}, ainsi que l'absence de standardisation dans le cadre de la classification ascendante hiérarchique (pour différentes mesures d'agrégation). Ils ont observé des performances complémentaires entre les z-scores et l'absence de standardisation. En effet, l'absence de standardisation semble efficace dans les cas où il n'y a pas de bruit dans les données, c'est-à-dire lorsque toutes les variables sont générées de sorte à séparer les clusters, tandis que les z-scores semblent plus robustes à l'ajout de variables bruitées. Néanmoins, toutes les méthodes de standardisation se sont révélées nettement moins performantes que z^{RANGE}. Ces résultats ont été confirmés par la seconde étude, celle de Steinley (2004*b*) avec l'algorithme des k-means. Selon eux, cette meilleure performance de z^{RANGE} par rapport aux z-scores classiques provient d'un effet trop prononcé de la réduction à la variance unitaire induite par les z-scores. La standardisation par la plage de valeurs prises par les observations a un effet moins prononcé sur les variances : elle permet de rendre les variables comparables entre elles en termes de magnitude et d'échelle tout en préservant les différences issues de la composition originale des données. En quelque sorte, z^{RANGE} représente un compromis entre l'absence de standardisation et les z-scores. Il est néanmoins important de noter que selon leur étude, le choix de la méthode de standardisation n'a jamais eu d'effet délétère sur le clustering. Une troisième étude, moins citée, menée par Schaffer & Green (1996), a également étudié l'impact de la standardisation, mais sur des données réelles, et a souligné de meilleures performances du clustering en l'absence de standardisation. Ces résultats contraires à ceux des deux études précédentes qui sont pourtant les études de références dans le cadre de la standardisation, peuvent s'expliquer par des difficultés à simuler des données réellement représentatives des données observées en pratique. Cependant, toutes les études sont unanimes sur un point : la standardisation par les z-scores n'est pas celle conduisant aux meilleures performances du clustering, que ce soit sur données simulées ou sur données réelles. Malgré ces résultats en défaveur des z-scores, ce sont pourtant eux qui se sont imposés comme la méthode de référence pour la standardisation des données en amont du clustering du fait de leur interprétabilité.

2.2.2 La standardisation dans le cadre des données RNA-seq

Les données RNA-seq, faisant généralement l'objet d'un clustering, ne font pas exception à la standardisation. Ainsi, elles sont couramment transformées en z-scores, conformément aux pratiques générales en analyse de données. Cette standardisation est alors automatiquement intégrée dans les pipelines de clustering les plus connues comme celle de `Seurat`. Outre le clustering, la standardisation dans le cadre de l'analyse de données

RNA-seq est nécessaire pour visualiser les données sous forme de *heatmaps*. En effet, cette représentation graphique n'est visuellement parlante que si les variances des comptes sont comparables, autrement ceux ayant les plus grandes variances dominent la représentation graphique, empêchant alors de repérer des variations moins prononcées de certains gènes. Aussi, la dimension des données RNA-seq est souvent réduite par une ACP. Là aussi la standardisation a un effet important puisqu'en utilisant les z-scores au lieu des données originales, l'ACP s'effectue sur la matrice de corrélations des gènes et non sur leur matrice de covariance.

La nécessité de standardiser les données dans le cadre de l'analyse de données RNA-seq est cependant discutable, comme souligné par Luecken & Theis (2019), qui mettent en évidence l'absence de consensus à ce sujet. Traditionnellement, la standardisation vise à rendre les variables comparables entre elles en termes de magnitude et d'échelle. Néanmoins, dans le cas des données RNA-seq, toutes les variables mesurent le même processus : l'expression génique, ce qui élimine le problème d'échelle. De plus, la magnitude de l'expression génique peut fournir des informations sur la biologie sous-jacente de l'échantillon. En particulier, la variance des gènes peut s'avérer informative comme en témoigne sa grande utilisation comme critère de sélection de caractéristiques. Or, l'impact trop prononcé des z-scores sur les variances a déjà été souligné par Steinley (2004*b*). Ainsi dans le cadre de l'analyse de données RNA-seq, la standardisation peut être justifiée uniquement pour des tâches telles que la visualisation sous forme de *heatmaps*, et s'explique pour le clustering par le simple fait que la pratique générale consiste à standardiser les données en amont d'un clustering. Mais, il n'y a ni preuves statistiques ni justifications biologiques solides pour expliquer son utilisation répandue.

Les conclusions des études menées par Milligan & Cooper (1988) et Steinley (2004*b*) sur la standardisation ne peuvent être directement extrapolées à l'analyse des données RNA-seq. Ces travaux ont été réalisées avant le changement de paradigme induit par l'avènement de la grande dimension. Elles se sont toutes deux déroulées avec un nombre limité de variables, ce qui implique une prédominance du signal et une faible présence de bruit. En revanche, la grande dimension se caractérise plutôt par un grand nombre de variables de bruit, surpassant même parfois le nombre de variables portant du signal. Jusqu'à présent, seuls Raymaekers & Zamar (2020) ont évalué l'impact de la standardisation sur le clustering en grande dimension. Leur étude compare leur méthode de standardisation, basée sur l'inertie intra-cluster, avec quatre autres méthodes, y compris l'absence de standardisation et les z-scores, à travers des simulations numériques contenant un grand nombre de variables bruitées et une application sur des données réelles. Leurs résultats soulignent les meilleures performances de leur approche en présence de bruit et des performances similaires entre les z-scores et l'absence de standardisation. Toutefois, tout comme les études antérieures, leur intérêt résidait davantage sur la comparaison des méthodes de standardisation entre elles. En particulier, Raymaekers & Zamar (2020) es-

timent que la standardisation est indispensable en amont du clustering pour les raisons de mise à l'échelle énoncées plus haut. Dans le contexte de l'analyse des données d'expression génique, la question concerne souvent la comparaison entre les z-scores, une pratique courante, et l'absence de standardisation, justifiée par la similarité naturelle des variables mesurant un même processus biologique.

2.3 Méthodes

Pour déterminer si la standardisation est réellement bénéfique pour le clustering dans le contexte de l'analyse des données RNA-seq, il est crucial d'étudier son impact dans des conditions où la dimensionnalité des données dépasse largement celle des études précédentes, menées par Milligan & Cooper (1988) et Steinley (2004*b*). Il est également nécessaire d'évaluer son impact sur des données réelles pour lesquelles une partition des observations est connue a priori, comme dans les études menées par Schaffer & Green (1996) et Raymaekers & Zamar (2020).

2.3.1 Étude de simulations

Pour évaluer l'impact de la standardisation sur les performances du clustering dans un contexte de grande dimension, nous nous sommes appuyés sur un exemple jouet basé sur un modèle de mélange gaussien à deux composantes : $\boldsymbol{X} \sim \pi_1 \mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1) + \pi_2 \mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$. Nous avons fixé les paramètres $\pi_1 = \pi_2 = 0.5$ ainsi que $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma} = \boldsymbol{I}_{p_1}$. Ce modèle représente une configuration de base en clustering, caractérisée par une matrice de variance-covariance partagée entre les composantes et une absence de corrélations intra-cluster. D'après les études précédentes sur l'impact de la standardisation, trois critères semblent particulièrement d'intérêt et d'autant plus dans le cadre de la grande dimension. Bien que très basique, notre exemple est suffisant pour les étudier tous les trois.

Le premier est la séparation entre les clusters : un fort degré de séparation entre les clusters conduit à de meilleures performances du clustering. Cette séparation entre les clusters est directement liée à la différence de moyenne $\delta = |[\boldsymbol{\mu}_2]_j - [\boldsymbol{\mu}_1]_j|$, $j = 1, \ldots, p_1$, des deux composantes du mélange, ainsi qu'à leur variance. Dans un contexte univarié, en considérant une variance unitaire (comme dans notre cas), une bonne séparation des composantes commence à être observable pour $\delta \geq 3$ (Siffer et al. 2018). Nous avons alors fixé $\boldsymbol{\mu}_1 = \mathbf{0}_{p_1}$ et $\boldsymbol{\mu}_2 = \begin{pmatrix} \delta \\ \vdots \\ \delta \end{pmatrix} \in \mathbb{R}^{p_1}$, c'est-à-dire une différence de moyenne constante de δ sur les p_1 variables du mélange. Ces p_1 variables issues du mélange sont des variables informatives, séparant effectivement les deux composantes. Ce nombre p_1 de variables portant de l'information est le second facteur d'intérêt de notre étude puisqu'il est directement lié au ratio signal versus bruit connu en grande dimension. En effet, si trop peu de variables sont informatives pour le clustering, même pour une grande valeur de δ, l'information se

dilue totalement dans le bruit, empêchant ainsi les algorithmes de clustering de correctement identifier la structure des données. Pour ajouter du bruit, nous avons ajouté aux p_1 variables informatives p_2 variables aléatoires gaussiennes de moyenne nulle et de variance σ_b^2, indépendantes entre elles, et indépendantes des p_1 variables informatives. Enfin, le dernier critère étudié correspond à la force du bruit par rapport à l'information, gérée au travers de la variance commune des p_2 variables de bruit. En particulier, si cette variance excède trop celles des variables informatives, alors les magnitudes des variables ne sont plus comparables, et la standardisation devient nécessaire comme illustré en Figure 2.1 et démontré dans l'étude de Steinley (2004*b*). Cependant, dans notre contexte applicatif, on ne s'attend pas à observer de telles différences de magnitudes entre les variables.

$n = 100$ observations et un total $p = p_1 + p_2$ de 1 000 variables ont été considérées. À p fixé, δ, la différence de moyenne entre les composantes sur les p_1 variables informatives variait entre 0.1, indiquant une très faible séparation entre les composantes, jusqu'à 10, témoignant d'une forte séparation des composantes. Le nombre p_1 de variables informatives, variait également, allant de 2 (très peu d'information) à 998 (très peu de bruit). À partir de la variance des p_1 variables informatives définie par $\mathbf{\Sigma}^* = \mathbf{\Sigma} + 0.25 \begin{pmatrix} -\delta \\ \vdots \\ -\delta \end{pmatrix} \times (-\delta \ \dots \ -\delta)$, la variance du bruit est définie comme $\sigma_b^2 = \rho {\sigma^*}^2$, où ${\sigma^*}^2$ est le terme commun de la diagonale de $\mathbf{\Sigma}^*$. Le paramètre ρ permet de contrôler le degré de comparabilité entre la variance du bruit et celle de l'information. 7 valeurs différentes de ce paramètre ont alors été examinées, allant de 0.01, où la variance du bruit est 100 fois inférieure à celle de l'information, à 3, où la variance du bruit est trois fois supérieure à celle de l'information.

2.3.2 Analyse de données réelles

Nous avons également évalué les performances du clustering en comparant les z-scores aux données originales à l'aide de quatre ensembles de données RNA-seq en masse. Parmi ces ensembles, trois ont été extraits de l'Atlas du génome du cancer (TCGA, pour *The Cancer Genome Atlas* (Weinstein et al. 2013)). Ce projet collaboratif, dirigé par l'Institut national du cancer et l'Institut national du génome humain aux États-Unis, a permis la caractérisation de l'expression génique dans environ 20 000 échantillons couvrant 33 types de cancer différents. Nous avons spécifiquement étudié trois types de cancer : le cancer du rein (TCGA-KIRP), le gliome (TCGA-LGG) et le cancer de l'estomac (TCGA-STAD). Les données d'expression génique correspondant à ces trois types de cancers ont été récupérées grâce au paquet `R TCGAbiolinks` (Mounir et al. 2019) disponible sur `Bioconductor`. Nous avons également ré-analysé les données RNA-seq de Lévy et al. (2021) dans le cadre de la COVID-19 présentées en Section 1.1.3. Ces quatre ensembles de données comportent des groupes de patients réels, connus en amont du clustering. Ils peuvent donc être utilisés comme référence pour évaluer l'effet de la standardisation. Une

description de chaque jeu de données se trouve en Table 2.1.

	TCGA-KIRP	**TCGA-LGG**	**TCGA-STAD**	**FrenchCOVID**
Nombre d'observations	135	534	450	54
Nombre de gènes	19947	19947	19947	30185
Nombre de sous-groupes	2	3	4	2
Sous-groupes	Type 1 Papillary RCC Type 2 Papillary RCC	IDHmut-codel IDHmut-non-codel IDHwt	CIN EBV GS MSI	Sain COVID19
Distribution des sous-groupes	75/60	169/250/94	138/25/54/60	10/44

TABLE 2.1 – **Caractéristiques des données RNA-seq en masse utilisées.**

Chaque jeu de données a été prétraité selon le même pipeline. Les observations pour lesquelles le sous-groupe d'origine était inconnu ont été exclues de l'analyse. Ensuite, parmi les observations restantes, les gènes présentant une expression constante ont été éliminés et les données ont été log2-normalisées par la méthode du compte par million (cpm). Nous avons effectué une analyse différentielle entre les groupes connus en utilisant la méthode `dearseq` (Gauthier et al. 2020), permettant ainsi d'estimer le nombre de variables informatives p_1 en considérant comme informatifs les gènes ayant une p-valeur ajustée significative au seuil de 5%.

2.3.3 Critère d'évaluation

Nous avons examiné les performances du clustering pour identifier la structure réelle des données et ainsi comparer l'utilisation des données originales (X) par rapport aux z-scores (Z). Pour ce faire, nous avons utilisé l'Indice de Rand Ajusté (ARI) (Hubert & Arabie 1985) qui quantifie la similitude entre deux partitions en tenant compte du fait que certaines correspondances peuvent être simplement dues au hasard (Steinley 2004*a*). L'ARI prend des valeurs comprises entre -1 et 1. Une valeur proche de 1 indique une forte concordance, tandis qu'une valeur nulle suggère une correspondance due au hasard, et une valeur négative indique une concordance pire que le hasard. Si la vraie partition des observations est donnée par $C^* = \{C_1^*, \ldots, C_K^*\}$ et celle résultant du clustering est donnée par $\widehat{C} = \{\widehat{C}_1, \ldots, \widehat{C}_L\}$ (où K peut être différent de L), alors on peut définir la table de contingence entre les deux partitions comme en Table 2.2.

	$\widehat{C}_1$	$\widehat{C}_2$	$\cdots$	$\widehat{C}_L$	**Total**
C_1^*	n_{11}	n_{12}	$\cdots$	n_{1L}	a_1
C_2^*	n_{21}	n_{22}	$\cdots$	n_{2L}	a_2
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
C_K^*	n_{K1}	n_{K2}	$\cdots$	n_{KL}	a_K
Total	b_1	b_2	$\cdots$	b_L	n

TABLE 2.2 – **Table de contingence entre la vraie partition C^* et la partition donnée par le clustering $\widehat{C}$.**

À partir de cette table, l'ARI se définit alors par :

$$\text{ARI} = \frac{\sum_{ij}\binom{n_{ij}}{2} - [\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}]/\binom{n}{2}}{\frac{1}{2}[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}] - [\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}]/\binom{n}{2}} \tag{2.1}$$

où n_{ij}, a_i, b_j, $i = 1, \ldots, K, j = 1, \ldots, L$ et n sont définis dans la table de contingence en Table 2.2.

Pour les deux analyses, l'étude de simulations et les analyses de données réelles, l'ARI a été calculé entre la partition connue des données et les résultats de deux méthodes de clustering, à savoir les k-means et la classification ascendante hiérarchique, lorsque soit les données originales X, soit les z-scores Z, ont été utilisés pour le clustering. Pour l'étude de simulations, les données ont été simulées 1 000 fois pour chaque valeur de δ, p_1 et ρ. Pour les applications, deux niveaux de bruits définis par p_1 (absence de bruit) et $p_1 + p_2$ (présence de bruit) ont été comparés.

2.4 Résultats

2.4.1 Étude de simulations

Les résultats de l'étude de simulations sont présentés dans la Figure 2.2. Dans l'ensemble, les k-means semblent plus performants que la classification ascendante hiérarchique de Ward, quel que soit le nombre de variables informatives p_1, la force de la séparation δ qu'elles portent et le ratio des variances ρ. Globalement, le clustering basé sur X donne de meilleurs résultats par rapport à celui sur Z lorsque peu de variables informatives sont présentes dans les données. Comme prévu, à mesure que le nombre de variables informatives p_1 augmente, les performances du clustering s'améliorent, atteignant même un ARI de 1 lorsque toutes les variables sont informatives. Ainsi, en l'absence de bruit dans les données, les deux méthodes de standardisation sont comparables, comme présenté dans le panneau A.

L'influence du ratio des variances $\rho = \sigma_b^2/\sigma^{*^2}$ est illustrée dans le panneau B. Seules les performances du clustering sur les données originales X sont impactées par ce facteur. En effet, la standardisation vise principalement à rendre les variables comparables entre elles, maintenant ainsi ce ratio constant à 1, quelle que soit la variance initiale des variables. Ainsi, si la variance des p_2 variables de bruit est faible par rapport à celle des p_1 variables informatives ($\rho < 1$), alors les variances des variables informatives deviennent, tout comme leur différence de moyenne δ, des signaux facilitant la découverte de la véritable partition des observations, rendant préférable le clustering sur X. Dans ce cas, la standardisation des données conduit à une augmentation artificielle du poids des variances des p_2 variables de bruit, ce qui nuit aux performances du clustering. Si $\rho = 1$, alors le clustering sur X et sur Z conduisent aux mêmes résultats puisque les variables sont déjà à la même échelle ;

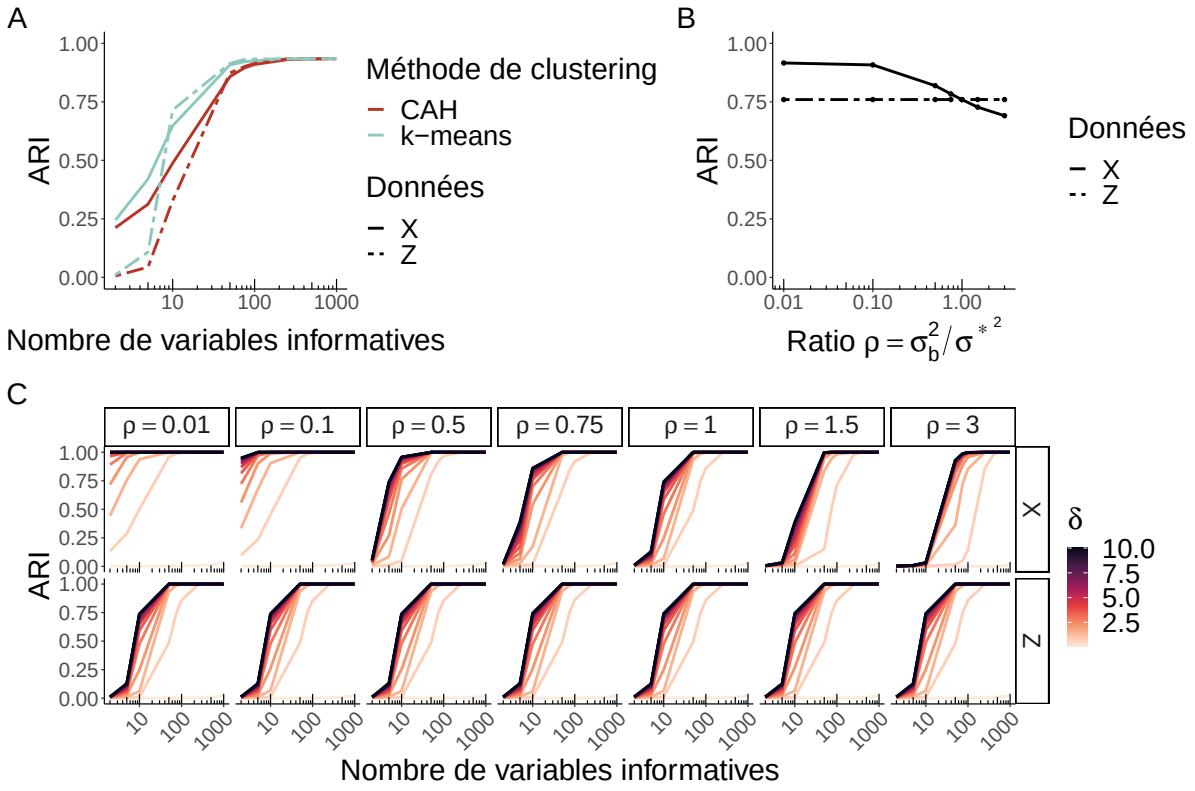


FIGURE 2.2 – **Comparaison des performances du clustering entre *z*-scores ($\boldsymbol{Z}$) et données originales ($\boldsymbol{X}$) en terme d'ARI calculé à partir de 1 000 simulations des données**. Panneau A : Évolution de l'ARI en fonction du nombre de variables informatives p_1 et de la méthode de clustering pour X et Z. Panneau B : Évolution de l'ARI en fonction du ratio $\rho = \sigma_b^2/\sigma^{*^2}$ pour X et Z. Panneau C : Évolution de l'ARI en fonction du nombre de variables informatives p_1, de la force de la séparation qu'elles portent δ et du ratio ρ pour X et Z.

la standardisation n'a alors aucun effet. Enfin, si $\rho > 1$, c'est la variance des p_2 variables de bruit qui est utilisée par les algorithmes de clustering pour construire les clusters. Dans ce cas, l'information est diluée dans le bruit et la standardisation devient nécessaire pour garantir de bonnes performances du clustering. Ces résultats sont conformes à ceux attendus et retrouvés dans la littérature (Steinley 2004*b*).

Le panneau C fournit des résultats plus détaillés sur l'impact conjoint du ratio des variances ρ et de la différence de moyenne δ portée par les p_1 variables informatives. Ces deux paramètres sont liés, car plus δ augmente, plus la variance des variables informatives augmente également, comme le montre le calcul de la variance de ces p_1 variables $\boldsymbol{\Sigma}^*$ en Section 2.3.1. De manière générale, plus δ augmente, meilleures sont les performances du clustering, indépendamment du choix de la méthode de standardisation. Seules les faibles différences de moyennes δ sont donc intéressantes : si trop peu de variables sont informatives, alors seul le clustering sur X parvient à identifier la partition des données. Dans ces cas, bien que peu présentes, les p_1 variables informatives et la faible séparation

qu'elles portent restent facilement identifiables en raison de leur variance plus élevée que celle du bruit.

Cette étude de simulations révèle que les performances du clustering sur les z-scores sont principalement impactées par le nombre de variables informatives et la force de la séparation qu'elles portent. Cependant, rendre toutes les variables comparables peut s'avérer être une perte d'information. Outre leur différence de moyenne, les variances des variables informatives peuvent également être source de signal et aider les algorithmes de clustering à retrouver la partition des observations. Il devient ainsi plus facile d'identifier de faibles signaux peu visibles en considérant les variables originales. Cependant, cela n'est vrai que dans les cas où la variance des variables informatives est supérieure à celle des variables de bruit, rendant les performances du clustering sur X dépendantes de ce ratio, en plus des deux autres paramètres δ et p_1. En pratique, standardiser ou non les données revient donc à faire une forte hypothèse : celle que les variables ont des échelles de mesures comparables et que leur variance est informative sur le clustering.

2.4.2 Analyse de données réelles

Les résultats de l'analyse des données réelles sont présentés dans la Figure 2.3. Le panneau A donne les résultats obtenus pour chaque jeu de données en termes d'ARI. La variabilité des résultats entre les jeux de données peut s'expliquer par le nombre de sous-groupes connus qui diffère allant de 2 pour TCGA-KRIP et FrenchCOVID à 4 pour TCGA-STAD, avec une dégradation des performances médianes au fur et à mesure que ce nombre augmente. Elle peut également être causée par la complexité de définir une partition a priori des données. En effet, l'évaluation des performances du clustering sur données réelles peut être sujette à des biais, particulièrement dans le cadre des données RNA-seq en raison de leur forte hétérogénité, car la partition de références choisie peut ne pas être la plus représentative de la structure réelle des données. Cependant, une tendance générale se dégage. Dans l'ensemble, sur les quatre jeux de données examinés, l'absence de standardisation au pire entraîne des performances similaires, et au mieux des performances supérieures aux z-scores (panneau A). Seul le clustering sur TCGA-LGG semble préférable sur les données standardisées, avec un ARI médian de 0.25 pour X et 0.43 pour Z.

En ce qui concerne l'algorithme de clustering, la classification ascendante hiérarchique de Ward semble être plus sensible à la standardisation des données que les k-means, avec des performances plus variables mais toujours en faveur de l'utilisation de X (panneau B). L'impact le plus visible sur les performances du clustering semble être la présence de bruit dans les données (panneau C). Les z-scores sont plus efficaces lorsque les données sont exemptes de bruit. En revanche, le clustering sur les données brutes semble moins affecté par le bruit, ce qui se traduit par des performances médianes relativement similaires à

celles obtenues avec uniquement les variables informatives.

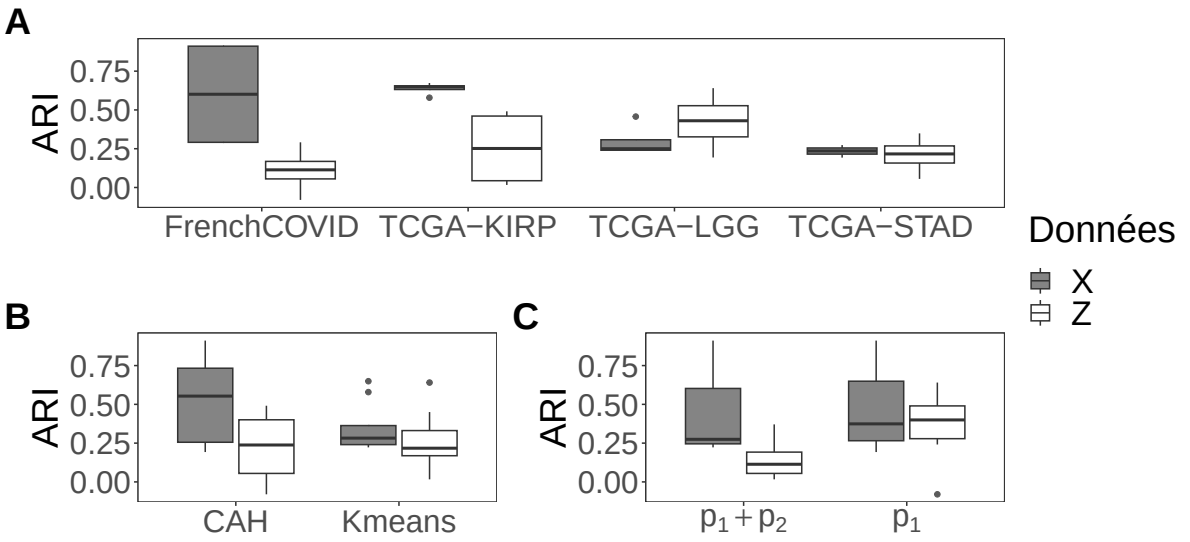


FIGURE 2.3 – **Impact de la standardisation par les z-scores ($\boldsymbol{Z}$) ou non ($\boldsymbol{X}$) sur données réelles**. Panneau A : Performances globales du clustering sur chaque jeu de données en fonction de la standardisation ou non des données. Panneau B : Performances des méthodes de clustering en fonction de la standardisation ou non des données. Panneau C : Impact de la présence de bruit sur les performances du clustering en fonction de la standardisation ou non des données.

En considérant, comme pour l'étude de simulation, le ratio ρ de la variance des p_2 gènes non différentiellement exprimés (de bruit) sur celle des p_1 gènes différentiellement exprimés (informatifs), on constate un ratio médian $\rho = 0.88$. Ainsi, les gènes non différentiellement exprimés ont tendance à avoir une variance inférieure à celle des gènes différentiellement exprimés. Cette variance devient alors une source d'information lorsque les données X sont utilisées pour le clustering, expliquant ses meilleures performances en présence de bruit. Ne considérer que les gènes informatifs n'améliore pas les performances du clustering sur X, car, dans tous les cas, en raison de leurs variances plus élevées, les gènes informatifs sont facilement identifiables. Seul le clustering sur Z semble gagner en performance lorsque uniquement les p_1 gènes informatifs sont utilisés pour le clustering. Dans le cas où il y a du bruit, les z-scores, en attribuant le même poids à tous les gènes, augmentent artificiellement celui des gènes non différentiellement exprimés, supprimant ainsi une source d'information importante et rendant le clustering plus difficile.

2.5 Discussion

Les données RNA-seq nécessitent un prétraitement important en amont du clustering, et chaque étape peut avoir un impact conséquent sur ses résultats. Il est donc crucial

de définir un pipeline optimal pour garantir les meilleures performances du clustering et réduire au maximum les incertitudes, surtout si les clusters doivent être réutilisés pour une analyse différentielle. Bien que de nombreuses recherches aient été menées sur l'impact des différentes étapes de prétraitement sur le clustering, la standardisation des données dans le contexte des données RNA-seq reste un sujet peu exploré mais néanmoins important. Cette étude a donc cherché à examiner en détail pourquoi la standardisation des données est sujette à débat et à comprendre son impact sur le clustering pour finalement déterminer sa réelle nécessité.

La standardisation des données dans le cadre du clustering a été largement étudiée avant l'avènement des données RNA-seq et de la grande dimension. Bien qu'essentielle pour rendre les variables comparables entre elles et éviter que le clustering ne soit influencé uniquement par les variables présentant la plus grande variance, nos résultats dans le cadre des données RNA-seq contredisent les recommandations habituelles, allant plutôt en faveur de l'absence de standardisation, comme préconisé par Luecken & Theis (2019). En effet, la nature de ces données suggère que les gènes ont une variance comparable entre eux, étant donné qu'ils mesurent un même processus : l'expression génique. Il est même probable que la variance des gènes différentiellement exprimés entre deux clusters soit supérieure à celle des gènes non différentiellement exprimés, simplement parce que les clusters sont séparés sur ces gènes, contrairement aux gènes non différentiellement exprimés.

Nos simulations numériques montrent que, au-delà de la force de la séparation, la variance peut également apporter de l'information pour le clustering lorsque celle des variables informatives est supérieure à celle du bruit. Cette hypothèse est cohérente avec nos observations sur des données réelles, qui révèlent une tendance à de meilleures performances du clustering sur les données originales par rapport aux z-scores, la méthode de standardisation la plus couramment utilisée. Nous avons d'ailleurs constaté que la variance des gènes informatifs était effectivement supérieure à celle des gènes de bruit, ce qui confirme son utilité pour la détection des clusters. En utilisant les z-scores, cette information portée par la variance est diluée, la rendant comparable à celle du bruit, ce qui conduit à des performances de clustering moins bonnes. Il s'agit là d'un résultat déjà étayé par Steinley (2004*b*). Cependant, nous avons pu mettre en avant que cette détérioration des performances a particulièrement lieu lorsque peu de variables sont informatives et que la force de la séparation est relativement faible.

Notre étude présente plusieurs limites. Elle repose sur un schéma de simulation simplifié, ne prenant notamment pas en compte les corrélations entre les variables informatives au sein d'un cluster, ni entre les variables de bruit entre elles. En se concentrant uniquement sur l'impact de la variance, nous négligeons alors d'autres sources potentielles d'information ou de bruit pour le clustering telles que les corrélations entre les variables. Cette simplification a l'avantage de rendre compréhensible et évidente l'effet de la standar-

disation sur chaque facteur étudié. Il convient de noter que des méthodes de simulations de données plus réalistes existent (Milligan 1985, Qiu & Joe 2006) et ont déjà été largement utilisées dans la littérature notamment dans le cadre d'études sur la standardisation (Milligan & Cooper 1988, Balakrishnan et al. 1994, Waller et al. 1998, Carmone Jr et al. 1999, Brusco & Cradit 2001). Bien que ces méthodes soient compréhensibles en petite dimension, leur comportement en grande dimension est moins clair, rendant la différenciation et l'interprétation de l'impact des différents facteurs entre eux plus complexes, comme le souligne notre expérience. Enfin, en ce qui concerne notre analyse des données réelles, elle est limitée par le fait qu'elle inclut seulement quatre jeux de données RNA-seq en masse, suivant ainsi le schéma de l'étude menée par Vidman et al. (2019). Une extension de cette étude à des données RNA-seq en cellule unique serait également nécessaire pour affiner les conclusions et élargir la portée de nos résultats.

Bien qu'en faveur de l'utilisation des données non-standardisées, il est difficile de conclure sur l'impact de la standardisation. Ce choix reste dépendant de l'analyse et du type d'hypothèses que l'on considère a priori. En effet, ne pas standardiser les données implique une hypothèse sous-jacente très forte : la variance des gènes est bel et bien comparable et informative sur la présence de cluster. Nous recommandons alors de comparer systématiquement les résultats du clustering sur des données non standardisées et standardisées afin d'évaluer en pratique l'impact de la standardisation lors des analyses. Mais dans ce cas, une question demeure : quelle méthode de standardisation utiliser ? Les z-scores sont-ils réellement la méthode de standardisation à privilégier ? Raymaekers & Zamar (2020) donnent des premiers éléments de réponses à ces questions en introduisant une méthode de standardisation inspirée des z-scores qui semble performante dans le cadre des données RNA-seq. Enfin, bien que cette étude aide à comprendre comment obtenir un bon clustering des observations, il n'en reste pas moins qu'en pratique la vraie partition des données est inconnue et qu'il reste donc toujours de fortes incertitudes. Conjointement à la double utilisation des données, ces recommandations ne sont pas suffisantes pour répondre aux enjeux de l'inférence post-clustering.

Chapitre 3

Etat de l'art des méthodes d'inférence post-clustering

Contenu

3.1 Introduction 59
3.2 Méthodes basées sur un conditionnement 60
3.2.1 Lemme polyédrique 60
3.2.2 Méthodes d'inférence post-clustering 61
3.3 Méthodes basées sur une décomposition explicite de l'information 65
3.3.1 Découpage de l'échantillon 65
3.3.2 Découpage d'une observation 66

3.1 Introduction

L'inférence post-clustering, fréquemment rencontrée lors de l'analyse des données RNA-seq et détaillée en Section 1.2.3, est un défi central pour garantir la fiabilité des découvertes dans ce domaine (Lähnemann et al. 2020). Elle s'inscrit dans le contexte plus large des problèmes d'inférence sélective, dont le cadre théorique a été défini par Fithian et al. (2014). L'inférence sélective englobe toutes les formes d'analyses où la sélection du modèle M et/ou de l'hypothèse de test $\mathcal{H}_0$ est déterminée par les données elles-mêmes. Elle offre alors une réponse à la double utilisation des données servant à poser les questions et à y répondre simultanément. En effet, dans ce contexte, les méthodes d'inférence ont tendance à être trop optimistes, rejetant trop facilement $\mathcal{H}_0$ et conduisant ainsi à une inflation de l'erreur de type I, comme montré en Figure 1.11 dans le cadre de l'inférence post-clustering. Pour corriger ce problème et ainsi s'assurer de la fiabilité des découvertes, c'est-à-dire empêcher des conclusions qui ne soient que le simple résultat de la sélection du modèle, Fithian et al. (2014) proposent alors de contrôler l'erreur de type I sélective au niveau α :

$$\mathbb{P}_{\mathcal{H}_0}\left(\text{rejeter } \mathcal{H}_0 \mid (M, \mathcal{H}_0) \text{ sélectionnés}\right) \leq \alpha \tag{3.1}$$

Fithian et al. (2014) proposent deux stratégies pour garantir un contrôle efficace de l'erreur de type I sélective définie en (3.1). La première stratégie consiste à définir un

test d'hypothèse qui contrôle directement l'erreur de type I sélective au niveau α. Cela se fait en conditionnant sur l'événement de sélection dans la définition de la p-valeur du test. La seconde stratégie, souvent utilisée en apprentissage automatique pour contrer les problèmes similaires de sur-apprentissage, consiste à construire le modèle et les hypothèses sur un jeu de données, puis à les tester sur un autre jeu de données, indépendant du premier. Dans ce cas, étant donné que l'événement de sélection est indépendant de celui de test, l'erreur de type I sélective en (3.1) devient simplement $\mathbb{P}_{\mathcal{H}_0}$ (rejeter $\mathcal{H}_0$), correspondant en fait à l'erreur de type I traditionnellement contrôlée par n'importe quel test d'hypothèse. La première stratégie, la stratégie conditionnelle, peut être vue comme une généralisation de la seconde stratégie, basée sur un partage des données, comme l'ont souligné Goeman & Solari (2023). En effet, ces deux stratégies ont pour objectif commun de décomposer l'information en deux parties, que ce soit par un partage explicite ou par le conditionnement, afin de ne pas réutiliser l'information permettant de construire le modèle et les hypothèses lors du test d'hypothèse. Ce Chapitre vise donc à présenter les méthodes d'inférence sélective existantes, qu'elles recourent à la première ou à la seconde stratégie, afin de répondre aux défis posés par l'inférence post-clustering.

3.2 Méthodes basées sur un conditionnement

3.2.1 Lemme polyédrique

Les méthodes d'inférence sélective basées sur un conditionnement sur l'évènement de sélection reposent sur un résultat clé démontré par Lee et al. (2016) : le lemme polyédrique. Ce lemme permet de définir une statistique de test qui, conditionnellement à l'évènement de sélection, a une distribution uniforme sous $\mathcal{H}_0$ et peut donc servir pour calculer une p-valeur (Tibshirani et al. 2016). Formellement, soit $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ avec $\boldsymbol{\mu} \in \mathbb{R}^n$ et $\boldsymbol{\Sigma} \in \mathbb{R}^{n \times n}$. Considérons un évènement de sélection du modèle et des hypothèses dépendant des données $\boldsymbol{X}$ et s'écrivant sous la forme d'un polyèdre : $\{\boldsymbol{AX} < \boldsymbol{b}\}$ avec $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ et $\boldsymbol{b} \in \mathbb{R}^m$. Supposons de plus que l'hypothèse nulle $\mathcal{H}_0$ du test considéré puisse s'écrire sous la forme $\mathcal{H}_0 : \boldsymbol{\eta}^t \boldsymbol{\mu} = 0$, pour un certain vecteur de contraste $\boldsymbol{\eta} \in \mathbb{R}^n$. Alors :

$$F^{[V^-, V^+]}_{\boldsymbol{\eta}^t \boldsymbol{\mu}, \boldsymbol{\eta}^t \boldsymbol{\Sigma} \boldsymbol{\mu}} \left(\boldsymbol{\eta}^t \boldsymbol{X} \right) \mid \{\boldsymbol{AX} < \boldsymbol{b}\} \overset{\mathcal{H}_0}{\sim} \mathcal{U}[0, 1] \tag{3.2}$$

où $F^{[a,b]}_{\mu, \sigma^2}$ correspond à la fonction de répartition d'une distribution gaussienne $\mathcal{N}(\mu, \sigma^2)$ tronquée à l'intervalle $[a, b]$ et où V^- et V^+ ont des expressions analytiques détaillées dans Lee et al. (2016) et Tibshirani et al. (2016). En particulier, il découle de l'équation (3.2)

et des propriétés de la distribution uniforme que :

$$\mathbb{P}_{\mathcal{H}_0}\left(\underbrace{F^{[V^-,V^+]}_{\boldsymbol{\eta}^t\boldsymbol{\mu},\boldsymbol{\eta}^t\boldsymbol{\Sigma}\boldsymbol{\mu}}\left(\boldsymbol{\eta}^t\boldsymbol{X}\right) \leq \alpha}_{\text{rejeter } \mathcal{H}_0} \,|\, \underbrace{\{\boldsymbol{A}\boldsymbol{X} < \boldsymbol{b}\}}_{(M,\mathcal{H}_0) \text{ sélectionnés}}\right) = \alpha, \quad \forall \alpha \in [0,1]\,.$$

Cette équation garantit donc le contrôle de l'erreur de type I sélective définie en (3.1).

Ce lemme est au cœur de nombreuses méthodes d'inférence sélective basées sur un conditionnement, notamment dans le contexte de la régression. En effet, la régression est sujette aux problèmes d'inférence sélective lorsqu'elle implique une sélection des variables s'effectuant sur la base des données. Si l'on teste ensuite l'association entre une variable sélectionnée et la variable réponse sur les mêmes données, il est probable d'obtenir une association significative simplement en raison de la sélection de la variable. Mais, il a été montré que les méthodes de sélection de variables les plus connues dans le contexte de la régression, comme l'approche séquentielle ou le LASSO, conduisent à des évènements de sélection qui peuvent s'écrire sous la forme d'un polyèdre (ou d'une union de polyèdres) (Lee et al. 2016, Tibshirani et al. 2016, Liu et al. 2018) d'où la large utilisation de ce lemme.

Cependant, d'autres domaines que la régression s'intéressent au contrôle de l'erreur de type I sélective. Parmi eux, l'inférence post-clustering, qui nous concerne ici, ainsi que la détection de points de changement (Hyun et al. 2021, Jewell et al. 2022) ou les arbres CART (Neufeld et al. 2022). Dans ces contextes, l'événement de sélection est parfois difficile à exprimer sous forme d'un polyèdre. Dans certains cas, il est possible de conditionner sur davantage que l'événement de sélection lui-même pour se ramener à un polyèdre, mais cela peut entraîner une perte d'information et donc de puissance, comme l'ont démontré Fithian et al. (2014). Ainsi, Jewell et al. (2022) ont été les premiers à s'affranchir du lemme polyédrique pour développer un test d'inférence sélective dans le cadre de la détection de points de changement, une approche adoptée ensuite par Gao et al. (2024) pour l'inférence post-clustering.

3.2.2 Méthodes d'inférence post-clustering

Bien que la littérature sur l'inférence sélective soit de plus en plus étoffée, celle portant spécifiquement sur l'inférence post-clustering demeure relativement limitée. Notre objectif ici est donc de présenter les méthodes existantes qui reposent sur un conditionnement explicite sur la sélection du modèle dans la définition de la p-valeur.

3.2.2.1 `TN-test` (Zhang et al. 2019)

Zhang et al. (2019) ont été les premiers à proposer un test d'analyse différentielle post-clustering : le test Normal-Tronquée (*Truncated-Normal test* ou TN-test). Soient

$\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$, n réalisations *i.i.d* d'une variable aléatoire $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ avec $\boldsymbol{\mu} \in \mathbb{R}^p$ et $\boldsymbol{\Sigma} \in \mathbb{R}^{p \times p}$. Leur approche s'appuie sur une divisision des n observations en deux parties, $\boldsymbol{x}_1^{(1)}, \ldots, \boldsymbol{x}_{n_1}^{(1)}$ et $\boldsymbol{x}_1^{(2)}, \ldots, \boldsymbol{x}_{n_2}^{(2)}$, avec $n = n_1 + n_2$. La première partie est utilisée pour le clustering. Ensuite, la différence de moyenne de chaque gène (c'est-à-dire une différence marginale sur chacune des p dimensions) entre les clusters construits sur les $\boldsymbol{x}_1^{(1)}, \ldots, \boldsymbol{x}_{n_1}^{(1)}$ est testée à l'aide de la seconde partie des observations : $\boldsymbol{x}_1^{(2)}, \ldots, \boldsymbol{x}_{n_2}^{(2)}$. Mais pour ce faire, il est nécessaire de labelliser les observations $\boldsymbol{x}_1^{(2)}, \ldots, \boldsymbol{x}_{n_2}^{(2)}$ en fonction des clusters construits sur $\boldsymbol{x}_1^{(1)}, \ldots, \boldsymbol{x}_{n_1}^{(1)}$. Zhang et al. (2019) font l'hypothèse que les clusters sont linéairement séparables, c'est-à-dire qu'il existe un hyperplan pouvant les séparer. Les observations $\boldsymbol{x}_1^{(2)}, \ldots, \boldsymbol{x}_{n_2}^{(2)}$ sont donc labellisées en fonction de leur position par rapport à cet hyperplan, typiquement à droite ou à gauche de l'hyperplan dans le cas d'un problème à $p = 2$ dimensions. Cet hyperplan est estimé à partir des résultats du clustering en entraînant un classifieur de type *Support Vector Machine* (SVM) pour apprendre les clusters construits sur la première partie des données et en utilisant ces résultats pour prédire les labels des observations de la seconde partie des données.

Pour se ramener à un conditionnement plus simple que celui induit par le clustering, Zhang et al. (2019) conditionnent leur test sur la position des observations par rapport à l'hyperplan ayant servi à les labelliser. Il découle immédiatement de ce conditionnement que, sachant l'hyperplan qui sépare deux clusters, les observations au sein d'un cluster suivent une distribution gaussienne tronquée (d'où le nom du test). La troncature de la distribution est donnée par le fait pour une observation d'être d'un côté ou de l'autre de l'hyperplan. Cela permet de directement caractériser la distribution des observations conditionnellement à l'évènement de sélection. Il est alors possible d'intégrer de manière explicite l'information issue du processus de clustering, donnée par cette distribution tronquée, dans le calcul de la statistique de test et d'ainsi contrôler l'erreur de type I sélective.

3.2.2.2 `clusterpval` (Gao et al. 2024)

Gao et al. (2024) ont également introduit un test d'inférence sélective post-clustering. Contrairement à l'approche de Zhang et al. (2019), leur méthode conditionne directement sur l'événement de clustering lui-même, plutôt que sur la position par rapport à l'hyperplan estimé par un classifieur. Cependant, leur test aborde une question différente : sachant qu'ils ont été estimés, est-ce que deux clusters sont effectivement différents l'un de l'autre. L'hypothèse nulle dans ce cas porte alors sur l'égalité des barycentres entre les deux clusters, et non pas sur une égalité de moyenne au niveau des variables (c'est-à-dire sur une composante marginale des barycentres) comme on le souhaiterait dans le cadre de l'analyse différentielle. Il s'agit donc d'une méthode permettant de valider les résultats du clustering, mais qui ne permet pas d'inférer quelles variables sont responsables de la

séparation des deux clusters.

Nous allons détailler un peu plus en profondeur cette méthode car elle est à l'origine d'une des méthodes proposées au Chapitre 4. Formellement, soit un ensemble de données $\boldsymbol{X}$ distribuées selon une loi matrice normale $\mathcal{MN}_{n\times p}(\boldsymbol{\mu}, \boldsymbol{I_n}, \sigma^2\boldsymbol{I_p})$, où $\boldsymbol{\mu} \in \mathbb{R}^{n\times p}$ est une matrice inconnue de moyennes et σ^2 est connu. Sur $\boldsymbol{X}$, une méthode de clustering $c(\ )$ est appliquée, conduisant à une partition en K clusters de $\boldsymbol{X}$, donnée par $c(\boldsymbol{X}) = \left\{\widehat{C}_1, \ldots, \widehat{C}_K\right\}$. Gao et al. (2024) s'intéressent alors à la différence entre les barycentres de deux clusters $\widehat{C}_k$ et $\widehat{C}_l$. L'hypothèse nulle de leur test est donc :

$$\mathcal{H}_0 : \boldsymbol{\mu}^t\boldsymbol{\eta} = \boldsymbol{0_p} \tag{3.3}$$

où $\boldsymbol{\eta}$ est un vecteur de contraste défini par : $\eta_i = \frac{\mathbb{1}_{i\in\widehat{C}_k}}{|\widehat{C}_k|} - \frac{\mathbb{1}_{i\in\widehat{C}_l}}{|\widehat{C}_l|} \forall i = 1, \ldots, n$. La p-valeur du test d'inférence sélective est définie comme :

$$p \equiv \mathbb{P}_{H_0}\left(\|\boldsymbol{X}^t\boldsymbol{\eta}\|_2 \geq \|\boldsymbol{x}^t\boldsymbol{\eta}\|_2 \,|\, \hat{C}_k, \hat{C}_l \in c(\boldsymbol{X})\right) \tag{3.4}$$

avec $\boldsymbol{x}$ une réalisation de $\boldsymbol{X}$. Cette p-valeur inclue donc un conditionnement explicite sur le fait que les deux clusters $\widehat{C}_k$ et $\widehat{C}_l$ ont été estimés à partir des données, garantissant un contrôle de l'erreur de type I sélective. Toutefois, pour des raisons computationnelles, Gao et al. (2024) conditionnent sur un peu plus que ce simple évènement, comme ce qui a été fait par Jewell et al. (2022) pour s'affranchir de l'hypothèse polyédrique. Cela ne compromet pas le contrôle de l'erreur de type I sélective, mais peut cependant entraîner une perte de puissance statistique. En conditionnant sur plus dans la p-valeur en (3.4), il devient possible d'écrire une nouvelle p-valeur :

$$p = \mathbb{P}_{H_0}\left(\phi \geq \|\boldsymbol{x}^t\boldsymbol{\eta}\|_2 \,|\, \phi \in S\right) \tag{3.5}$$

où $\phi \sim (\sigma\|\boldsymbol{\eta}\|_2).\chi_p$ et $S = \left\{\phi : \widehat{C}_k, \widehat{C}_l \in c\left(\boldsymbol{x}'(\phi)\right)\right\}$ est l'ensemble des perturbations des données observées $\boldsymbol{x}$ par ϕ qui conduisent à la ré-estimation de $\widehat{C}_k$ et $\widehat{C}_l$ lorsque $c(\ )$ y est appliquée.

La partie la plus compliquée du calcul de cette p-valeur réside donc dans la description analytique de cet ensemble S des perturbations ϕ des données observées conduisant à la ré-estimation de $\widehat{C}_k$ et $\widehat{C}_l$. Bien sûr, elle dépend de la méthode de clustering $c(\ )$ considérée et à ce jour, une description explicite de cet ensemble n'est disponible que pour deux méthodes de clustering : la classification ascendante hiérarchique sur les distances euclidiennes au carré et certaines mesures d'agrégations (dont la mesure de Ward) (Gao et al. 2024) ainsi que les k-means (Chen & Witten 2023). Pour toutes les autres méthodes de clustering, la p-valeur en (3.5) reste tout de même calculable grâce à une approche de Monte-Carlo, en échantillonnant N perturbations $\phi^1, \ldots, \phi^N \sim (\sigma\|\boldsymbol{\eta}\|_2).\chi_p$ et en ne conservant que celles préservant $\widehat{C}_k$ et $\widehat{C}_l$. Cette approche et son interprétation seront détaillées dans le

Chapitre 4.

3.2.2.3 `poclin` (Bachoc et al. 2023)

Bachoc et al. (2023) ont également proposé un test d'inférence post-clustering en s'appuyant sur les travaux effectués dans le cadre de l'inférence sélective pour le LASSO de Lee et al. (2016). Cette connexion entre les deux problématiques est rendue possible grâce aux fortes similarités entre une méthode $c(\ )$ de clustering particulière : le clustering convexe, et le LASSO. Leur test d'inférence post-clustering s'inscrit directement dans le cadre de l'analyse différentielle puisqu'il permet de tester une différence de moyenne entre deux clusters $\widehat{C}_k$ et $\widehat{C}_l$ (obtenus grâce au clustering convexe) à l'échelle de la variable, et non plus à l'échelle du barycentre comme le test de Gao et al. (2024). Bachoc et al. (2023) se placent dans un premier temps dans le cadre univarié et s'intéressent à tester une hypothèse nulle similaire à celle en (3.3), c'est-à-dire s'écrivant sous la forme :

$$\mathcal{H}_0 : \boldsymbol{\mu}^t \boldsymbol{\eta} = 0$$

Cependant, de part la nature univariée des données ($p = 1$), il s'agit bien ici d'un test de différence de moyenne à l'échelle de la variable.

Grâce aux propriétés du clustering convexe, il est possible d'écrire l'évènement de sélection sous la forme d'un polyèdre. Cependant, pour obtenir un polyèdre facilement calculable, il est nécessaire de conditionner sur un peu plus que juste l'évènement de clustering en lui-même. C'est pourquoi Bachoc et al. (2023) conditionnent également sur l'ordre des observations. Cela permet ainsi d'adapter le lemme polyédrique de Lee et al. (2016) et d'utiliser la gaussienne tronquée en (3.2) pour définir des p-valeurs garantissant le contrôle de l'erreur de type I sélective dans ce contexte d'inférence post-clustering. Comme pour le test de Gao et al. (2024), ce conditionnement sur plus que juste l'évènement de clustering peut engendrer une perte de puissance statistique, mais permet en contre partie de réduire le coût computationel de la p-valeur sans pour autant affecter le contrôle de l'erreur de type I sélective du test.

Pour étendre leur test au contexte multivarié ($p > 1$), Bachoc et al. (2023) proposent d'agréger les résultats des p clusterings convexes univariés. Ils réalisent cela en utilisant l'algorithme de classification ascendante hiérarchique sur la matrice à n lignes et p colonnes contenant les labels des n observations obtenus pour chacun des p clusterings univariés. Ensuite, ils étendent leur test univarié pour évaluer les différences de moyennes sur chacune des p variables, mais cette fois-ci en considérant les clusters résultant de cette agrégation (et non plus ceux résultant du clustering convexe de la variable testée).

3.3 Méthodes basées sur une décomposition explicite de l'information

Une alternative pour maîtriser l'erreur de type I sélective, présentée par Fithian et al. (2014), consiste à diviser les données en deux ensembles indépendants. Dans cette configuration, la première partie des données est utilisée pour construire le modèle M et les hypothèses nulles $\mathcal{H}_0$, tandis que la seconde partie est réservée au test de $(M, \mathcal{H}_0)$. Comme précédemment souligné, cette méthode tire parti de l'indépendance entre les données utilisées pour la sélection du modèle et les hypothèses, et celles employées pour les tests, permettant alors de simplifier l'erreur de type I sélective à l'erreur de type I traditionnelle. En particulier, cela implique que n'importe quel test statistique peut être utilisé pour inférer la séparation entre deux clusters quelque soit la méthode de clustering utilisée pour les construire.

3.3.1 Découpage de l'échantillon

Soit n observations $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n$ d'une variable aléatoire $\boldsymbol{X} \in \mathbb{R}^p$. Une approche naïve consiste à partager aléatoirement les observations de cet échantillon de taille n en deux parties $\boldsymbol{x}_1^{(1)}, \dots, \boldsymbol{x}_{n_1}^{(1)}$ et $\boldsymbol{x}_1^{(2)}, \dots, \boldsymbol{x}_{n_2}^{(2)}$, avec $n = n_1 + n_2$ à la manière de ce qui a été proposé par Zhang et al. (2019). Il s'agit là d'une application des découpages en apprentissage/test couramment utilisés en apprentissage automatique pour résoudre les problèmes analogues de sur-apprentissage. Dans le cadre général, ce découpage permet effectivement de résoudre les problèmes d'inférence sélective, utilisant $\boldsymbol{x}_1^{(1)}, \dots, \boldsymbol{x}_{n_1}^{(1)}$ pour la sélection du modèle et des hypothèses et $\boldsymbol{x}_1^{(2)}, \dots, \boldsymbol{x}_{n_2}^{(2)}$ pour les tester. Cependant, dans le cas particulier de l'inférence post-clustering, ce type de découpage n'est pas exploitable. Il est bien possible de partitionner les n_1 observations $\boldsymbol{x}_1^{(1)}, \dots, \boldsymbol{x}_{n_1}^{(1)}$ à l'aide d'une méthode de clustering $c(\ )$, et d'ainsi générer les hypothèses de test. Mais, il est impossible de directement tester ces hypothèses sur les $\boldsymbol{x}_1^{(2)}, \dots, \boldsymbol{x}_{n_2}^{(2)}$ puisque les observations $\boldsymbol{x}_i^{(2)}, i = 1, \dots, , n_2$ que contient ce sous-échantillon ne sont pas encore labellisées : il faut d'abord transférer les clusters obtenus sur la première partie des observations. C'est ce que font Zhang et al. (2019) en entraînant un classifieur sur $\boldsymbol{x}_1^{(1)}, \dots, \boldsymbol{x}_{n_1}^{(1)}$ pour s'en servir à labelliser les observations de $\boldsymbol{x}_1^{(2)}, \dots, \boldsymbol{x}_{n_2}^{(2)}$.

Le transfert des clusters est donc indispensable mais ne peut s'effectuer sans l'utilisation d'un classifieur $\mathscr{C}(\ )$, peu importe lequel est choisi. Toutefois, cette étape de classification supervisée a pour conséquence d'apprendre le clustering, et donc les hypothèses générées sur les $\boldsymbol{x}_1^{(1)}, \dots, \boldsymbol{x}_{n_1}^{(1)}$, et de les prédire sur les $\boldsymbol{x}_1^{(2)}, \dots, \boldsymbol{x}_{n_2}^{(2)}$, qui sont ensuite réutilisées pour les tester. En particulier, pour apprendre la partition, $\mathscr{C}(\ )$ doit également apprendre les incertitudes et les potentielles différences artificielles créées dans ce sous-échantillon. Pour illustrer ce phénomène, nous avons généré $n = 200$ réalisations

d'une variable aléatoire gaussienne multivariée à $p = 2$ dimensions. D'après ce processus de génération, il est clair qu'il n'existe aucun vrai processus latent permettant de séparer les observations en 2 clusters. Nous avons ensuite partagé aléatoirement ces $n = 200$ observations en deux sous-échantillons de taille $n_1 = n_2 = 100$. Sur le premier échantillon, nous avons appliqué une classification ascendante hiérarchique de Ward pour construire $K = 2$ clusters qui ne sont donc pas représentatifs de la vraie nature des données. Puis, pour labelliser les observations du second sous-échantillon, nous avons entraîné une forêt aléatoire sur le premier sous-échantillon et avons prédit les labels des observations dans le second sous-échantillon. Finalement, nous avons testé sur chacunes des $p = 2$ variables de ce sous-échantillon une différence de moyenne entre les deux clusters prédits à l'aide du test t. Il est évident que puisque ces deux clusters n'existent pas réellement, les tests sont sous $\mathcal{H}_0$. La Figure 3.1 donne le QQ-plot des p-valeurs ainsi obtenues par rapport à la distribution uniforme sur $[0, 1]$ pour 1 000 simulations des données. Il est clair que les p-valeurs sont très éloignées de l'uniformité, en faveur d'un rejet constant de l'hypothèse nulle, révélant bien que les différences artificielles générées sur le premier sous-échantillon ont été transférées sur le second.

Dans ce contexte, l'utilisation d'un classifieur $\mathscr{C}(\,)$ soulève des problèmes d'inférence sélective. En effet, l'événement de sélection des hypothèses prend la forme $\left\{c\left(\boldsymbol{x}_1^{(1)}, \ldots, \boldsymbol{x}_{n_1}^{(1)}\right)\right\}$, mais cela est uniquement valable pour $\boldsymbol{x}_1^{(1)}, \ldots, \boldsymbol{x}_{n_1}^{(1)}$, qui ne sont pourtant pas utilisées pour les tests. Étant donné qu'il est nécessaire de passer par un classifieur pour labelliser les observations $\boldsymbol{x}_1^{(2)}, \ldots, \boldsymbol{x}_{n_2}^{(2)}$, sur cet ensemble de données, cet événement devient de la forme $\left\{\mathscr{C}\left(c\left(\boldsymbol{x}_1^{(1)}, \ldots, \boldsymbol{x}_{n_1}^{(1)}\right), \boldsymbol{x}_1^{(2)}, \ldots, \boldsymbol{x}_{n_2}^{(2)}\right)\right\}$. Ainsi, la sélection des hypothèses devient une fonction du sous-échantillon $\boldsymbol{x}_1^{(2)}, \ldots, \boldsymbol{x}_{n_2}^{(2)}$ utilisé pour les tester, ce qui ramène à nouveau les problèmes d'inférence sélective. C'est pourquoi, malgré la division en deux sous-échantillons, Zhang et al. (2019) conditionnent leurs tests sur les résultats de la SVM, afin de prendre en compte la double utilisation des $\boldsymbol{x}_1^{(2)}, \ldots, \boldsymbol{x}_{n_2}^{(2)}$.

3.3.2 Découpage d'une observation

Le problème majeur des découpages en deux sous-échantillons provient du fait que les deux sous-échantillons ne contiennent pas les mêmes observations, rendant l'utilisation d'un classifieur $\mathscr{C}(\,)$ indispensable pour transférer les résultats du clustering d'un sous-échantillon à l'autre. Motivés par ce type de découpages, Leiner et al. (2023) ont introduit la fission de données. Il s'agit d'une approche qui a également pour but de partager l'information en deux parties indépendantes, sauf que le découpage ne s'effectue plus à l'échelle de l'échantillon, mais à l'échelle de l'observation. L'idée principale de la fission de données est donc de générer, à partir d'une variable aléatoire $\boldsymbol{X}$, deux nouvelles variables aléatoires $\boldsymbol{X}^{(1)}$ et $\boldsymbol{X}^{(2)}$ contenant chacune de l'information issue de $\boldsymbol{X}$, mais incomplète. Dans ce contexte, $\boldsymbol{X}^{(1)}$ peut être utilisée pour construire le modèle et les hypothèses, et

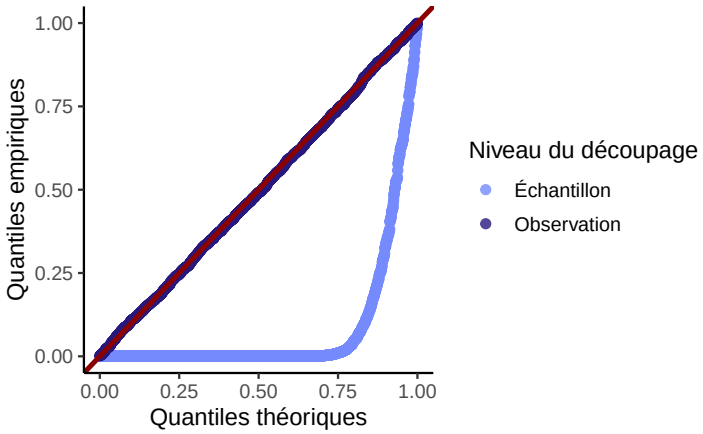


FIGURE 3.1 – **QQ-plot des p-valeurs du test t contre la distribution uniforme $\mathcal{U}[0,1]$ lorsque deux approches basées sur un découpage de l'information sont mises en oeuvre face à l'inférence post-clustering.**

$\boldsymbol{X}^{(2)}$ pour les tester. Cependant, pour que le découpage donné par la fission soit valide dans un contexte d'inférence sélective, il faut que les deux nouvelles variables aléatoires obtenues respectent l'une des deux propriétés ci-dessous :

- $\boldsymbol{\mathcal{P}_1}$: $\boldsymbol{X}^{(1)}$ et $\boldsymbol{X}^{(2)}$ sont indépendantes de distributions connues.
- $\boldsymbol{\mathcal{P}_2}$: $\boldsymbol{X}^{(1)}$ a une distribution (marginale) connue et $\boldsymbol{X}^{(2)} \mid \boldsymbol{X}^{(1)}$ a une distribution (conditionnelle) connue.

En effet, si le processus de fission satisfait la propriété $\boldsymbol{\mathcal{P}_1}$, alors, de la même manière que pour les découpages en deux sous-échantillons, l'erreur de type I sélective se ramène à l'erreur de type I classique (par indépendance), ce qui rend n'importe quel test d'hypothèse valide. Si le processus respecte la propriété $\boldsymbol{\mathcal{P}_2}$, alors l'inférence sur $\boldsymbol{X}^{(2)} \mid \boldsymbol{X}^{(1)}$ tient compte du fait qu'une partie de l'information de $\boldsymbol{X}^{(1)}$ a déjà été utilisée pour construire le modèle et les hypothèses, contrôlant ainsi l'erreur de type I sélective. La connaissance de la distribution conditionnelle de $\boldsymbol{X}^{(2)} \mid \boldsymbol{X}^{(1)}$ est donc indispensable pour développer un test contrôlant l'erreur de type I sélective. La propriété $\boldsymbol{\mathcal{P}_1}$ étant plus forte que la propriété $\boldsymbol{\mathcal{P}_2}$, elle est plus compliquée à obtenir. En particulier, Leiner et al. (2023) ont réussi à donner un processus de fission respectant $\boldsymbol{\mathcal{P}_2}$ pour toutes les distributions de la famille exponentielle alors que seules deux distributions, la gaussienne et la Poisson, possèdent un processus de fission respectant $\boldsymbol{\mathcal{P}_1}$. Neufeld et al. (2024) ont alors proposé une généralisation de la fission de données, appelée la dilution de données (*data thinning*), qui permet des décompositions garantissant $\boldsymbol{\mathcal{P}_1}$ pour une famille de distributions de probabilité plus large (incluant notamment la loi de Poisson, la loi gaussienne et la loi binomiale négative). Son objectif est similaire : décomposer une variable aléatoire $\boldsymbol{X}$ en deux nouvelles variables aléatoires $\boldsymbol{X}^{(1)}$ et $\boldsymbol{X}^{(2)}$ telles que :

$$\boldsymbol{X^{(1)} + X^{(2)} = X} \quad \text{et} \quad \boldsymbol{X}^{(1)} \perp\!\!\!\perp \boldsymbol{X}^{(2)} \tag{3.6}$$

En plus d'élargir les possibilités de décomposition en termes de distribution par rapport à la fission de données, la dilution de données permet également des décompositions en $m \geq 2$ parties indépendantes. Ces deux approches reposent toutes deux sur un compromis d'information issue de $\boldsymbol{X}$ conservée dans $\boldsymbol{X}^{(1)}$ ou $\boldsymbol{X}^{(2)}$ qui est géré par un hyperparamètre τ. Cet hyperparamètre est analogue à n_1 et n_2 de l'approche basée sur un découpage de l'échantillon, permettant d'allouer plus ou moins d'information soit pour le clustering soit pour les tests d'hypothèses.

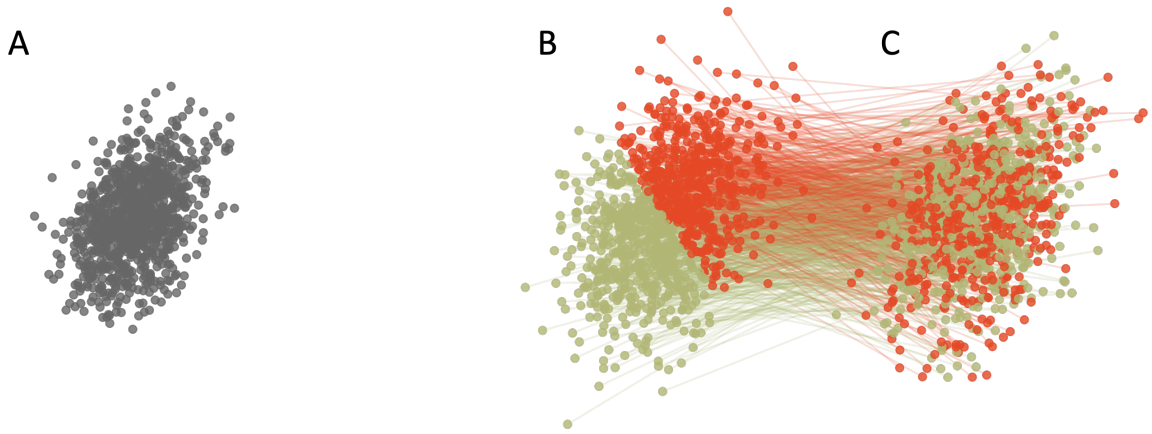


FIGURE 3.2 – **Décomposition de l'information au niveau de l'observation dans le cadre de l'inférence post-clustering**. Le panneau A montre les données originales. Le panneau B correspond aux réalisations de $\boldsymbol{X}^{(1)}$ sur lesquelles $K = 2$ clusters ont été estimés, représentés par la coloration des points. Le panneau C correspond aux réalisations de $\boldsymbol{X}^{(2)}$ dont les labels ont été assignés par un simple transfert depuis $\boldsymbol{X}^{(1)}$.

Pour répondre aux enjeux de l'inférence post-clustering qui nous intéresse, l'indépendance entre les deux nouvelles variables aléatoires est indispensable. La première variable aléatoire, $\boldsymbol{X}^{(1)}$, est alors utilisée pour construire K clusters $\widehat{C}_1, \ldots, \widehat{C}_K$ à partir de ses n réalisations $\boldsymbol{x}_1^{(1)}, \ldots, \boldsymbol{x}_n^{(1)}$ grâce à $c(\,)$ (Figure 3.2 panneau B). Les réalisations $\boldsymbol{x}_1^{(2)}, \ldots, \boldsymbol{x}_n^{(2)}$ de la seconde variable aléatoire $\boldsymbol{X}^{(2)}$, indépendante de $\boldsymbol{X}^{(1)}$, peuvent ensuite être employées pour tester les différences entre ces clusters. Grâce à cette décomposition, chaque observation $\boldsymbol{x}_i$ est scindée en deux parties indépendantes mais homologues : $\boldsymbol{x}_i^{(1)}$ et $\boldsymbol{x}_i^{(2)}$. Le transfert des clusters devient donc immédiat puisque si $\boldsymbol{x}_i^{(1)} \in \widehat{C}_k$, alors $\boldsymbol{x}_i^{(2)} \in \widehat{C}_k$, pour tout $i = 1, \ldots, n$ et pour tout $k = 1, \ldots, K$ (Figure 3.2 panneau C). L'événement de sélection des hypothèses (donné par le clustering) qui sont testées sur les n observations de $\boldsymbol{X}^{(2)}$ est donc de la forme $\left\{c\left(\boldsymbol{x}_1^{(1)}, \ldots, \boldsymbol{x}_n^{(1)}\right)\right\}$ et ne dépend que des réalisations de $\boldsymbol{X}^{(1)}$. Les données décomposées ne sont donc utilisées qu'une seule fois chacune, et grâce à leur indépendance, l'erreur de type I est contrôlée quel que soit le test d'hypothèse et la méthode $c(\,)$ de clustering utilisés. Ce bon comportement de la fission (et de la dilution) de données est représenté sur la Figure 3.1. Les méthodes basées sur un découpage de l'information au niveau de l'observation même permettent bien de conduire à des p-valeurs

uniformes (traduisant un bon contrôle de l'erreur de type I), là où, à schéma de simulation identique, l'approche basée sur un découpage de l'information au niveau de l'échantillon conduisait à une inflation de l'erreur de type I. La fission et la dilution de données semblent donc être des approches très attractives face à l'inférence post-clustering. Leur application dans ce contexte sera détaillée mais également discutée en Chapitre 5.

Chapitre 4

Tests de différences post-clustering : inférence valide et considérations pratiques avec des applications aux données écologiques et biologiques

---

Ce Chapitre est adapté de l'article : Hivert, B., Agniel, D., Thiébaut, R., & Hejblum, B. P. (2024). Post-clustering difference testing : Valid inference and practical considerations with applications to ecological and biological data. Computational Statistics & Data Analysis, 193, 107916

DOI : 0.1016/j.csda.2023.107916

Un paquet `VALIDICLUST` est disponible sur le CRAN

Contenu

4.1 Introduction 71
4.2 Méthodes 73
4.2.1 Test d'inférence sélective 74
4.2.2 Un test plus puissant en présence de clusters intercalés 77
4.2.3 Test de multimodalité 79
4.3 Résultats 80
4.3.1 Étude de simulations 80
4.3.2 Différences morphologiques chez les manchots de l'archipel de Palmer 87
4.3.3 Clustering de cellules immunitaires issues de mesures de cytométrie en flux 91
4.4 Discussion 92

4.1 Introduction

L'inférence post-clustering dans le contexte des données scRNA-seq se pose en raison de la nécessité d'identifier les gènes marqueurs de chaque cluster. Cela s'interprète donc

comme l'identification des variables responsables de la séparation entre les clusters. Face aux défis d'inférence sélective qui émergent de la double utilisation des données, une première approche pour contrôler l'erreur de type I sélective consiste à conditionner sur l'événement de clustering (ayant servi à générer les hypothèses) lors du test. Cependant, comme l'a révélé l'état de l'art des méthodes d'inférence post-clustering dans le Chapitre 3, seules deux méthodes permettent effectivement de tester la séparation de deux clusters à l'échelle de la variable tout en garantissant un contrôle de l'erreur de type I sélective : le `TN-test` (Zhang et al. 2019) et `poclin` (Bachoc et al. 2023). Néanmoins, le `TN-test` repose sur une décomposition de l'échantillon, puis l'utilisation d'un classifieur et enfin un conditionnement sur l'événement de sélection lors du test et le cumul de ces trois étapes engendre une grande perte d'information tout en complexifiant le processus de test. La méthode `poclin`, quant à elle, contraint l'utilisation du clustering convexe.

Notre objectif est donc de proposer de nouvelles méthodes pour l'inférence post-clustering valables pour toute méthode de clustering $c(\ )$ pouvant être appliquée pour partitionner les données en K clusters. De plus, nous cherchons à tester l'hypothèse nulle selon laquelle une variable donnée ne sépare pas réellement deux clusters estimés. Cette hypothèse nulle admet que la variable :

i) ne participe pas à la séparation des deux clusters et n'est pas affectée par l'étape de clustering ;

ii) ne participe à cette séparation que parce que la méthode de clustering appliquée aux données a créé des différences artificielles entre les clusters.

Il est important de noter que des objectifs similaires peuvent être abordés grâce aux méthodes de sélection de variables pour le clustering, qui permettent d'identifier simultanément le clustering optimal des données et les variables informatives (c'est-à-dire celles expliquant la partition obtenue). Ces méthodes reposent généralement sur la maximisation de critères de qualité d'ajustement du modèle (Raftery & Dean 2006, Maugis et al. 2009), ou sur une vraisemblance pénalisée (Guo et al. 2010). Néanmoins, l'inférence post-clustering – comme présentée dans ce travail – suppose que les clusters ont déjà été estimés par un algorithme de clustering. L'accent principal ici est d'évaluer la séparation entre ces clusters estimés, en tenant compte à la fois de l'incertitude introduite pendant le processus de clustering et de la double utilisation des données. Cela pose des défis qui diffèrent de l'estimation simultanée de clusters et des variables informatives pour ce partitionnement.

Pour s'affranchir du cadre strict de l'inférence sélective qui complexifie le développement des tests, nous proposons également d'étudier la séparation entre deux clusters sur une variable donnée via le prisme de la multimodalité. En effet, selon Kim et al. (2021), la multimodalité est un marqueur de la séparation des clusters : chaque mode correspond à un groupe homogène d'observations (donc à un cluster), séparé par des régions moins denses dans la distribution. Dans le contexte du clustering, Kalogeratos & Likas (2012) ont proposé un algorithme de clustering basé sur des tests d'unimodalité incrémentaux, et

Siffer et al. (2018) ont développé un test d'unimodalité pour évaluer la "clusterabilité" des données basé sur leur distribution multivariée. Ameijeiras-Alonso et al. (2021) donnent un aperçu récent des tests d'unimodalité existants, et parmi tous ces tests, trois sont plus fréquemment utilisés que les autres : *i*) le test de Silverman (Silverman 1981) basé sur une estimation à noyau de la densité f des données, *ii*) le test Dip (Hartigan et al. 1985) basé sur la fonction de répartition F des données, et *iii*) le test de masse excédentaire (Müller & Sawitzki 1991). Le test Dip est le seul qui ne nécessite pas l'estimation d'hyperparamètres ou de faire des hypothèses distributionnelles. Il a, par ailleurs, déjà été appliqué dans le cadre du clustering (Kalogeratos & Likas 2012, Wasserman et al. 2014, Schelling & Plant 2020). De plus, par rapport à plusieurs tests de multimodalité disponibles dans le package `R multimode` (Ameijeiras-Alonso et al. 2021), le test Dip offre un bon compromis entre temps de calcul et performances statistiques (Figure Supplémentaire A.1 en Annexe A).

Dans ce Chapitre, nous présentons alors trois nouvelles méthodes pour l'inférence post-clustering. Tout d'abord, en s'appuyant sur les concepts rigoureux d'inférence sélective, nous adaptons la méthode proposée par Gao et al. (2024) détaillée en Section 3.2.2.2 pour tester des hypothèses univariées, et non plus des hypothèses portant sur les barycentres des clusters, afin d'étudier si une variable donnée contient de l'information relative à la séparation de deux clusters. Pour traiter le cas où $K > 2$ clusters sont estimés, nous présentons également une extension de ce premier test basé sur une agrégation des p-valeurs entre clusters adjacents. Nous proposons enfin une autre approche qui tient compte de l'étape de clustering en examinant la présence d'un continuum dans la distribution de la variable et qui exploite le test Dip. Le Chapitre se déroule comme suit : dans la Section 4.2, nous décrivons les méthodes que nous proposons pour l'inférence post-clustering. Ces approches sont ensuite évaluées et comparées dans la Section 4.3 à l'aide de simulations numériques et de deux jeux de données réels. Quelques commentaires finaux et conseils sont disponibles dans la Section 4.4.

4.2 Méthodes

Dans ce qui suit, soit $\boldsymbol{X}$ une variable aléatoire de dimensions $n \times p$ dont la $g^{\text{ème}}$ colonne est notée $\boldsymbol{X}_g$. Sur $\boldsymbol{X}$, nous appliquons une méthode de clustering $c(\ )$ pour créer $c(\boldsymbol{X}) = \left\{\widehat{C}_1, \ldots, \widehat{C}_K\right\}$ une partition des n observations en K clusters disjoints. Nous nous intéressons à la capacité de la variable $\boldsymbol{X}_g$ à séparer deux clusters $\widehat{C}_k$ et $\widehat{C}_l$ estimés en utilisant l'ensemble des informations contenues dans $\boldsymbol{X}$ avec la méthode de clustering $c(\ )$.

4.2.1 Test d'inférence sélective

Les tests d'inférence sélective reposent généralement sur une hypothèse gaussienne des données. Ainsi, pour adapter le test d'inférence post-clustering de Gao et al. (2024) introduit au Chapitre 3 au cas univarié, nous nous appuyons également sur cette hypothèse. Cependant, alors que Gao et al. (2024) supposent une distribution matrice-normale pour l'ensemble de $\boldsymbol{X}$, dans notre cas nous avons seulement besoin que chacunes des n observations de $\boldsymbol{X}_g$ proviennent de distributions gaussiennes indépendantes avec une moyenne inconnue $\mu_{gi} \in \mathbb{R}$ et une variance connue $\sigma_g^2 \in \mathbb{R}$. Ainsi, pour tout $i \in \{1, \dots, n\}$, $X_{gi} \sim \mathcal{N}(\mu_{gi}, \sigma_g^2)$. En raison de l'indépendance supposée entre chaque X_{gi}, la distribution multivariée de $\boldsymbol{X}_g$ est une distribution gaussienne multivariée $\mathcal{N}_n\left(\boldsymbol{\mu_g}, \sigma_g^2 \boldsymbol{I_n}\right)$ avec une moyenne $\boldsymbol{\mu_g} = \begin{pmatrix} \mu_{g1} \\ \vdots \\ \mu_{gn} \end{pmatrix}$ et une matrice de covariance $\boldsymbol{\Sigma}_g = \sigma_g^2 \boldsymbol{I_n}$. Soit $\boldsymbol{x}_g$ la réalisation de $\boldsymbol{X}_g$ observée dans $\boldsymbol{X}$. Maintenant, pour un cluster $\widehat{C}_k$, soit :

$$\overline{\mu}_g^{\widehat{C}_k} = \frac{1}{|\widehat{C}_k|} \sum_{i \in \widehat{C}_k} \mu_{gi} \qquad \text{et} \qquad \overline{x}_g^{\widehat{C}_k} = \frac{1}{|\widehat{C}_k|} \sum_{i \in \widehat{C}_k} x_{gi}$$

respectivement la vraie moyenne et la moyenne empirique de la variable $\boldsymbol{X}_g$ dans le cluster $\widehat{C}_k$. Un moyen classique permettant de tester une séparation entre deux clusters $\widehat{C}_k$ et $\widehat{C}_l$ le long de $\boldsymbol{X}_g$ est de tester une différence de moyennes entre les deux clusters sur cette variable. Ainsi, les hypothèses du test sont définies comme :

$$\mathcal{H}_0 : \overline{\mu}_g^{\widehat{C}_k} = \overline{\mu}_g^{\widehat{C}_l} \qquad \text{vs} \qquad \mathcal{H}_1 : \overline{\mu}_g^{\widehat{C}_k} \neq \overline{\mu}_g^{\widehat{C}_l}. \tag{4.1}$$

En définissant un vecteur de contraste $\boldsymbol{\eta} \in \mathbb{R}^n$ par : $\eta_i = \frac{\mathbb{1}_{i \in \widehat{C}_k}}{|\widehat{C}_k|} - \frac{\mathbb{1}_{i \in \widehat{C}_l}}{|\widehat{C}_l|}$, pour tout $i = 1, \dots, n$, à la manière de Jewell et al. (2022), Gao et al. (2024), il devient alors possible de réécrire les hypothèses de test en (4.1) par :

$$\mathcal{H}_0 : \boldsymbol{\mu}_g^t \boldsymbol{\eta} = 0 \qquad \text{vs} \qquad \mathcal{H}_1 : \boldsymbol{\mu}_g^t \boldsymbol{\eta} \neq 0. \tag{4.2}$$

$\mathcal{H}_0$ dans (4.2) est donc générée par une fonction $c\left(\boldsymbol{X}\right)$ des données, ce qui nous place dans le contexte de l'inférence sélective. Conditionner sur cet événement de clustering lors du test d'hypothèse devient donc nécessaire. Pour cela, nous nous basons sur les travaux de Jewell et al. (2022) et Gao et al. (2024) pour dériver une adaptation de leur p-valeur à notre objectif d'inférence post-clustering univarié :

$$p_g^{\widehat{C}_k, \widehat{C}_l} \equiv \mathbb{P}_{\mathcal{H}_0}\left(|\boldsymbol{X}_g^t \boldsymbol{\eta}| > |\boldsymbol{x}_g^t \boldsymbol{\eta}| \mid \widehat{C}_k, \widehat{C}_l \in c(\boldsymbol{X}) \right). \tag{4.3}$$

Dans cette p-valeur, nous conditionnons sur l'estimation de $\widehat{C}_k$ et $\widehat{C}_l$ par $c(\boldsymbol{X})$, qui permettent de générer $\mathcal{H}_0$. Ainsi, la p-valeur tient compte du clustering et des incertitudes

qui lui sont associées, garantissant un contrôle de l'erreur de type I sélective. Cette p-valeur $p_g^{\widehat{C}_k,\widehat{C}_l}$ quantifie la probabilité que la différence de moyennes entre $\widehat{C}_k$ et $\widehat{C}_l$ soit aussi grande que la différence observée sous $\mathcal{H}_0$ étant donné la structure de clustering estimée. Son calcul repose alors sur toutes les réalisations possibles de $\boldsymbol{X}_g$ conduisant à la ré-estimation de $\widehat{C}_k$ et $\widehat{C}_l$ lorsque nous appliquons $c(\ )$ à $\boldsymbol{X}$. Cependant, décrire tous ces ensembles de données est difficile. Pour obtenir une p-valeur plus facilement calculable, nous suivons Jewell et al. (2022) et Gao et al. (2024) en contraignant l'aléatoire dans la variable aléatoire $\boldsymbol{X}_g$ et nous définissons notre p-valeur comme suit :

$$\tilde{p}_g^{\widehat{C}_k,\widehat{C}_l} \equiv \mathbb{P}_{\mathcal{H}_0}\left(|\boldsymbol{X}_g^t\boldsymbol{\eta}| > |\boldsymbol{x}_g^t\boldsymbol{\eta}| \ |\widehat{C}_k,\widehat{C}_l \in c(\boldsymbol{X}), \boldsymbol{\pi}_{\boldsymbol{\eta}}^{\perp}\boldsymbol{X}_g = \boldsymbol{\pi}_{\boldsymbol{\eta}}^{\perp}\boldsymbol{x}_g\right), \tag{4.4}$$

où $\boldsymbol{\pi}_{\boldsymbol{\eta}}^{\perp} = \boldsymbol{I}_n - \frac{\boldsymbol{\eta}\boldsymbol{\eta}^t}{\|\boldsymbol{\eta}\|_2^2}$ restreint la variable aléatoire $\boldsymbol{X}_g$ à un espace défini par le scalaire $\boldsymbol{\pi}_{\boldsymbol{\eta}}^{\perp}\boldsymbol{x}_g$ sans perdre le contrôle de l'erreur de type I sélective (Gao et al. 2024).

Nous souhaitons maintenant calculer la p-valeur du test d'inférence sélective définie en (4.4). Pour ce faire, nous devons écrire notre matrice de données $\boldsymbol{X}$ en fonction de notre statistique ${\boldsymbol{X}_g}^t\boldsymbol{\eta}$ et du terme résiduel $\boldsymbol{\pi}_{\boldsymbol{\eta}}^{\perp}\boldsymbol{X}_g$ où $\boldsymbol{\pi}_{\boldsymbol{\eta}}^{\perp} = \boldsymbol{I}_n - \frac{\boldsymbol{\eta}\boldsymbol{\eta}^t}{\|\boldsymbol{\eta}\|_2^2}$. Puisque, $\boldsymbol{X}_g = \boldsymbol{\pi}_{\boldsymbol{\eta}}^{\perp}\boldsymbol{X}_g + \left(\boldsymbol{I}_n - \boldsymbol{\pi}_{\boldsymbol{\eta}}^{\perp}\right)\boldsymbol{X}_g$, nous avons alors :

$$\begin{aligned}
c(\boldsymbol{X}) &= c\left([\boldsymbol{x}_1|\dots|\boldsymbol{X}_g|\dots|\boldsymbol{x}_p]\right)\\
&= c\left([\boldsymbol{x}_1|\dots|\boldsymbol{0}_n|\dots|\boldsymbol{x}_p] + [\boldsymbol{0}_n|\dots|\boldsymbol{X}_g|\dots|\boldsymbol{0}_n]\right)\\
&= c\left([\boldsymbol{x}_1|\dots|\boldsymbol{0}_n|\dots|\boldsymbol{x}_p] + \left[\boldsymbol{0}_n|\dots|\boldsymbol{\pi}_{\boldsymbol{\eta}}^{\perp}\boldsymbol{X}_g + (\boldsymbol{I}_n - \boldsymbol{\pi}_{\boldsymbol{\eta}}^{\perp})\boldsymbol{X}_g|\dots|\boldsymbol{0}_n\right]\right)\\
&= c\left([\boldsymbol{x}_1|\dots|\boldsymbol{0}_n|\dots|\boldsymbol{x}_p] + \left[\boldsymbol{0}_n|\dots|\boldsymbol{\pi}_{\boldsymbol{\eta}}^{\perp}\boldsymbol{X}_g + \left(\boldsymbol{I}_n - \boldsymbol{I}_n + \frac{\boldsymbol{\eta}\boldsymbol{\eta}^t}{\|\boldsymbol{\eta}\|_2^2}\right)\boldsymbol{X}_g|\dots|\boldsymbol{0}_n\right]\right)\\
&= c\left([\boldsymbol{x}_1|\dots|\boldsymbol{0}_n|\dots|\boldsymbol{x}_p] + \left[\boldsymbol{0}_n|\dots|\boldsymbol{\pi}_{\boldsymbol{\eta}}^{\perp}\boldsymbol{X}_g + \frac{\boldsymbol{\eta}\boldsymbol{\eta}^t}{\|\boldsymbol{\eta}\|_2}\boldsymbol{X}_g|\dots|\boldsymbol{0}_n\right]\right)\\
&= c\left([\boldsymbol{x}_1|\dots|\boldsymbol{0}_n|\dots|\boldsymbol{x}_p] + \left[\boldsymbol{0}_n|\dots|\boldsymbol{\pi}_{\boldsymbol{\eta}}^{\perp}\boldsymbol{X}_g + \frac{\boldsymbol{\eta}\phi_g}{\|\boldsymbol{\eta}\|_2^2}|\dots|\boldsymbol{0}_n\right]\right) \qquad \text{avec } \phi_g = {\boldsymbol{X}_g}^t\boldsymbol{\eta}\\
&= c\left(\left[\boldsymbol{x}_1|\dots|\boldsymbol{x}_g - \frac{\boldsymbol{\eta}\boldsymbol{\eta}^t\boldsymbol{x}_g}{\|\boldsymbol{\eta}\|_2^2} + \frac{\boldsymbol{\eta}\phi_g}{\|\boldsymbol{\eta}\|_2^2}|\dots|\boldsymbol{x}_p\right]\right)
\end{aligned}$$

Nous avons aussi :

$$\boldsymbol{X}_g^t\boldsymbol{\eta} \perp\!\!\!\perp \boldsymbol{\pi}_{\boldsymbol{\eta}}^{\perp}\boldsymbol{X}_g$$

car $\boldsymbol{\pi}_{\boldsymbol{\eta}}^{\perp}$ est la matrice de projection orthogonale sur le sous-espace orthogonal à span($\boldsymbol{\eta}$) (Jewell et al. 2022, Gao et al. 2024).

Enfin, nous avons :

$$\begin{aligned}
\boldsymbol{X}_g \sim \mathcal{N}_n(\boldsymbol{\mu_g}, \sigma_g^2\boldsymbol{I}_n) &\Rightarrow \boldsymbol{X}_g^t\boldsymbol{\eta} \sim \mathcal{N}(\boldsymbol{\mu}_g^t\boldsymbol{\eta}, \sigma_g^2\|\boldsymbol{\eta}\|_2^2)\\
&\Rightarrow \phi_g \overset{H_0}{\sim} \mathcal{N}\left(0, \sigma_g^2\|\boldsymbol{\eta}\|_2^2\right)
\end{aligned}$$

La p-valeur (4.4) est donc égale à :

$$\tilde{p}_g^{\widehat{C}_k,\widehat{C}_l} = \mathbb{P}_{\mathcal{H}_0}\left(|\phi_g| > |\boldsymbol{x}_g^t\boldsymbol{\eta}| \,|\phi_g \in S_g\right), \tag{4.5}$$

où $S_g = \left\{\phi_g : \widehat{C}_k, \widehat{C}_l \in c\left(\boldsymbol{X}\left(\phi_g\right)\right)\right\}$ est l'ensemble des perturbations de la $g^{\text{ème}}$ variable de $\boldsymbol{X}$ où à la fois $\widehat{C}_k$ et $\widehat{C}_l$ sont préservés lorsque $c(\ )$ y est ré-appliquée, et $\phi_g = \boldsymbol{X}_g^t\boldsymbol{\eta} \overset{\mathcal{H}_0}{\sim} \mathcal{N}\left(0, \sigma_g^2\|\boldsymbol{\eta}\|_2^2\right)$. Ainsi, $\boldsymbol{X}(\phi_g)$ représente une version perturbée des données $\boldsymbol{X}$ sur la $g^{\text{ème}}$ variable, où cette perturbation est définie par :

$$\boldsymbol{x_g} - \frac{\boldsymbol{\eta}\boldsymbol{\eta}^t\boldsymbol{x}_g}{\|\boldsymbol{\eta}\|_2^2} + \frac{\boldsymbol{\eta}\phi_g}{\|\boldsymbol{\eta}\|_2^2}.$$

Cette perturbation a une interprétation claire : si $|\phi_g| > |\boldsymbol{x}_g^t\boldsymbol{\eta}|$, les observations contenues dans les deux clusters sont encore plus séparées le long de $\boldsymbol{X}_g$ que ce qui est observé dans les données. Au contraire, si $|\phi_g| < |\boldsymbol{x}_g^t\boldsymbol{\eta}|$, elles sont rapprochées le long de $\boldsymbol{X}_g$ (et si $\phi_g = \boldsymbol{x}_g^t\boldsymbol{\eta}$, les données ne sont pas perturbées car dans ce cas $\boldsymbol{X}(\phi_g) = \boldsymbol{X}$). Il est important de noter que la p-valeur en (4.5) peut être réécrite comme $\mathbb{P}_{\mathcal{H}_0}\left(|\phi_g| > |\boldsymbol{x}_g^t\boldsymbol{\eta}|, \phi_g \in S_g\right) / \mathbb{P}_{\mathcal{H}_0}\left(\phi_g \in S_g\right)$. Il en découle que si $\widehat{C}_k$ et $\widehat{C}_l$ sont préservés seulement lorsque les observations sont davantage séparées les unes des autres, alors la p-valeur en (4.5) sera grande car :

$$\mathbb{P}_{\mathcal{H}_0}\left(|\phi_g| > |\boldsymbol{x}_g^t\boldsymbol{\eta}|, \phi_g \in S_g\right) \simeq \mathbb{P}_{\mathcal{H}_0}\left(\phi_g \in S_g\right).$$

Ce test d'inférence sélective s'interprète donc en termes de séparabilité entre les deux clusters considérés, même s'il se base sur une différence de moyennes, car il se réduit à quantifier la possibilité de rapprocher les observations des deux clusters tout en préservant leur séparation le long de la variable d'intérêt.

Pour explicitement décrire l'ensemble S_g tout en conservant autant de généralité que possible concernant la méthode de clustering $c(\ )$, nous suivons Gao et al. (2024) et utilisons une procédure de Monte-Carlo pour estimer $\tilde{p}_g^{\widehat{C}_k,\widehat{C}_l}$. Cette stratégie repose sur la réécriture de (4.5) sous la forme suivante :

$$\tilde{p}_g^{\widehat{C}_k,\widehat{C}_l} = \frac{\mathbb{E}\left[\mathbb{1}\left\{|\phi_g| > |\boldsymbol{x}_g^t\boldsymbol{\eta}|, \phi_g \in S_g\right\}\right]}{\mathbb{E}\left[\mathbb{1}\left\{\phi_g \in S_g\right\}\right]}. \tag{4.6}$$

Nous échantillonnons ainsi $\phi_g^1, \ldots, \phi_g^N \overset{\text{i.i.d}}{\sim} \mathcal{N}\left(0, \sigma_g^2\|\boldsymbol{\eta}\|_2^2\right)$ pour un grand nombre N de tirages, et remplaçons les espérances dans (4.6) par les sommes sur tous les échantillons. Cette procédure de Monte-Carlo évite ainsi de devoir décrire formellement l'ensemble S_g (qui dépend de $c(\ )$). Pour améliorer l'efficacité numérique, Gao et al. (2024) utilisent également une approche d'échantillonnage préférentiel initialement proposée par Yang et al. (2016) pour améliorer la probabilité de conserver le clustering dans les données

perturbées. Notre estimation de $\tilde{p}_g^{\widehat{C}_k,\widehat{C}_l}$ devient donc :

$$\tilde{p}_g^{\widehat{C}_k,\widehat{C}_l} \approx \frac{\sum_{i=1}^{N} \pi^i \mathbb{1}\left\{|\omega_g^i| \geq |\boldsymbol{x}_g^t \boldsymbol{\eta}|, \widehat{C}_k, \widehat{C}_l \in c(\boldsymbol{X}(\omega_g^i))\right\}}{\sum_{i=1}^{N} \pi^i \mathbb{1}\left\{\widehat{C}_k, \widehat{C}_l \in c(\boldsymbol{X}(\omega_g^i))\right\}}, \tag{4.7}$$

où $\omega_g^1, \ldots, \omega_g^N \sim \mathcal{N}\left(\boldsymbol{x}_g^t \boldsymbol{\eta}, \sigma_g^2 \|\boldsymbol{\eta}\|_2^2\right)$, et $\pi^i = \frac{f_1(\omega_g^i)}{f_2(\omega_g^i)}$ sont les probabilités d'échantillonnage préférentiel avec f_1 la densité d'une distribution $\mathcal{N}\left(0, \sigma_g^2 \|\boldsymbol{\eta}\|_2^2\right)$ et f_2 la densité d'une distribution $\mathcal{N}\left(\boldsymbol{x}_g^t \boldsymbol{\eta}, \sigma_g^2 \|\boldsymbol{\eta}\|_2^2\right)$. Il est important de noter que nous adaptons la méthode de Phipson & Smyth (2010) pour obtenir des estimations non biaisées de cette p-valeur de Monte-Carlo (voir Annexe A.1).

Au cœur du test ci-dessus se trouve le paramètre de variance σ_g^2, qui représente la variance de $\boldsymbol{X}_g$. Bien que σ_g^2 soit supposée connue dans le test, ce n'est pas le cas en pratique et nous proposons d'utiliser à la place l'estimation suivante :

$$\widehat{\sigma}_g^2 = \frac{1}{|\widehat{C}_k| + |\widehat{C}_l| - 1} \sum_{i \in \widehat{C}_k \cup \widehat{C}_l} \left(x_{gi} - \overline{x}_g^{\widehat{C}_k,\widehat{C}_l}\right)^2 \qquad \text{avec} \quad \overline{x}_g^{\widehat{C}_k,\widehat{C}_l} = \frac{1}{|\widehat{C}_k| + |\widehat{C}_l|} \sum_{i \in \widehat{C}_k \cup \widehat{C}_l} x_{gi}.$$

Cette estimation de variance ne prend en compte pour le test que les observations des deux clusters d'intérêt, en accord avec l'hypothèse nulle de non-séparation des deux clusters, la variance elle-même peut informer sur la séparation des données (Liu et al. 2010) comme expliqué en Chapitre 2. Cette estimation correspond à la variance estimée uniquement à partir des clusters testées, négligeant donc la variabilité supplémentaire induite par la présence d'autres clusters externes non pertinents pour le test. Cependant, dans certains cas, cette estimation de $\widehat{\sigma}_g^2$ peut sous-estimer la variance de $\boldsymbol{X}_g$, particulièrement si le clustering induit de nombreuses différences artificielles. Gao et al. (2024) s'appuient sur une estimation différente, qui surestime plutôt la variance dans la plupart des cas. Ils ont montré que le contrôle de l'erreur de type I est garanti, avec une variance surestimée, au prix d'avoir un test excessivement conservateur (voir la Figure A.2 en Annexe A pour plus de détails). Cela met en lumière la difficulté intrinsèque de l'estimation de la variance dans le contexte de l'inférence post-clustering qui sera notamment rediscutée en Chapitre 5.

4.2.2 Un test plus puissant en présence de clusters intercalés

Le test de Gao et al. (2024) a été conçu pour comparer une paire de clusters, testant une différence de barycentres. Mais lorsqu'il s'agit de tester des différences à l'échelle univarié, comme notre test d'inférence sélective, des clusters intermédiaires supplémentaires peuvent compliquer la comparaison entre deux clusters. En pratique, les données comportent souvent plus de 2 clusters, et, dans une dimension de l'espace, cela peut sé-

rieusement compromettre la puissance statistique du test, même pour des clusters bien séparés. En effet, s'il existe un ou plusieurs clusters intercalés entre les deux clusters comparés dans la dimension testée, il devient rapidement impossible de les rapprocher sans changer le regroupement (voir Figure A.3 en Annexe A). Pour surmonter cette limitation, nous proposons d'étendre notre test d'inférence sélective en supposant que deux clusters estimés $\widehat{C}_k$ et $\widehat{C}_l$ sont séparés sur $\boldsymbol{X}_g$ si et seulement si au moins une paire intercalée de clusters adjacents entre eux est séparée. Cela signifie qu'à l'inverse, s'il y a un continuum sur $\boldsymbol{X}_g$ pour passer de $\widehat{C}_k$ à $\widehat{C}_l$, alors $\widehat{C}_k$ et $\widehat{C}_l$ ne sont pas séparés. En ne considérant que des tests qui portent sur des paires de clusters adjacents, nous conservons la puissance statistique du test d'inférence sélective. Nous proposons alors de combiner toutes les p-valeurs des tests d'inférence sélective entre paires de clusters adjacents intercalés pour en déduire une p-valeur agrégée évaluant globalement la séparation de $\widehat{C}_k$ et $\widehat{C}_l$ sur $\boldsymbol{X}_g$.

Pour identifier les clusters intercalés entre $\widehat{C}_k$ et $\widehat{C}_l$ sur $\boldsymbol{X}_g$, nous définissons l'ensemble :

$$\mathcal{C}_g^{k:l} := \left\{ \widehat{C}_j, j = 1, \dots, K \ / \ \overline{x}_g^{\widehat{C}_j} \in \left[\min\left(\overline{x}_g^{\widehat{C}_k}, \overline{x}_g^{\widehat{C}_l} \right), \max\left(\overline{x}_g^{\widehat{C}_k}, \overline{x}_g^{\widehat{C}_l} \right) \right] \right\},$$

où $\overline{x}_g^{\widehat{C}_j} = \frac{1}{|\widehat{C}_j|} \sum_{i \in \widehat{C}_j} x_{gi}$ permet implicitement de trier les clusters selon leur moyenne empirique sur $\boldsymbol{X}_g$. Nous définissons deux clusters $\widehat{C}_{m_1}$ et $\widehat{C}_{m_2}$ comme adjacents sur $\boldsymbol{X}_g$ si $\mathcal{C}_g^{m_1:m_2} = \left\{ \widehat{C}_{m_1}, \widehat{C}_{m_2} \right\}$. Ainsi, si $|\mathcal{C}_g^{k:l}| = M$, il existe $M-1$ paires de clusters adjacents dans l'ensemble $\mathcal{C}_g^{k:l}$ des clusters intercalés entre $\widehat{C}_k$ et $\widehat{C}_l$, traçant alors un chemin entre ces deux clusters. Nous pouvons alors définir :

$$p_g^{\widehat{C}_k:\widehat{C}_l} := f\left(p_g^1, \dots, p_g^{M-1}\right),$$

où f est une fonction d'agrégation de p-valeurs comme décrite par Vovk & Wang (2020). Sur la base de simulations numériques, nous favorisons l'utilisation de la moyenne harmonique, recommandée par Vovk & Wang (2020) en cas de dépendances potentielles entre les p-valeurs. C'est ici notre cas puisque chaque cluster contribue au calcul de deux p-valeurs. L'utilisation de cette moyenne présente alors un bon compromis entre erreur de type I et puissance statistique (voir Figure A.4 en Annexe A). Ainsi, la p-valeur du test d'inférence sélective par agrégation de clusters adjacents est définie par :

$$p_g^{\widehat{C}_k:\widehat{C}_l} = \min\left(e \log(M-1) \frac{M-1}{\sum_{i=1}^{M-1} \frac{1}{p_g^i}}, 1 \right).$$

Il convient de noter que toutes les $p_g^1, \dots, p_g^{M-1}$ sont calculées ici en utilisant une même estimation de la variance. Nous proposons donc dans ce cas d'utiliser une estimation de

σ_g^2 tenant compte de toutes les observations appartenant à l'ensemble $\mathcal{C}_g^{k:l}$ des clusters intercalés entre $\widehat{C}_k$ et $\widehat{C}_l$:

$$\hat{\sigma}_g^2 = \frac{1}{|\mathcal{C}_g^{k:l}| - 1} \sum_{C \in \mathcal{C}_g^{k:l}} \sum_{i \in C} \left(x_{gi} - \overline{x}_g^{\mathcal{C}_g^{k:l}} \right)^2 \qquad \text{avec} \quad \overline{x}_g^{\mathcal{C}_g^{k:l}} = \frac{1}{|\mathcal{C}_g^{k:l}|} \sum_{C \in \mathcal{C}_g^{k:l}} \sum_{i \in C} x_{gi}.$$

4.2.3 Test de multimodalité

Nous proposons d'exploiter la notion de continuum pour décrire l'absence de séparation entre deux clusters : si $\widehat{C}_k$ et $\widehat{C}_l$ sont séparés, alors il doit exister un creux dans la distribution de la variable entre ces deux clusters, c'est-à-dire que sa distribution est multimodale. En revanche, s'il y a un continuum, alors ils ne peuvent pas réellement être séparés, c'est-à-dire que la distribution de cette variable est unimodale. Bien qu'un mauvais clustering puisse induire une multimodalité artificielle, la présence ou non d'un tel continuum dans la distribution de la variable ne peut être causée par le clustering.

Cette deuxième approche peut alors être considérée comme une version simplifiée du test d'inférence sélective. En effet, lorsque l'on perturbe les données pour vérifier s'il est possible de rapprocher les deux clusters sans changer le clustering, nous évaluons en réalité à quel point il est probable d'observer un continuum entre $\widehat{C}_k$ et $\widehat{C}_l$. S'il y a un continuum entre $\widehat{C}_k$ et $\widehat{C}_l$ sur $\boldsymbol{X}_g$, alors sa distribution doit être unimodale. Ainsi, pour étudier la séparabilité de ces deux clusters sur $\boldsymbol{X}_g$, il suffit d'appliquer un test d'unimodalité à sa distribution restreinte uniquement aux individus des clusters de l'ensemble $C_g^{k:l}$. En effet, si $\boldsymbol{X}_g$ sépare $\widehat{C}_k$ et $\widehat{C}_l$, alors il y a au moins deux clusters dans $C_g^{k:l}$ qui sont séparés l'un des autres, et en particulier, puisque ces clusters sont entre $\widehat{C}_k$ et $\widehat{C}_l$, il y a aussi une séparation entre eux sur $\boldsymbol{X}_g$.

Un test d'unimodalité teste l'hypothèse nulle selon laquelle « la distribution de $\boldsymbol{X}_g$ est unimodale » par rapport à une hypothèse l'alternative selon laquelle « la distribution de $\boldsymbol{X}_g$ est multimodale ». Nous proposons alors d'utiliser le test Dip introduit par Hartigan et al. (1985). Ce test repose sur la statistique du creux : $\text{dip}(F) = \min_{G \in \mathcal{U}} \rho(F, G)$, où $\rho(F, G) = \sup_x |F(x) - G(x)|$ et $\mathcal{U}$ est la classe des distributions unimodales. Ainsi, la statistique du creux s'interprète comme la distance de F à l'ensemble des fonctions unimodales et mesure donc l'écart de notre distribution par rapport à l'unimodalité. Si F est unimodale, alors $\text{dip}(F) = 0$, et inversement si F est multimodale, alors $\text{dip}(F) > 0$. En pratique, les p-valeurs sont calculées comme :

$$p_{\widehat{D}_n} := \mathbb{P}\left(d_{U_n} \geq \widehat{D}_n \right),$$

où $d_{U_n} = \text{dip}(\hat{F}_{U_n})$ est la statistique du creux calculée pour un échantillon de taille n issu de la distribution uniforme $\mathcal{U}[0, 1]$ de fonction de répartition empirique $\hat{F}_{U_n}$, $\widehat{D}_n = \text{dip}(\hat{F}_n)$ est la statistique du creux observée, calculée à l'aide de la fonction de répartition empirique

$\hat{F}_n$ des données, et n est la taille de l'échantillon. Hartigan et al. (1985) ont montré que la distribution uniforme est la distribution unimodale avec la plus grande statistique asymptotique du creux parmi les distributions unimodales, il s'agit donc intuitivement du candidat le moins favorable pour l'unimodalité. Ainsi, une distribution avec une statistique du creux plus grande que celle de la distribution uniforme ne peut pas être considérée comme unimodale (au sens de la statistique du creux). $p_{\widehat{D}_n}$ est alors interprétée comme la probabilité sous l'hypothèse nulle d'unimodalité que la distribution uniforme ait une statistique du creux plus grande que la statistique du creux observée de $\widehat{F}_n$.

Dans notre contexte, nous appliquons donc le test Dip à la distribution de la variable $\boldsymbol{X}_g$ restreinte aux individus qui sont dans les clusters de l'ensemble $\mathcal{C}_g^{k:l}$ pour tester l'existence d'un continuum entre $\widehat{C}_k$ et $\widehat{C}_l$ et ainsi tester leur séparation.

4.3 Résultats

4.3.1 Étude de simulations

Nous présentons ici les résultats évaluant le comportement des tests d'inférence post-clustering proposés dans la Section 4.2, tant en termes de contrôle de l'erreur de type I que de puissance statistique. Sauf indication contraire, toutes les analyses sont réalisées au niveau de significativité $\alpha = 5\%$, et le clustering se fait par la classification ascendante hiérarchique de Ward associée à la distance euclidienne.

4.3.1.1 Cadre bidimensionnel

Nous avons généré des données bidimensionnelles ($p = 2$) selon deux scénarios : (i) d'abord sans vrais clusters séparés, où les observations sont issues d'une distribution gaussienne $\mathcal{N}(\mathbf{0}, \boldsymbol{I}_2)$; et (ii) avec trois clusters des mêmes tailles $n/3$ provenant de distributions gaussiennes $\mathcal{N}(\boldsymbol{\mu}^{C_j}, \boldsymbol{I}_2)$ tel que $\boldsymbol{\mu}^{C_1} = (-5, 0)$, $\boldsymbol{\mu}^{C_2} = (5, 0)$ et $\boldsymbol{\mu}^{C_3} = (0, 10)$, ainsi X_1 sépare les trois clusters tandis que X_2 ne sépare que C_3 des deux autres clusters, ce qui signifie que X_2 est sous l'hypothèse nulle pour la comparaison entre C_1 et C_2. Dans les deux cas, nous avons appliqué la classification ascendante hiérarchique de Ward pour construire trois clusters. La Figure 4.1A montre un exemple de données sous chaque scénario. Dans le premier scénario, les clusters ont été créés en forçant des différences entre des groupes d'observations, engendrant ainsi des différences artificielles, tandis que dans le deuxième scénario, les clusters estimés sont représentatifs de la véritable structure des données, chaque cluster étant un groupe homogène et séparé d'observations issues d'une composante du mélange.

La Figure 4.1B montre les résultats des trois tests proposés comparés à ceux obtenus à l'aide du test t habituel pour 2 000 simulations des données, chacune avec une

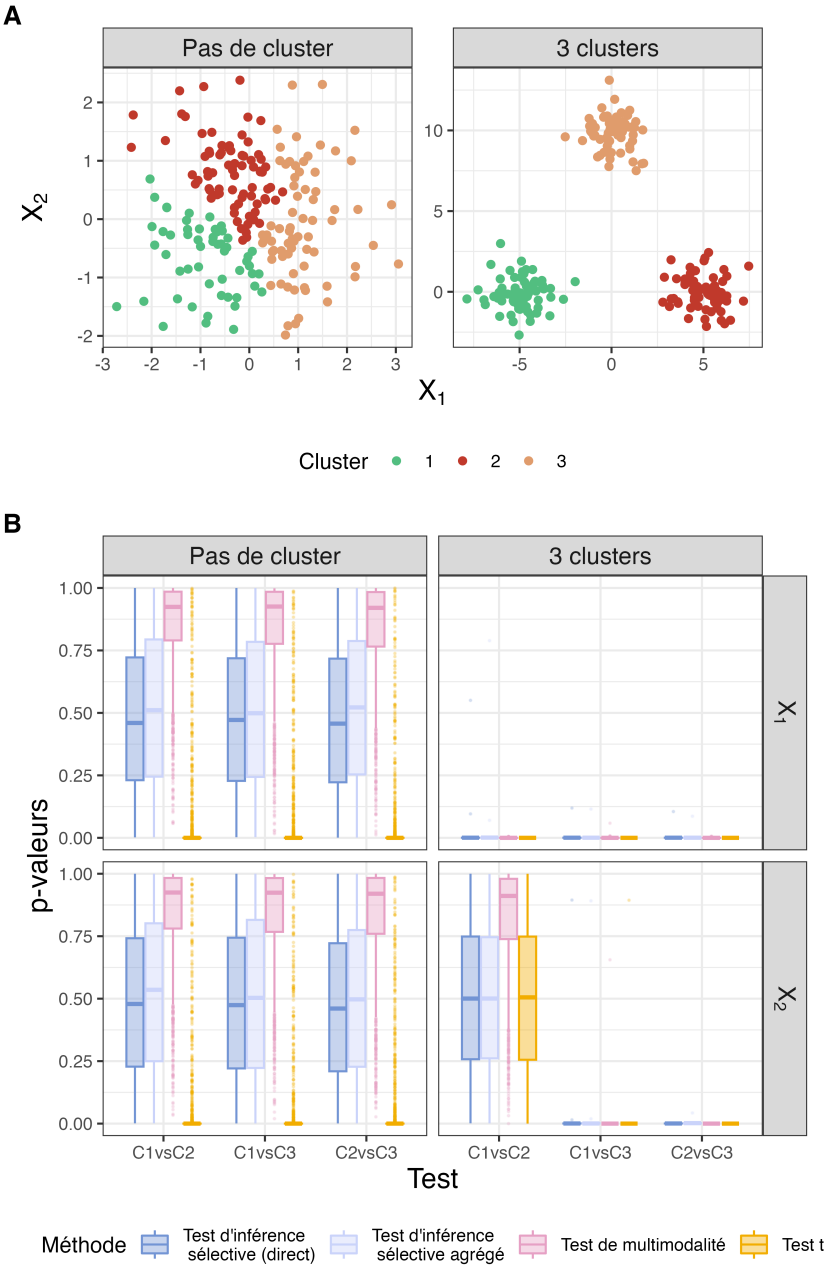


FIGURE 4.1 – **Validité des *p*-valeurs renvoyées par nos tests, comparaison avec le test *t*.** Panneau A : Processus de génération des données. Deux cas sont étudiés : un cas sous un $\mathcal{H}_0$ global traduisant une absence totale de clusters dans les données (Pas de cluster) et un cas avec trois vrais clusters (3 clusters). Dans les deux cas, la classification ascendante hiérarchique de Ward est appliquée sur les données pour construire $K = 3$ clusters. Panneau B : Distribution des p-valeurs pour les tests concernant chaque paire de clusters possible sur chacune des deux variables pour 2 000 simulations des données.

taille d'échantillon de $n = 200$. Pour le scénario sans cluster, le test t renvoie des p-valeurs extrêmement petites qui traduisent une inflation de l'erreur de type I $(90,70\%)$. En fait, le test t identifie les différences artificielles créées lors du clustering. En tenant compte de cette étape de clustering, les p-valeurs du test d'inférence sélective $p_g^{\widehat{C}_k,\widehat{C}_l}$ et

celles résultant de son extension par agrégation $p_g^{\widehat{C}_k:\widehat{C}_l}$ sont relativement bien uniformément réparties sur l'intervalle $[0, 1]$. Cela garantit un bon calibrage des p-valeurs et un contrôle effectif de l'erreur de type I (respectivement $6,53\%$ et $5,98\%$). Quant au test de multimodalité, ses p-valeurs sont très conservatrices mais cohérentes avec l'absence de séparation des clusters (avec une erreur de type I égale à $0,11\%$). Cela s'explique par la distribution unimodale de référence utilisée par le test Dip, qui est la distribution uniforme, alors que les données sont générées à partir d'une distribution gaussienne, qui a donc une statistique de creux inférieure à celle de la distribution uniforme. Ces bons résultats se confirment dans le scénario à 3 clusters lors de la comparaison entre C_1 et C_2 le long de X_2 (où $\mathcal{H}_0$ est vérifiée). Pour toutes les autres comparaisons dans ce scénario, les 4 tests détectent correctement la séparation, rejettant à raison $\mathcal{H}_0$ (avec une puissance statistique supérieure à $99,9\%$ tous cas confondus).

Il convient de noter que si le clustering ne force pas de différences entre des groupes d'observations, en identifiant correctement la structure en clusters des données, le test t contrôle également l'erreur de type I, comme le montre le cas à 3 clusters dans la Figure 4.1B ($4,65\%$). Cela illustre davantage le lien entre différences artificielles et estimation du nombre de clusters. Cependant, bien que le test t puisse donner des résultats corrects lorsque les méthodes de clustering identifient correctement le processus latent réel sous-jacent aux données, il n'est jamais valide car il ne tient pas compte des incertitudes dans la partition obtenue et du fait que les hypothèses de tests soient basées sur les données elles-mêmes.

La taille de l'échantillon a un impact relativement faible sur la puissance statistique des tests, à l'exception du test de multimodalité qui est non paramétrique. La puissance statistique des tests d'inférence sélective est plus impactée par la différence de moyennes δ entre les deux clusters que par la taille de l'échantillon. Tous les tests atteignent une puissance statistique assez raisonnable pour une taille d'échantillon modérée ($n = 50$, voir Figure A.5 en Annexe A).

4.3.1.2 Focus sur la puissance statistique dans un cadre univarié

Nous avons également généré des données à partir d'un mélange univarié de deux distributions gaussiennes avec des proportions et des variances égales : $0.5\mathcal{N}(0, 1) + 0.5\mathcal{N}(\delta, 1)$, où les deux composantes sont séparées par une différence de moyennes δ – également appelé modèle de contamination (Laurent et al. 2018). Chacune des composantes du mélange représente donc un cluster, et la force de leur séparation est gérée par δ. Nous avons appliqué la classification ascendante hiérarchique de Ward pour construire soit $K = 2$ soit $K = 4$ clusters. La Figure 4.2A montre une réalisation de cette simulation. Dans le cas à 2 clusters, le clustering a effectivement identifié la véritable structure des données, tandis que dans le cas à 4 clusters, de faux clusters ont été estimés. Nous avons évalué la

puissance statistique des trois tests proposés pour détecter la séparation entre les clusters 1 et 2, les deux clusters les plus extrêmes dans la distribution des données, selon δ, en utilisant $N = 2\,000$ générations des données.

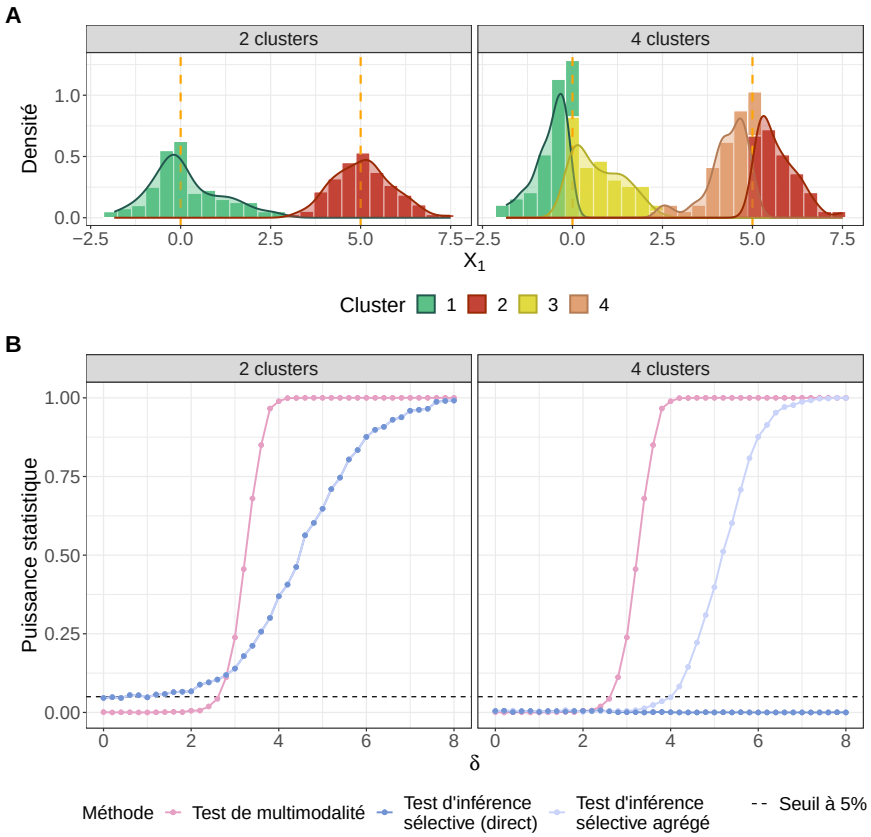


FIGURE 4.2 – **Puissance statistique des tests proposés.** Panneau A : Illustration du processus de génération de données. Les données sont générées selon un mélange gaussien univarié à deux composantes (avec proportion et variance égales) séparées par une différence de moyennes δ (modèle de contamination). Deux cas sont étudiés : un cas où le nombre réel de clusters est estimé ($K = 2$ clusters) et un cas où plus de clusters que le vrai nombre sont estimés ($K = 4$ clusters). La ligne pointillée orange représente la moyenne de chaque composante. Panneau B : Puissance statistique pour la séparation du cluster 1 et du cluster 2 (au seuil de 5%) des tests proposés selon la différence de moyennes δ séparant les deux composantes du mélange gaussien.

La Figure 4.2B synthétise les résultats pour les 2 000 simulations des données avec une taille d'échantillon $n = 200$ (100 observations par classe). La puissance statistique augmente avec δ. Le test de multimodalité semble être le plus puissant dans ce contexte, surtout lorsque $\delta \geq 3$, une valeur remarquable pour la multimodalité dans ce genre de mélange comme précédemment expliqué (Siffer et al. 2018). De plus, comme dans le cas à 4 clusters, tous les clusters sont entre le Cluster 1 et le Cluster 2, la puissance statistique atteinte par le test de multimodalité est exactement la même que dans le cas à 2 clusters. Dans le cas à 2 clusters, le test d'inférence sélective direct et le test d'inférence sélective par agrégation des p-valeurs adjacentes ont une puissance statistique identique, puisque seuls

deux clusters sont estimés, ils sont nécessairement adjacents et donc le test d'inférence sélective direct est exactement le même que le test par agrégation. En revanche, le test d'inférence sélective direct échoue dans le cas à 4 clusters, quelle que soit la valeur de δ. En effet, il est impossible de rapprocher le cluster 1 du cluster 2 sans les mélanger avec les clusters 3 et 4. Le test par agrégation des p-valeurs adjacentes permet de corriger ce manque de puissance. À noter qu'il reste valide lorsque le nombre de p-valeurs à agréger augmente.

4.3.1.3 Cadre multidimensionnel avec différents algorithmes de clustering

Pour étudier plus en détail le comportement des tests post-clustering proposés, il est essentiel d'étudier leurs performances lorsqu'ils sont appliqués avec différents algorithmes de clustering. Nous avons sélectionné les algorithmes de clustering les plus courants en pratique et majoritairement présentés en Section 1.2.1, à savoir la classification ascendante hiérarchique de Ward, les k-means, les fuzzy $c-$means et le clustering basé sur un modèle gaussien, avec `Mclust` (Scrucca et al. 2016).

En s'inspirant des travaux de Maugis et al. (2009) dans le cadre de la sélection de variables pour le clustering, nous avons généré des données selon un modèle de mélange gaussien multivarié à deux composantes :

$$\boldsymbol{X}^{1,2} \sim \pi_1 \mathcal{N}_2\left(\boldsymbol{\mu}^{C_1}, \boldsymbol{\Sigma}\right) + \pi_2 \mathcal{N}_2\left(\boldsymbol{\mu}^{C_2}, \boldsymbol{\Sigma}\right) \text{ où } \boldsymbol{\mu}^{C_1} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \boldsymbol{\mu}^{C_2} = \begin{pmatrix} 0 \\ \delta \end{pmatrix} \text{ et } \Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$$

Nous avons augmenté ce mélange en introduisant trois variables supplémentaires : $X^{3,4,5} = \boldsymbol{X}^{1,2}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ où $\boldsymbol{\varepsilon} \sim \mathcal{N}_3\left(\mathbf{0}_3, \boldsymbol{I}_3\right)$. Nous avons alors considérés trois scénarios permettant d'étudier l'impact de différentes caractéristiques liées i) à la taille des clusters (π_i), ii) aux corrélations entre les variables informatives (ρ), et iii) au nombre de variables informatives pour la séparation des clusters ($\boldsymbol{\beta}$). Le premier scénario présente une configuration relativement simple avec des tailles de clusters équilibrées, une seule variable informative sur la séparation des clusters (X^2), et aucune corrélation entre les variables. En revanche, les deux scénarios restants introduisent une plus grande complexité, avec des tailles de clusters déséquilibrées, des corrélations entre les variables, et deux variables informatives pour la séparation (X^2 et X^4). La distinction entre les scénarios $Sc2$ et $Sc3$ se trouve au niveau des variables corrélées : dans le scénario 2, aucune variable informative n'est corrélée avec une variable de bruit (ne contribuant pas à la séparation des clusters), seules les variables informatives et celles de bruit sont respectivement corrélées entre elles. Au contraire, dans le scénario 3, une variable de bruit (X^1) est corrélée avec une variable informative (X^2) à cause de la valeur non-nulle de ρ. Les valeurs détaillées des paramètres pour chaque scénario sont fournies dans la Table 4.1. Nous avons effectué 1 000 simulations des données avec une taille d'échantillon de $n = 200$ pour évaluer la capacité de nos tests à retrouver les variables qui séparent effectivement les deux clusters. Ces simulations

couvrent une gamme de forces de séparation caractérisées par $\delta \in \{0, 2.5, 4, 8\}$. $\delta = 0$ correspond à l'hypothèse nulle globale d'absence de vrais clusters, $\delta = 2.5$ illustre le cas d'une faible séparation, où les algorithmes de clustering risquent d'échouer à identifier la vraie structure en clusters des données, et $\delta = 4$ et $\delta = 8$ sont respectivement les cas d'une séparation modérée et d'une forte séparation, assurant de bonnes performances des algorithmes de clustering.

Scénario	π_1	ρ	$\boldsymbol{\beta}$
$Sc1$	0.5	0	$\begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$
$Sc2$	0.3	0	$\begin{pmatrix} 0.5 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}$
$Sc3$	0.3	0.3	$\begin{pmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}$

TABLE 4.1 – **Valeurs des paramètres dans les trois scénarios de simulations inspirés des travaux de Maugis et al. (2009).**

Les quatre algorithmes de clustering montrent des performances relativement comparables en termes d'ARI pour identifier la vraie partition des données (Figure 4.3A). Comme expliqué au Chapitre 2, dans ce cas où la variance du bruit est identique à la variance des variables informatives, δ est le principal déterminant de la qualité du clustering. Les performances de nos tests post-clustering sont donc principalement influencées par les caractéristiques des données plutôt que par les algorithmes de clustering eux-mêmes.

Dans la Figure 4.3B, une variable est considérée comme sélectionnée si sa p-valeur est inférieure à $\alpha = 5\%$, évaluant l'erreur de type I sous $\mathcal{H}_0$ et la puissance statistique sous $\mathcal{H}_1$. Lorsqu'il n'y a aucune corrélation entre les variables (scénario $Sc1$), les deux tests post-clustering i) contrôlent correctement l'erreur de type I, et ii) parviennent à identifier les variables effectivement sous $\mathcal{H}_1$, avec une puissance dépendant de δ. Ceci est cohérent avec les résultats observés dans la Figure 4.2. Dans le cas sans séparation ($\delta = 0$) ou lorsqu'elle est faible ($\delta = 2.5$), toutes les méthodes de clustering montrent de mauvaises performances pour identifier la vraie partition des données. Cela signifie que le test de séparation s'effectue entre des groupes d'observations ne correspondant pas à de vrais clusters. Dans ce cas, nos tests post-clustering contrôlent efficacement l'erreur de type I, pour $\delta = 2.5$, ils ont une puissance très faible pour identifier la variable X^2 qui est effectivement sous $\mathcal{H}_1$. Dans les scénarios comprenant des corrélations entre les variables, c'est-à-dire les scénarios $Sc2$ et $Sc3$, le comportement du test de multimodalité reste inchangé. Au contraire, le test d'inférence sélective peut avoir du mal à contrôler l'erreur de type I associée aux variables corrélées. Notamment dans le scénario $Sc3$, lorsqu'une variable de bruit (c'est-à-dire sous $\mathcal{H}_0$) est corrélée avec une variable informative, le test d'inférence sélective ne parvient plus à contrôler l'erreur de type I associée à la variable

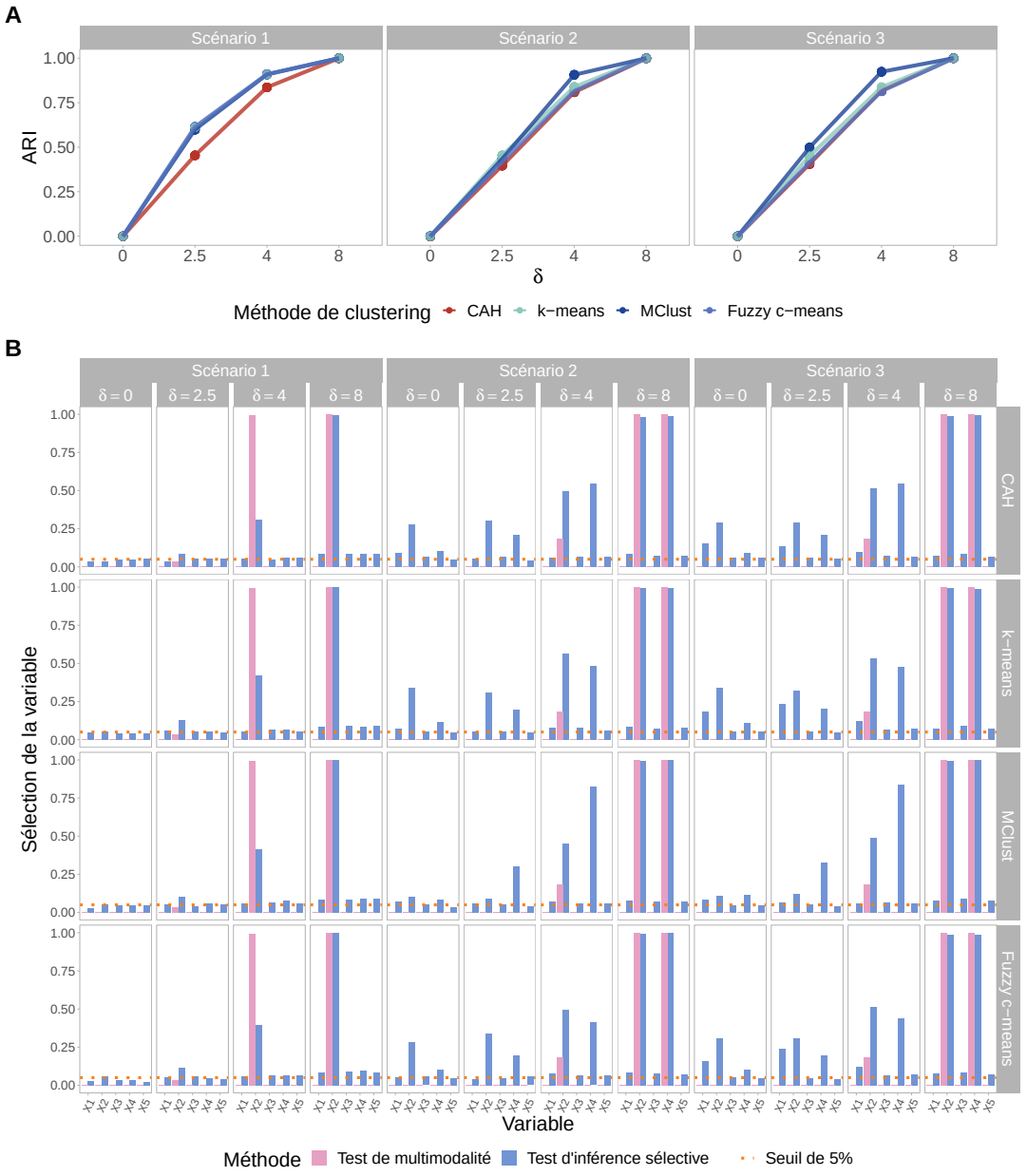


FIGURE 4.3 – **L'inférence post-clustering est principalement influencée par la structure des données plutôt que par le choix de l'algorithme de clustering.** Panneau A : Performances du clustering pour identifier la véritable structure des données en fonction de la séparation entre les clusters donnée par δ. Panneau B : Proportion moyenne de fois où les tests post-clustering ont sélectionné chacune des cinq variables.

de bruit. En revanche, dans le scénario $Sc2$, lorsque les données présentent des clusters clairs, avec $\delta \in \{4, 8\}$, et que des corrélations existent uniquement entre les variables informatives, l'erreur de type I est bien calibrée. Cela met en évidence l'importance de l'hypothèse d'indépendance pour le test d'inférence sélective. En effet, en présence de corrélations, les perturbations des données, qui ont seulement lieu à l'échelle univariée, sont insuffisantes pour garantir un bon calibrage du test, voir la Section A.13 de l'Annexe A pour plus de détails.

Nous avons également comparé nos tests aux méthodes de sélection de variables pour le clustering basé sur un modèle de mélange gaussien (voir la Figure A.6 en Annexe A). Nos deux tests, qui tiennent compte de la double utilisation des données, offrent non seulement un meilleur contrôle de l'erreur de type I pour toutes les variables sous une hypothèse nulle globale d'absence de clusters, mais démontrent également une précision supérieure dans la sélection des variables informatives lorsqu'une vraie séparation existe. De plus, le test de multimodalité présente systématiquement de meilleures performances computationnelles par rapport à la sélection de variables dans le cadre des mélanges gaussiens, même lorsqu'il est utilisé conjointement avec un clustering basé sur ce modèle. En revanche, le test d'inférence sélective, qui nécessite que le clustering soit appliqué sur chaque version perturbée des données, se révèle plus sensible au choix de l'algorithme de clustering. Ses temps de calculs sont, par contre, nettement plus élevé lorsqu'il est associé à un clustering basé sur un modèle gaussien.

4.3.2 Différences morphologiques chez les manchots de l'archipel de Palmer

Pour évaluer davantage les tests d'inférence post-clustering proposés, nous avons également analysé des données réelles issues de l'écologie et disponibles dans le package `R` `palmerpenguins` (Horst et al. 2020). Ce jeu de données de référence pour le clustering comprend $p = 4$ variables : longueur du bec (en mm), profondeur du bec (en mm), longueur de la nageoire (en mm) et masse corporelle (en g), pour 344 manchots. Après avoir supprimé les observations contenant des valeurs manquantes pour au moins l'une des 4 variables, $n = 333$ observations ont été conservées dans notre analyse. Les manchots proviennent de trois espèces différentes : Adélie, *Chinstrap* et *Gentoo*, avec respectivement 146, 68 et 119 individus.

4.3.2.1 Contrôle négatif

Nous avons commencé par sélectionner uniquement des manchots femelles de l'espèce *Gentoo* afin de créer un jeu de données de contrôle négatif. En effet, étant donné que ce jeu de données ne contient que des observations de la même espèce et du même sexe, il ne devrait pas y avoir de différence morphologiques notables entre les individus. Nous avons appliqué une classification ascendante hiérarchique de Ward sur les z-scores, en raison des problèmes d'unité entre les variables, pour construire 3 clusters. Comme aucun réel processus latent définissant une partition n'existe dans ce sous-ensemble d'individus, le clustering a donc forcé des différences entre les clusters pour les construire. La Table A.1 en Annexe A présente les p-valeurs de chacun des 3 tests proposés ainsi que celles du test t pour toutes les paires de clusters le long de chacune des quatre variables. Une fois de plus, le test t conduit à de nombreuses fausses découvertes. Au contraire, nos 3 tests

d'inférence post-clustering se comportent correctement en n'identifiant aucune des quatre mesures comme séparant significativement une paire de clusters.

4.3.2.2 Analyse complète des données

Nous avons ensuite inclu les $n = 333$ manchots dans notre analyse et avons examiné les données comme si nous ne connaissions pas l'espèce des manchots. Puisque les trois espèces sont dorénavant présentes, nous voulons identifier quelles caractéristiques morphologiques les différencient (les séparent). La Figure 4.4A présente la densité des quatre variables (z-scores) au sein de chacunes des trois espèces de manchots. Les manchots Adélie et *Chinstrap* semblent plus difficiles à distinguer, uniquement la longueur de leur bec semble les différencier : les manchots *Chinstrap* ont des becs plus grands, comparables à ceux des manchots *Gentoo*, que les manchots Adélie. L'espèce de manchot *Gentoo* est la plus facile à identifier, ayant des caractéristiques morphologiques clairement différentes des deux autres espèces en ce qui concerne les trois autres variables mesurées. Une fois de plus, nous avons appliqué la classification ascendante hiérarchique de Ward sur les z-scores. La Figure 4.4B présente les résultats de ce clustering. Le dendrogramme a été coupé de telle sorte à construire trois clusters. Ils réflètent bien la véritable composition des vraies espèces : le Cluster 2 et le Cluster 3 contiennent chacun uniquement des manchots *Gentoo* et *Chinstrap* (respectivement), tandis que le Cluster 1 contient un mélange de deux espèces (100% des manchots Adélie et 11 manchots *Chinstrap*).

La Table 4.2 présente les p-valeurs de chacun des trois tests d'inférence post-clustering proposés comparées à celles du test t pour toutes les paires de clusters le long de chacunes des quatre variables. Puisque les clusters identifiés correspondent bien aux trois véritables espèces de manchots, l'étape de clustering ne peut pas induire de différences artificielles, et donc les résultats du test t peuvent être utilisés comme référence. Seule une comparaison n'est pas significative au seuil de 5% selon le test t : la profondeur du bec n'est pas significativement différente entre les manchots Adélie (Cluster 1) et les manchots *Chinstrap* (Cluster 3), ce qui est cohérent avec l'observation visuelle en Figure 4.4A. Le test de multimodalité semble manquer de puissance statistique ici, mais il est clair d'après la distribution des variables représentée dans la Figure 4.4A que seulement quelques comparaisons présentent une multimodalité notables (le Cluster 1/Adélie comparé aux deux autres clusters pour la longueur de la nageoire par exemple). Les deux tests d'inférence sélective identifient davantage de différences significatives (6/11 pour le test direct et 7/11 pour le test par agrégation qui est plus robuste lorsque plus de 2 clusters sont estimés). Les séparations manquées peuvent s'expliquer par le manque de puissance statistique pour détecter de petites différences (voir Table A.2 en Annexe A). En tenant compte de l'étape de clustering, le test d'inférence sélective perd en puissance notamment par rapport à d'autres tests comme le test t, qui est le test uniformément le plus puissant

(Lehmann 2012). Mais cela vient notamment du fait qu'ils tiennent compte de l'étape de clustering, le rendant valide dans ce contexte d'inférence post-clustering.

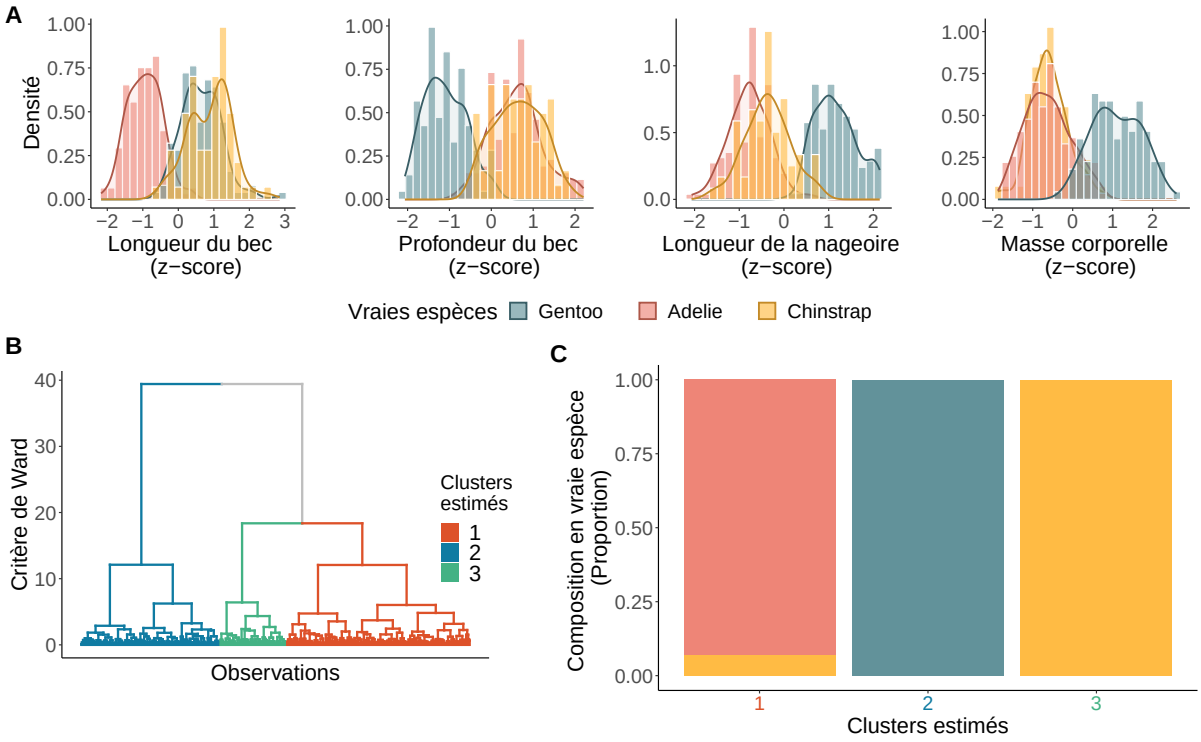


FIGURE 4.4 – **Clustering des 333 manchots de Palmer.** Panneau A : Distribution de chaque variable (z-scores) au sein des trois véritables espèces de manchots. Panneau B : 3 clusters sont estimés grâce à la classification ascendante hiérarchique. Les couleurs dans le dendrogramme représentent les clusters estimés. Panneau C : Concordance entre les résultats du clustering et les espèces connues des manchots.

TABLE 4.2 – **p-valeurs pour tous les tests concernant chaque paire de clusters le long de chacune des 4 variables à partir des données réelles du contrôle positif**.
* met en évidence les p-valeurs significatives au seuil $\alpha = 5\%$

Paire de clusters testée Variable testée	**Test d'inférence sélective direct**	**Test d'inférence sélective par agrégation**	**Test de multimodalité**	**Test t**
Cluster 1 vs Cluster 2 ***(Adélie vs Gentoo)***				
Longueur du bec	0,0024*	0,0023*	0,1647	0*
Profondeur du bec	0,0015*	0,0017*	0,3687	0*
Longueur de la nageoire	0,0725	0,1832	0,0047*	0*
Masse corporelle	0,0439*	0,0008*	0,6402	0*
Cluster 1 vs Cluster 3 ***(Adélie vs Chinstrap)***				
Longueur du bec	0,1748	0,0191*	0,0674	0*
Profondeur du bec	0,2266	0,2323	0,2373	0,0702
Longueur de la nageoire	0,4318	0,4434	0,0168*	0*
Masse corporelle	0,7036	0,7027	0,3311	0,0267*
Cluster 2 vs Cluster 3 ***(Gentoo vs Chinstrap)***				
Longueur du bec	0,2263	0,2115	0,0927	0*
Profondeur du bec	0,0084*	0,0051*	0,2245	0*
Longueur de la nageoire	0,0186*	0,0205*	0,1585	0*
Masse corporelle	0,0002*	0,0002*	0,4174	0*

4.3.3 Clustering de cellules immunitaires issues de mesures de cytométrie en flux

Nous avons également analysé les données du panel de cellules T de Finak et al. (2016) pour le *Human Immunology Project Consortium* (HIPC). Ce jeu de données, disponible publiquement sur *ImmuneSpace*, contient des mesures de cytométrie en flux de 6 marqueurs de surface (CCR7, CD4, CD45RA, HLADR, CD38 et CD8) pour 33 992 cellules T. Dix sous-populations de cellules T ont été identifiées et annotées manuellement par des immunologistes en se basant sur la présence ou l'absence de chacun de ces marqueurs.

Nous avons étudié la capacité des trois tests d'inférence post-clustering présentés à identifier les marqueurs cellulaires spécifiques connus, après un clustering des cellules basé sur ces mêmes marqueurs. Étant donné la taille des données et la charge computationnelle de certains des tests proposés, nous avons limité notre analyse à un sous-ensemble de 5% des cellules annotées provenant uniquement de 4 des 10 sous-populations cellulaires : les CD8 Naïfs, les CD8 Mémoires Effectrices, les CD4 Naïfs et les CD4 Mémoires Effectrices. Au total, nous avons analysé 1 051 cellules (voir Figure A.7 en Annexe A pour une représentation graphique des données). Nous avons utilisé la classification ascendante hiérarchique de Ward sur les z-scores pour estimer $K = 4$ clusters, correspondant au nombre réel de sous-populations cellulaires. Ce clustering a révélé de bonnes performances pour identifier les sous-populations cellulaires avec un ARI de 0.98. Ensuite, nous avons testé chaque marqueur comme séparateur potentiel de chaque paire de clusters estimés, en utilisant les trois tests proposés. Les p-valeurs issues de la comparaison du Cluster 2 (contenant 90% des cellules CD8 Mémoires Effectrices) et du Cluster 4 (contenant 99% des cellules CD4 Mémoires Effectrices) sont données dans la Table 4.3 (toutes les autres comparaisons peuvent être trouvées dans la Table A.3 en Annexe A).

Les deux tests d'inférence sélective ont identifié la plupart des marqueurs comme séparant significativement les paires de clusters. Cela illustre l'une des principales limitations de ces tests d'inférence sélective univariée en présence de variables corrélées. Effectivement, comme presque n'importe quel couple de marqueurs est suffisant pour discriminer les quatre sous-populations cellulaires, la perturbation d'un marqueur permet toujours de retrouver la structure de clustering originale des données (en se basant sur l'information apportée par les marqueurs restants) et conduit donc presque toujours à des résultats significatifs. Ce phénomène est particulièrement prononcé pour le marqueur CD4, estimé comme séparant significativement les lymphocytes T CD4 Naïfs des lymphocytes T CD4 Mémoire Effectrice, bien qu'il soit censé être exprimé dans les deux populations cellulaires. Le test de multimodalité, qui est basé uniquement sur la séparation entre les clusters, n'utilise pas l'information liée aux corrélations, et renvoie alors des résultats biologiquement plus intéressants (Figure 4.5). Par exemple, le marqueur CD4 sépare de manière significative le Cluster 1 et le Cluster 2 du Cluster 3 et du Cluster 4, mais pas

le Cluster 3 du Cluster 4. Il était très facile de conclure que le Cluster 3 et le Cluster 4 contiennent tous deux des lymphocytes T CD4, tandis que le Cluster 1 et le Cluster 2 contiennent tous deux des lymphocytes T CD8. Il est plus difficile pour le test de multimodalité de distinguer les cellules Naïves des cellules Mémoires Effectrices au sein de ces deux populations. Cela peut s'expliquer par le fait que CCR7 et CD45RA ne sont pas les seuls marqueurs canoniques généralement utilisés pour différencier ces sous-types cellulaires. De plus, aucun marqueur spécifique des lymphocytes T Mémoire Effectrice n'a été identifié jusqu'à présent (Saxena et al. 2019).

TABLE 4.3 – **Comparaison entre le Cluster 2 (90% de lymphocytes T CD8 Mémoire Effectrice) et le Cluster 4 (99% de lymphocytes T CD4 Mémoire Effectrice) estimés à partir des données HIPC.**
$*$ met en évidence les p-valeurs significatives au seuil $\alpha = 5\%$

	Test d'inférence sélective direct	**Test d'inférence sélective par agrégation**	**Test de multimodalité**
CCR7	0,0005*	0,0005*	0,8611
CD4	0,0005*	0,0005*	0,0000*
CD45RA	0,0005*	0,0005*	0,9973
HLADR	0,2042	0,2231	0,9960
CD38	0,0013*	0,0013*	0,7312
CD8	0,0005*	0,0005*	0,0000*

4.4 Discussion

Dans ce Chapitre, nous proposons trois nouveaux tests statistiques pour l'inférence post-clustering pouvant être utilisés pour identifier des variables séparant une paire de clusters estimés à partir des données. Ces tests, qui prennent en compte l'impact du clustering sur l'inférence, donnent des p-valeurs valides, n'identifiant pas une séparation induite par l'algorithme de clustering, mais émanant d'un réel processus sous-jacent à la génération de données, contrôlant alors l'erreur de type I sélective. Nos approches peuvent être utilisées indépendamment de l'algorithme de clustering choisi et peuvent facilement être incorporées dans de nombreux pipelines d'analyse de données où les résultats de clustering sont utilisés post-hoc pour décrire et interpréter les clusters.

Le test de multimodalité se distingue comme une approche complètement non paramétrique, offrant une polyvalence sur un large éventail de distributions de données continues. En revanche, le test d'inférence sélective fait une hypothèse gaussienne pour la distribution des données. Bien qu'il puisse être robuste à une certaine asymétrie des données, un écart excessif par rapport à la symétrie de l'hypothèse gaussienne génère des valeurs aberrantes qui forment, à elles seules, un cluster supplémentaire (Figure A.8 en Annexe A).

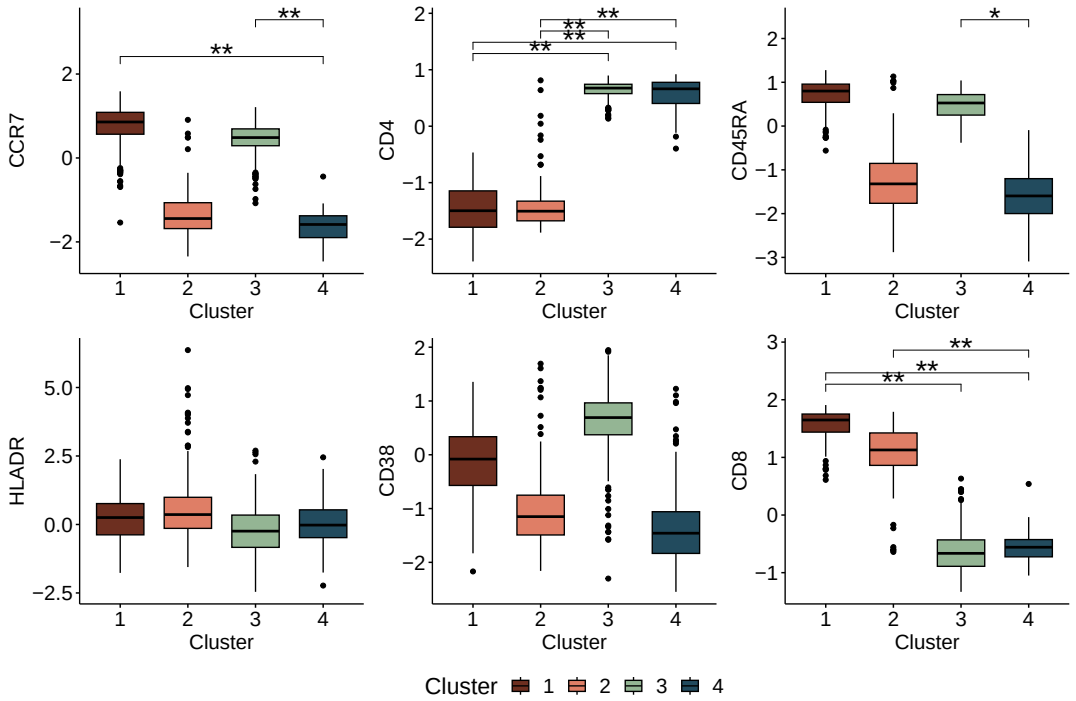


FIGURE 4.5 – **L'inférence post-clustering avec le test de multimodalité appliqué sur les données HIPC permet l'annotation des populations cellulaires.**
Chaque graphique donne la distribution des clusters sur un marqueur de surface. Une séparation significative entre deux clusters au seuil $\alpha = 1\%$ sur un marqueur est mise en évidence par $**$ et une séparation significative entre deux clusters au seuil $\alpha = 5\%$ par $*$.

En pratique, le test de multimodalité peut souvent être préféré, notamment lorsque le signal séparant les clusters est fort, comme le démontrent nos résultats de simulations. Dans ce cas, il identifie les variables qui sont vraiment pertinentes pour l'interprétation des clusters. Il pourrait également être particulièrement attrayant car il est efficace sur le plan computationnel. Cependant, si les données sont très hétérogènes, la multimodalité pourrait devenir un mauvais indicateur de la séparation des clusters lorsque l'hétérogénéité se produit au sein même des clusters, un phénomène souvent observé pour les données scRNA-seq. Dans ce cas, le test de multimodalité n'est pas recommandé et les deux tests d'inférence sélective devraient être préférés — tant que le nombre de variables et leur corrélation ne sont pas trop élevés, sans quoi les tests d'inférence sélective peuvent présenter un faible contrôle du taux de faux positifs. Ils peuvent également être préférés pour les petites tailles d'échantillon, étant plus puissants que le test de multimodalité pour identifier un faible signal.

Tous les codes et les données nécessaires pour reproduire les résultats présentés dans l'article original sont accessibles librement depuis Zenodo avec le `DOI10.5281/zenodo.7660128` (Hivert et al. 2023). Les tests proposés sont implémentés dans un paquet R disponible sur CRAN à l'adresse `https://CRAN.R-project.org/package=VALIDICLUST`.

Depuis la publication de nos méthodes, Chen & Gao (2023) ont également proposé une

extension de la méthode de Gao et al. (2024) pour tester la séparation entre deux clusters au niveau de la variable similaire à notre test d'inférence sélective. Comparé à notre test, leur approche présente l'avantage de ne plus supposer l'absence de corrélations entre les variables, une limite majeure de notre méthode. De plus, ils parviennent à fournir une description explicite de l'ensemble S_g des perturbations des données sur la $g^{\text{ème}}$ variable qui préservent le clustering pour la classification ascendante hiérarchique basée sur les distances euclidiennes au carré, et certaines mesures d'agrégation, ainsi que pour les k-means. Cela permet alors d'accélérer les temps de calculs, la p-valeur pouvant alors être explicitement calculée sans avoir à utiliser une approximation de Monte-Carlo. Cependant, cela rend leur test spécifique à ces méthodes de clustering, contraignant par la même occasion à utiliser la distance euclidienne au carré comme mesure de similarité entre les observations pour la classification ascendante hiérarchique. Enfin, une limite de leur test partagée par le notre est l'hypothèse de variances connues. En effet, en pratique, elles sont inconnues et à estimer avec les données. Néanmoins, ces paramètres sont informatifs sur la présence, ou non, de clusters dans les données, voire même parfois cluster-spécifiques, rendant leur estimation complexe.

Chapitre 5

Limites pratiques des approches basées sur une décomposition de l'information pour répondre aux problèmes d'inférence post-clustering

Ce chapitre est adapté du *preprint* : Hivert, B., Agniel, D., Thiébaut, R., & Hejblum, B. P. (2024). Running in circles : practical limitations for real-life application of data fission and data thinning in post-clustering differential analysis. arXiv preprint, arXiv:2405.13591

Contenu

5.1 Introduction 95
5.2 Méthodes 97
5.2.1 Rappels : fission de données et dilution de données 97
5.2.2 Limites de l'application pratique de la fission et de la dilution de données 99
5.2.3 Connaissance a priori et estimation du paramètre d'échelle 101
5.2.4 Solutions pratiques 103
5.3 Résultats 105
5.3.1 Définition de l'erreur de type I comme une fonction du biais dans l'estimation de la variance 105
5.3.2 Performances de l'estimateur de variance locale 111
5.3.3 Application à l'analyse de données scRNA-seq 113
5.4 Discussion 115

5.1 Introduction

Les méthodes d'inférence post-clustering présentées au Chapitre 4 reposent directement, pour le test d'inférence sélective, ou indirectement, pour le test de multimodalité,

sur un conditionnement lié à l'événement de clustering, garantissant ainsi leur validité. Ces tests d'hypothèses sont spécifiquement conçus pour répondre aux enjeux de l'inférence post-clustering. Toutefois, bien qu'ils soient utilisables pour tester des hypothèses d'expression différentielle entre deux clusters, ils ne sont pas vraiment adaptés à la nature des données RNA-seq. Premièrement, toutes les méthodes d'inférence post-clustering basées sur le concept d'inférence sélective reposent sur une hypothèse gaussienne, une hypothèse qui est discutable dans le contexte des données RNA-seq, comme expliqué en Section 1.1.2 du Chapitre 1. Cela est particulièrement vrai pour nos tests d'inférence sélective qui, en plus de s'appuyer sur l'hypothèse gaussienne, supposent l'indépendance des variables. Cette indépendance n'est jamais vérifiée dans le cas des données RNA-seq, en raison de la forte corrélation entre les gènes, due à leur co-expression dans de mêmes réseaux biologiques. De plus, étant basés sur des perturbations des données, inefficaces en grande dimension, nos tests d'inférence sélective sont difficilement applicables aux données RNA-seq. Bien que non paramétrique, le test de multimodalité n'est pas non plus bien adapté à la nature de ces données. En effet, la forte hétérogénéité qu'on y observe entraîne souvent une multimodalité intra-clusters (causée par exemple par l'inflation en 0), ce qui contredit l'hypothèse fondamentale de ce test.

En raison de ces limitations, nous nous sommes tournés vers des approches basées sur la décomposition de l'information au niveau de l'observation : la fission et la dilution de données, présentées en Section 3.3.2 du Chapitre 3. Ces approches, se concentrant uniquement sur la décomposition de l'information en deux (ou plusieurs) parties indépendantes, peuvent être utilisées conjointement avec n'importe quelle méthode de clustering et d'analyse différentielle. De plus, bien que paramétriques, des décompositions sont connues pour des distributions couramment utilisées pour les données RNA-seq, telles que la distribution de Poisson ou la distribution binomiale négative. Elles offrent alors la possibilité de mieux respecter la nature des données d'expression génique tout en exploitant la forte littérature existante relative à l'analyse de ces données.

La fission et la dilution de données, bien qu'initialement attrayantes pour l'analyse différentielle post-clustering, présentent de fortes limitations. Ces méthodes manquent de justifications théoriques lorsqu'elles sont appliquées à des distributions de mélange, couramment utilisées pour modéliser des données ayant une structure en clusters, comme expliqué en Section 1.2.1.1 du Chapitre 1. L'absence de tels résultats suppose une hypothèse nulle globale d'absence de clusters dans les données, ce qui restreint leur applicabilité. De plus, ces méthodes reposent sur une connaissance préalable de paramètres d'échelles, tels que la variance pour la distribution gaussienne ou la surdispersion pour la distribution binomiale négative. Bien que des estimateurs robustes pourraient théoriquement garantir la validité de ces méthodes, cela ajoute une complexité supplémentaire dans le cadre des modèles de mélanges où chaque composante (chaque cluster) peut avoir des paramètres distincts. Sans connaissance de la structure sous-jacente des données, l'estimateur

basé sur l'échantillon complet, qui utilise toutes les observations indépendamment de la composante du mélange à laquelle elles appartiennent, se révèle inefficace pour estimer les paramètres intra-composante. Même en modélisant chaque observation comme une réalisation de variables aléatoires distinctes ayant chacune leurs propres paramètres, permettant alors de s'affranchir du cadre des modèle de mélange, l'estimation précise des paramètres d'échelle reste difficile sans connaître les composantes du mélange.

Dans ce Chapitre, nous établissons un lien direct entre le biais dans l'estimation de la variance de la distribution gaussienne et l'erreur de type I attendue du test t post-clustering (Welch 1947), soulignant l'importance d'un estimateur robuste de ce paramètre pour la fission de données. Nous employons alors une méthode non paramétrique pour estimer la variance locale dans un contexte gaussien, en utilisant la proximité entre les observations comme un proxy du mélange sous-jacente inconnu. Cependant, les performances de cette approche démontrent qu'une estimation précise de la variance, et donc le contrôle de l'erreur de type I, reste difficile sans connaître le mélange, c'est-à-dire le clustering, en amont de la décomposition.

5.2 Méthodes

5.2.1 Rappels : fission de données et dilution de données

Soit $\boldsymbol{X}$ une variable aléatoire avec une distribution connue. La fission de données (Leiner et al. 2023) et sa généralisation, la dilution de données (Neufeld et al. 2024), visent à décomposer une variable aléatoire $\boldsymbol{X}$ en deux (ou plus, dans le cas de la dilution) nouvelles variables aléatoires $\boldsymbol{X}^{(1)}$ et $\boldsymbol{X}^{(2)}$. Ces nouvelles variables sont conçues pour i) conserver de l'information sur la variable originale $\boldsymbol{X}$, et ii) être indépendantes, dans le cadre de l'inférence post-clustering. La quantité d'information issue de $\boldsymbol{X}$ injectée dans $\boldsymbol{X}^{(1)}$ ou $\boldsymbol{X}^{(2)}$ est alors gérée par un hyperparamètre τ. De telles décompositions sont connues pour différentes distributions de probabilité de $\boldsymbol{X}$ comme la distribution gaussienne, et les distributions de Poisson ou encore binomiale négative, courament utilisées pour modéliser des données RNA-seq. Les décompositions de ces distributions en deux variables aléatoires indépendantes sont détaillées dans la Table 5.1.

Dans le cadre gaussien, soit $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}_{p\times p})$. En considérant la fission de données gaussienne décrite dans la Table 5.1, nous pouvons décomposer $\boldsymbol{X}$ en deux nouvelles variables aléatoires $\boldsymbol{X}^{(1)}$ et $\boldsymbol{X}^{(2)}$ à l'aide d'une nouvelle variable aléatoire $\boldsymbol{Z} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma})$. Il découle de cette décomposition que :

$$\boldsymbol{X}^{(1)} \sim \mathcal{N}\left(\boldsymbol{\mu}, \left(1+\tau^2\right)\boldsymbol{\Sigma}\right) \quad \text{et} \quad \boldsymbol{X}^{(2)} \sim \mathcal{N}\left(\boldsymbol{\mu}, \left(1+\frac{1}{\tau^2}\right)\boldsymbol{\Sigma}\right)$$

Nous souhaitons alors prouver que $\boldsymbol{X}^{(1)}$ et $\boldsymbol{X}^{(2)}$ sont indépendantes.

Distribution de $\boldsymbol{X}$	$\boldsymbol{\tau}$	Fission de données	Dilution de données
$\mathcal{P}(\lambda)$	$\tau \in [0,1]$	$Z \sim \text{Binom}(X,\tau)$ $X^{(1)} = Z$ $X^{(2)} = X - Z$	$X^{(1)} \mid X = x \sim \text{Binom}(x,\tau)$ $X^{(2)} = X - X^{(1)}$
$\mathcal{N}(\boldsymbol{\mu}_p, \boldsymbol{\Sigma}_{p\times p})$	$\tau \in]0,+\infty)$ $\tau_2 \in]0,1[$	$\boldsymbol{Z} \sim \mathcal{N}(\boldsymbol{0}_p, \boldsymbol{\Sigma}_{p\times p})$ $\boldsymbol{X}^{(1)} = \boldsymbol{X} + \tau\boldsymbol{Z}$ $\boldsymbol{X}^{(2)} = \boldsymbol{X} - \frac{1}{\tau}\boldsymbol{Z}$	$\boldsymbol{X}^{(1)} \mid \boldsymbol{X} = \boldsymbol{x} \sim \mathcal{N}(\tau_2\boldsymbol{x}, \tau_2(1-\tau_2)\boldsymbol{\Sigma}_{p\times p})$ $\boldsymbol{X}^{(2)} = \boldsymbol{X} - \boldsymbol{X}^{(1)}$
$\text{NegBin}(\mu,\theta)$	$\tau \in [0,1]$	Pas de fission respectant $\mathcal{P}_1$	$X^{(1)} \mid X = x \sim \text{BetaBin}(x, \tau\theta, (1-\tau)\theta)$ $X^{(2)} = X - X^{(1)}$

TABLE 5.1 – **Fission et dilution de données pour trois distributions usuelles dans le cadre des données d'expression génique : les distributions Poisson, gaussienne et binomiale négative.**

Pour commencer, nous avons :

$$\begin{aligned}
\text{Cov}(\boldsymbol{X}^{(1)}, \boldsymbol{X}^{(2)}) &= \mathbb{E}\left[\left(\boldsymbol{X}^{(1)} - \mathbb{E}[\boldsymbol{X}^{(1)}]\right)\left(\boldsymbol{X}^{(2)} - \mathbb{E}[\boldsymbol{X}^{(2)}]\right)^t\right] \\
&= \mathbb{E}\left[\left(\boldsymbol{X}^{(1)} - \boldsymbol{\mu}\right)\left(\boldsymbol{X}^{(2)} - \boldsymbol{\mu}\right)^t\right] \qquad \text{car} \qquad \mathbb{E}[\boldsymbol{X}^{(1)}] = \mathbb{E}[\boldsymbol{X}^{(2)}] = \boldsymbol{\mu} \\
&= \mathbb{E}\left[\boldsymbol{X}^{(1)}\boldsymbol{X}^{(2)^t} - \boldsymbol{X}^{(1)}\boldsymbol{\mu}^t - \boldsymbol{\mu}\boldsymbol{X}^{(2)^t} + \boldsymbol{\mu}\boldsymbol{\mu}^t\right] \\
&= \mathbb{E}[\boldsymbol{X}^{(1)}\boldsymbol{X}^{(2)^t}] - \mathbb{E}[\boldsymbol{X}^{(1)}]\boldsymbol{\mu}^t - \boldsymbol{\mu}\mathbb{E}[\boldsymbol{X}^{(2)^t}] + \boldsymbol{\mu}\boldsymbol{\mu}^t \\
&= \mathbb{E}[\boldsymbol{X}^{(1)}\boldsymbol{X}^{(2)^t}] - \boldsymbol{\mu}\boldsymbol{\mu}^t - \boldsymbol{\mu}\boldsymbol{\mu}^t + \boldsymbol{\mu}\boldsymbol{\mu}^t \\
&= \mathbb{E}[\boldsymbol{X}^{(1)}\boldsymbol{X}^{(2)^t}] - \boldsymbol{\mu}\boldsymbol{\mu}^t
\end{aligned}$$

Mais :

$$\begin{aligned}
\mathbb{E}[\boldsymbol{X}^{(1)}\boldsymbol{X}^{(2)^t}] &= \mathbb{E}\left[(\boldsymbol{X} + \tau\boldsymbol{Z})\left(\boldsymbol{X} - \frac{1}{\tau}\boldsymbol{Z}\right)^t\right] \\
&= \mathbb{E}\left[\boldsymbol{X}\boldsymbol{X}^t - \frac{1}{\tau}\boldsymbol{X}\boldsymbol{Z}^t + \tau\boldsymbol{Z}\boldsymbol{X}^t - \boldsymbol{Z}\boldsymbol{Z}^t\right] \\
&= \mathbb{E}[\boldsymbol{X}\boldsymbol{X}^t] - \frac{1}{\tau}\mathbb{E}[\boldsymbol{X}\boldsymbol{Z}^t] + \tau\mathbb{E}[\boldsymbol{Z}\boldsymbol{X}^t] - \mathbb{E}[\boldsymbol{Z}\boldsymbol{Z}^t] \\
&= \mathbb{E}[\boldsymbol{X}\boldsymbol{X}^t] - \frac{1}{\tau}\mathbb{E}[\boldsymbol{X}]\mathbb{E}[\boldsymbol{Z}^t] + \tau\mathbb{E}[\boldsymbol{Z}]\mathbb{E}[\boldsymbol{X}^t] - \mathbb{E}[\boldsymbol{Z}\boldsymbol{Z}^t] \qquad \text{car} \qquad \boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Z} \\
&= \mathbb{E}[\boldsymbol{X}\boldsymbol{X}^t] - \mathbb{E}[\boldsymbol{Z}\boldsymbol{Z}^t] \qquad \text{car} \qquad \mathbb{E}[\boldsymbol{Z}] = 0
\end{aligned}$$

De plus, pour calculer $\mathbb{E}[\boldsymbol{X}\boldsymbol{X}^t]$ et $\mathbb{E}[\boldsymbol{Z}\boldsymbol{Z}^t]$, nous pouvons nous appuyer sur leur variance :

$$\begin{aligned}
&\mathbb{V}\mathrm{ar}(\boldsymbol{X}) = \mathbb{E}[\boldsymbol{X}\boldsymbol{X}^t] - \mathbb{E}[\boldsymbol{X}]\mathbb{E}[\boldsymbol{X}^t] = \boldsymbol{\Sigma} \\
&\iff \mathbb{E}[\boldsymbol{X}\boldsymbol{X}^t] = \boldsymbol{\Sigma} + \mathbb{E}[\boldsymbol{X}]\mathbb{E}[\boldsymbol{X}^t] \\
&\iff \mathbb{E}[\boldsymbol{X}\boldsymbol{X}^t] = \boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}^t \quad \text{car} \quad \mathbb{E}[\boldsymbol{X}] = \boldsymbol{\mu}
\end{aligned}$$

et

$$\begin{aligned}
&\mathbb{V}\mathrm{ar}(\boldsymbol{Z}) = \mathbb{E}[\boldsymbol{Z}\boldsymbol{Z}^t] - \mathbb{E}[\boldsymbol{Z}]\mathbb{E}[\boldsymbol{Z}^t] = \boldsymbol{\Sigma} \\
&\iff \mathbb{E}[\boldsymbol{Z}\boldsymbol{Z}^t] = \boldsymbol{\Sigma} + \mathbb{E}[\boldsymbol{Z}]\mathbb{E}[\boldsymbol{Z}^t] \\
&\iff \mathbb{E}[\boldsymbol{Z}\boldsymbol{Z}^t] = \boldsymbol{\Sigma} \quad \text{puisque} \quad \mathbb{E}[\boldsymbol{Z}] = 0
\end{aligned}$$

Ainsi finalement,

$$\mathrm{Cov}(\boldsymbol{X}^{(1)}, \boldsymbol{X}^{(2)}) = \mathbb{E}[\boldsymbol{X}\boldsymbol{X}^t] - \mathbb{E}[\boldsymbol{Z}\boldsymbol{Z}^t] - \boldsymbol{\mu}\boldsymbol{\mu}^t = \boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}^t - \boldsymbol{\Sigma} - \boldsymbol{\mu}\boldsymbol{\mu}^t = 0 \tag{5.1}$$

Les preuves d'indépendance entre $\boldsymbol{X}^{(1)}$ et $\boldsymbol{X}^{(2)}$ pour la distribution de Poisson et la distribution binomiale négative, ont déjà été établies, respectivement, par Neufeld et al. (2024) et Neufeld et al. (2023). Il est clair d'après la Table 5.1 que la fission et la dilution de données gaussiennes et binomiales négatives nécessitent la connaissance des paramètres d'échelle, notamment $\boldsymbol{\Sigma}$ ou θ, pour effectuer la décomposition. Les garanties théoriques de la décomposition, et en particulier l'indépendance entre $\boldsymbol{X}^{(1)}$ et $\boldsymbol{X}^{(2)}$, reposent sur l'utilisation des vraies valeurs de ces paramètres. Cependant, dans les applications réelles, ces valeurs sont inconnues et doivent être estimées à partir des données comme décrit par Neufeld et al. (2023).

5.2.2 Limites de l'application pratique de la fission et de la dilution de données

5.2.2.1 Modèle de mélange

Comme détaillé en Section 1.2.1.1 du Chapitre 1, la présence de clusters dans les données se modélise par des distributions de mélanges, dont la densité est donnée en équation (1.1). Chaque cluster est représenté par une composante du mélange de densité $f(.|\boldsymbol{\theta}_k)$ ayant des paramètres $\boldsymbol{\theta}_k$ qui lui sont propres. Pour simplifier, nous ne considérons que le cas où toutes les composants appartiennent à la même distribution paramétrique de densité f. Cependant, il est impossible d'appliquer la fission de données ou la dilution de données à de telles distributions de mélange. En effet, toutes les distributions pour lesquelles une décomposition est connue sont des distributions simples, qui traduisent

une homogénéité totale des observations et donc l'absence de clusters. Ces distributions simples sont en réalité celles décrivant la distribution des observations au sein d'une composante du mélange. Par conséquent, la fission et la dilution de données ne sont applicables qu'au niveau de la composante individuelle du mélange et non de manière globale.

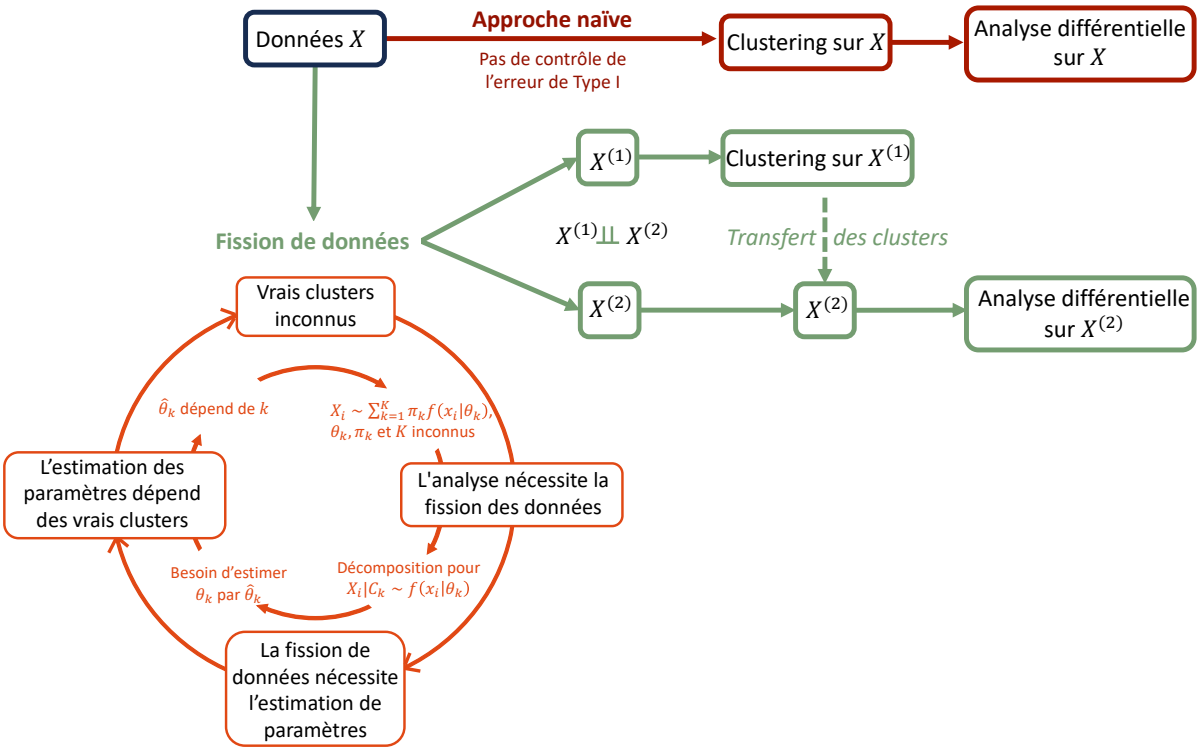


FIGURE 5.1 – **Vue schématique illustrant la circularité induite par la fission de données pour l'analyse différentielle post-clustering.**

La fission de données et la dilution de données créent donc une situation circulaire, comme illustré dans la Figure schématique 5.1. Ces méthodes n'étant applicables qu'à l'échelle d'une composante du mélange, il devient nécessaire de connaître les valeurs des paramètres intra-composante $\boldsymbol{\theta}_k$. Mais, ces paramètres dépendent eux-mêmes des composantes qui sont inconnues puisqu'à estimer à l'aide du clustering. Une approche naïve pourrait consister à estimer un paramètre global $\hat{\boldsymbol{\theta}}$ basé sur toutes les observations. Cependant, cela supposerait que la valeur du paramètre est la même pour toutes les composantes, c'est-à-dire $\boldsymbol{\theta} = \boldsymbol{\theta}_k$, pour tout k. Toutes les observations proviendraient alors d'une même distribution, indépendamment de leur composante :

$$f(\mathbf{x}_i) = f(\mathbf{x}_i|\boldsymbol{\theta}_k) = f(\mathbf{x}_i|\boldsymbol{\theta}). \tag{5.2}$$

Cela ne peut donc être vérifié que pour $K = 1$, c'est-à-dire dans le cas des distributions simples, donc sous une hypothèse nulle globale d'absence de cluster. Yun & Foygel Barber (2023) ont d'ailleurs souligné des défis similaires pour l'estimation de la variance des tests

d'inférence sélective post-clustering existants.

5.2.3 Connaissance a priori et estimation du paramètre d'échelle

Nous restreignons ici notre analyse au cadre gaussien, c'est-à-dire en considérant que :

$$f\left(\boldsymbol{x}_i|\boldsymbol{\theta}_k\right) = f\left(\boldsymbol{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right) = \frac{1}{(2\pi)^{p/2}\left|\Sigma_k\right|^{1/2}} \exp\left(-\frac{1}{2}\left(\boldsymbol{x}_i - \boldsymbol{\mu}_k\right)^t \boldsymbol{\Sigma}_k^{-1}\left(\boldsymbol{x}_i - \boldsymbol{\mu}_k\right)\right)$$

où $k = 1, \ldots, K$ est le nombre de composantes du mélange, $\boldsymbol{\mu}_k \in \mathbb{R}^p$ et $\boldsymbol{\Sigma}_k \in \mathbb{R}^{p\times p}$ sont respectivement le vecteur de moyennes et la matrice de variance de la $k^{\text{ème}}$ composante. Un grand nombre des résultats présentés dans cette partie peuvent cependant être extrapolés à la distribution binomiale négative, en considérant que son paramètre de surdispersion est analogue à la variance dans le cas gaussien.

La Figure 5.2 offre un aperçu détaillé des défis liés à l'estimation de la variance dans la fission de données gaussienne. Le panneau A présente un exemple illustratif avec $n = 300$ réalisations d'une distribution gaussienne multivariée ($K = 1$) avec un vecteur de moyennes $\boldsymbol{\mu} = \left(\begin{smallmatrix}0\\0\end{smallmatrix}\right)$ et une matrice de covariance $\boldsymbol{\Sigma} = \left(\begin{smallmatrix}4 & 0\\0 & 1\end{smallmatrix}\right)$. L'algorithme des k-means a été appliqué sur les n réalisations de $\boldsymbol{X}^{(1)}$ pour construire deux clusters, C_1 et C_2. Les performances de la fission de données ont été évaluées pour différentes estimations de $\boldsymbol{\Sigma}$. Tout d'abord, nous avons considéré la véritable matrice de covariance intra-composante $\boldsymbol{\Sigma}_k$, qui est en fait $\boldsymbol{\Sigma}$ puisque $K = 1$. Nous avons également considéré la matrice de covariance empirique de l'échantillon global, définie par $\widehat{\boldsymbol{\Sigma}} = \frac{1}{n-1}\sum_{i=1}^{n}\left(\boldsymbol{x}_i - \overline{\boldsymbol{x}}\right)\left(\boldsymbol{x}_i - \overline{\boldsymbol{x}}\right)^t$ où $\overline{\boldsymbol{x}}$ est le vecteur des moyennes de l'échantillon. Enfin, nous avons utilisé les résultats des k-means pour calculer une matrice de covariance intra-cluster définie par $\widehat{\boldsymbol{\Sigma}}_{\hat{k}} = \frac{1}{|C_k|-1}\sum_{i\in C_k}\left(\boldsymbol{x}_i - \overline{\boldsymbol{x}}_{C_k}\right)\left(\boldsymbol{x}_i - \overline{\boldsymbol{x}}_{C_k}\right)^t$ où $\overline{\boldsymbol{x}}_{C_k}$ est le vecteur des moyennes calculé à partir des observations appartenant au cluster C_k. Étant donné que les deux clusters proviennent de la même composante, il n'y a pas de différences réelles entre eux. Le test t entre ces deux clusters s'effectue donc sous $\mathcal{H}_0$, et ses p-valeurs devraient alors présenter une distribution uniforme. Le panneau B présente le QQ-plot de la distribution des p-valeurs issues de ce test t par rapport à la distribution uniforme lorsque l'on teste une différence de moyenne sur la variable X_1 à partir des réalisations de $\boldsymbol{X}^{(2)}$ pour 1 000 réplicats de l'expérience. Dans ce scénario où le mélange n'a qu'une seule composante ($K = 1$), la matrice de covariance de l'échantillon $\widehat{\boldsymbol{\Sigma}}$ fournit une estimation non biaisée de la véritable matrice $\boldsymbol{\Sigma}$, ce qui conduit à des p-valeurs uniformément distribuées sur l'intervalle $[0, 1]$. Cependant, en utilisant les matrices de covariance intra-cluster $\widehat{\boldsymbol{\Sigma}}_{\hat{k}}$ (pour $k = 1, 2$) issues des résultats des k-means, les p-valeurs ne présentent plus cettte distribution uniforme. Dans ce cas, les matrices estimées $\widehat{\boldsymbol{\Sigma}}_{\hat{k}}$ pour chaque cluster sous-estiment considérablement la véritable matrice de covariance $\boldsymbol{\Sigma}$, compromettant l'indépendance entre $\boldsymbol{X}^{(1)}$ et $\boldsymbol{X}^{(2)}$. En raison de cette déviation de l'indépendance, il de-

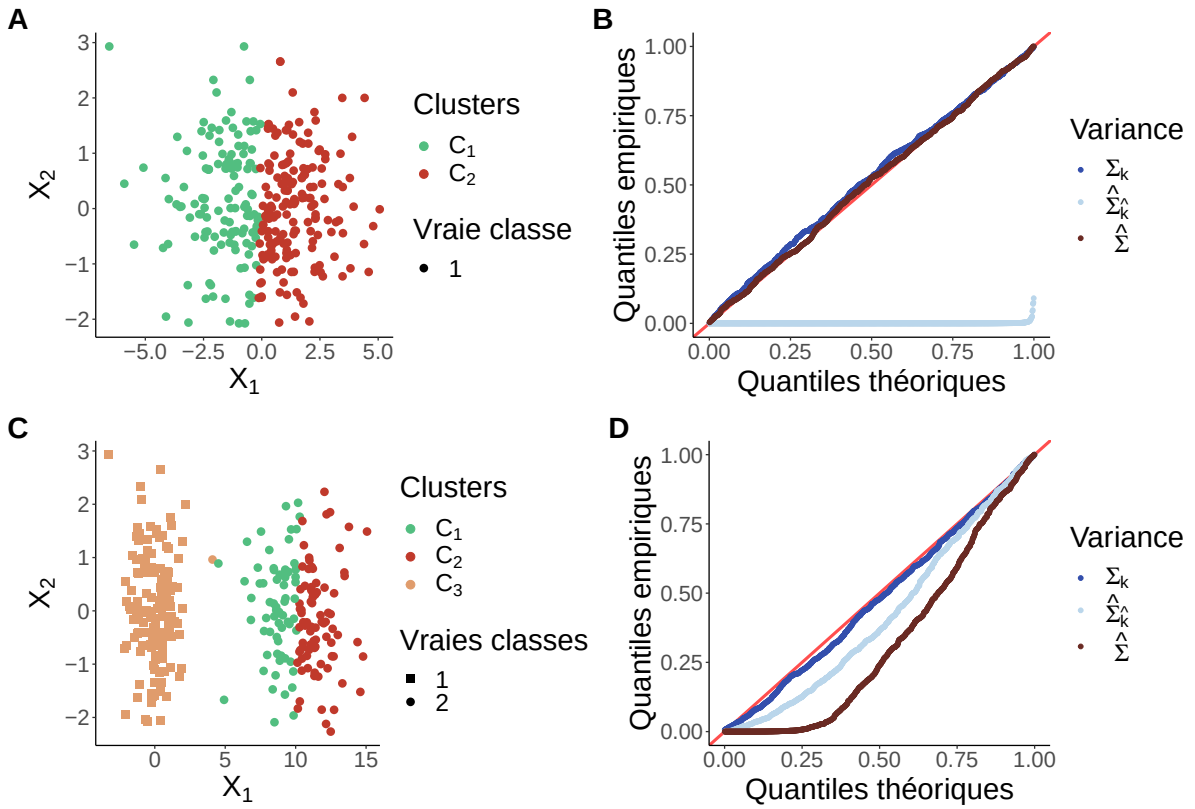


FIGURE 5.2 – **Exemple jouet illustrant l'impact de l'estimation de la variance sur les p-valeurs post-clustering du test t après une fission de données.** Panneau A : Distribution gaussienne bi-dimensionnelle incorrectement découpée en $K = 2$ clusters. Panneau B : QQ-plot des p-valeurs du test t pour la comparaison entre les $K = 2$ clusters estimés à travers 1 000 réplicats lorsque la fission des données est effectuée avec trois estimateurs de variance différents. Panneau C : Extension du problème à un mélange gaussien bidimensionnel à deux composantes incorrectement partagé en $K = 3$ clusters. C_1 et C_2 proviennent de la même composante, qui est divisée à tort en 2 clusters. Panneau D : p-valeurs du test t pour la comparaison entre C_1 et C_2 pour 1 000 réplicats de données utilisant les mêmes 3 estimateurs de variance pour la fission de données.

vient facile de répliquer les clusters identifiés dans $\boldsymbol{X}^{(1)}$ sur $\boldsymbol{X}^{(2)}$, entraînant une inflation directe de l'erreur de type I.

Le panneau C présente un scénario où deux véritables clusters sont générés à l'aide d'un mélange gaussien à deux composantes ($K = 2$). À des fins d'illustration, nous appliquons aux n réalisations de $\boldsymbol{X}^{(1)}$ l'algorithme des k-means pour construire 3 clusters, divisant ainsi incorrectement une composante du mélange en deux clusters, C_1 et C_2. Nous avons de nouveau comparé les performances de la fission de données en utilisant les mêmes 3 estimateurs de covariance que dans le scénario précédent. Le panneau D présente les p-valeurs issues du test t appliqué entre C_1 et C_2 sur X_1 au cours de 1 000 réplicats de l'expérience. Un contrôle efficace de l'erreur de type I est réalisable uniquement en considérant la véritable matrice de covariance intra-composante. Les covariances globales

et intra-clusters estimées avec les résultats des k-means sont des estimateurs biaisés de cette variance intra-composante, entraînant des corrélations entre $\boldsymbol{X}^{(1)}$ et $\boldsymbol{X}^{(2)}$ et compromettant ainsi le contrôle de l'erreur de type I. Cela souligne les défis complexes liés à l'estimation de la covariance pour la fission de données dans des scénarios pratiques, principalement en raison de la mauvaise spécification du modèle génératif de la fission de données lorsque de réels clusters existent.

5.2.4 Solutions pratiques

5.2.4.1 Fission ou dilution de données individuelles pour les modèles de mélange

Dans les modèles de mélange gaussien, les paramètres sont généralement spécifiques aux composantes, ce qui signifie qu'ils sont supposés être partagés par tous les individus au sein d'une même composante. Comme expliqué ci-dessus et illustré dans la Figure 5.1, cette hypothèse crée un défi circulaire. En effet, elle nécessite une connaissance préalable de la structure sous-jacente, encore inconnue, des données pour estimer avec précision les matrices de covariance spécifiques aux composantes, elles-mêmes indispensables à la réalisation de la fission ou de la dilution de données. Pour surmonter cette limite, nous proposons une approche alternative dans laquelle chaque observation est modélisée comme une réalisation de sa propre distribution gaussienne, c'est-à-dire $f_i(\boldsymbol{x}_i) = f\left(\boldsymbol{x_i}|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\right)$. Grâce à cette modélisation qui rompt avec l'hypothèse de réalisations $i.i.d$, la matrice de covariance $\boldsymbol{\Sigma}_i$ n'est plus spécifique aux composantes mais aux observations elles-mêmes. Malgré cette définition au niveau individuel, deux individus issus d'une même composante doivent avoir des paramètres de variance très proches. Cette stratégie de modélisation englobe alors théoriquement à la fois le cas nul global d'absence de clusters ($K = 1$) et le cas de mélange ($K \geq 1$), et ouvre la voie à la fission et à la dilution de données au niveau individuel. Comme souligné ci-dessus, l'estimation de la variance reste cruciale pour la mise en oeuvre pratique de ces méthodes. Cette nouvelle modélisation suppose des variances individuelles qui doivent encore être connues (idéalement), ou estimées avec précision dans des contextes réels.

5.2.4.2 Estimation non paramétrique de la variance locale

Pour estimer $\boldsymbol{\Sigma}_i$, nous proposons d'utiliser des variances pondérées où les poids sont déterminés par des estimations non paramétriques, à l'aide d'un lissage par noyau. Le principe sous-jacent à cette approche est que, malgré la nature spécifique de la variance à chaque individu, deux observations au sein de la même composante du mélange, c'est-à-dire deux voisins, doivent présenter des schémas de variances similaires. Les poids non paramétriques attribués à chaque individu reflètent alors leur contribution à l'estimation

de la variance pour l'observation i, capturant ainsi efficacement la proximité entre les individus. Supposons d'abord que nous sommes dans un cadre univarié, c'est-à-dire que $X_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$ pour $i = 1, \ldots, n$. Nous définissons $\widehat{\sigma^2}_i$, l'estimation de la variance σ_i^2 associée à l'individu x_i, comme suit :

$$\widehat{\sigma^2}_i = \frac{\sum_{j=1}^{n} w_{ij} \left(x_i - \widehat{m}_i\right)^2}{\left(\sum_{j=1}^{n} w_{ij}\right) - 1} \tag{5.3}$$

où w_{ij} sont les poids individuels et $\widehat{m}_i = \sum_{j=1}^{n} w_{ij} x_j \Big/ \sum_{j=1}^{n} w_{ij}$ sont les moyennes pondérées spécifiques à chaque individu. Idéalement, w_{ij} doit être nul, ou très petit, pour toutes les observations x_j qui ne sont pas dans la même composante que x_i.

L'enjeu du calcul de la variance définie en (5.3) réside dans la détermination de chaque poids individuel w_{ij}. Étant donné qu'il est essentiel que les w_{ij} capturent adéquatement la proximité entre les observations, nous avons opté pour une définition de ces poids basée sur un noyau de sorte que $w_{ij} = K(x_i - x_j)$, où $K(u) = \frac{1}{\sqrt{2\pi}h} e^{-\frac{u^2}{2h^2}}$ est le noyau gaussien permettant de définir une mesure de proximité entre x_i et x_j. Cette définition garantit que les poids individuels reflètent les relations locales au sein des données. Ce choix de noyau se concentre sur les points voisins, en mettant en avant leur influence sur l'estimation de la variance pondérée. Le paramètre h dans la définition de K sert de paramètre de lissage, contrôlant la largeur du noyau et, par conséquent, le voisinage autour de chaque observation x_i qui contribue à son estimation pondérée de la variance. Une petite valeur de h conduit à des estimations plus localisées, mettant l'accent sur les points les plus proches, mais peut conduire à une sous-estimation de la variance. À l'inverse, de grandes valeurs de h permettent d'inclure un éventail plus large d'observations, mais une valeur excessivement grande peut conduire à considérer presque toutes les observations dans l'estimation de la variance, produisant un estimateur proche de celui de l'échantillon global. Par conséquent, un choix optimal de h implique de ne considérer que les observations issues d'une même composante du mélange sous-jacent. Cependant, atteindre ce scénario idéal est impraticable, car cela équivaut à connaître les véritables composantes du mélange.

La calibration de ce paramètre de lissage est en faite une étape cruciale de toutes les méthodes à noyau (Heidenreich et al. 2013). Nous proposons d'utiliser un paramètre de lissage spécifique à chaque individu $h_i = h(x_i)$ pour réduire le biais dans l'estimation. Pour ce faire, nous estimons d'abord le point de changement, noté i^*, dans la répartition des distances de l'observation x_i par rapport à toutes les autres observations. Ce point de changement délimite en faite le changement de composante du mélange : les observations avec des distances précédant ce point sont considérées comme faisant partie de la même

composante que x_i, tandis que celles qui le suivent sont considérées comme faisant partie de composantes différentes. Ensuite, nous définissons $h_i = |x_i - x_{i^*}|$ comme la distance entre x_i et l'observation x_{i^*}, qui est la plus éloignée de x_i avant le changement de composante. Ainsi, h_i est déterminé de manière à ce que le lissage individuel ne comprenne que les observations de la même composante que x_i (Chacón & Duong 2020).

5.3 Résultats

5.3.1 Définition de l'erreur de type I comme une fonction du biais dans l'estimation de la variance

L'indépendance entre $\boldsymbol{X}^{(1)}$ et $\boldsymbol{X}^{(2)}$ est garantie uniquement lorsque la véritable variance est utilisée pour la décomposition. Remplacer cette véritable variance $\boldsymbol{\Sigma}_i$ par une estimation $\widehat{\boldsymbol{\Sigma}}_i$ introduit des corrélations entre les nouvelles variables aléatoires qui sont construites. Considérons une variable aléatoire gaussienne $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}_p, \boldsymbol{\Sigma}_{p\times p})$ que nous visons à décomposer en utilisant les processus de fission de données ou de la dilution de données décrits dans la Table 5.1, mais en utilisant une matrice de covariance estimée, $\widehat{\boldsymbol{\Sigma}}$, à la place de $\boldsymbol{\Sigma}$. En effet, supposons que nous utilisions une variable aléatoire $\boldsymbol{Z} \sim \mathcal{N}\left(\boldsymbol{0}, \widehat{\boldsymbol{\Sigma}}\right)$ pour effectuer la fission de données d'une variable aléatoire $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Il découle de l'équation (5.1) que, comme $\mathbb{V}\mathrm{ar}(\boldsymbol{Z}) = \hat{\boldsymbol{\Sigma}}$, nous avons :

$$\mathrm{Cov}(\boldsymbol{X}^{(1)}, \boldsymbol{X}^{(2)}) = \boldsymbol{\Sigma} - \hat{\boldsymbol{\Sigma}}$$

La Table 5.2 résume les valeurs de covariance entre $\boldsymbol{X}^{(1)}$ et $\boldsymbol{X}^{(2)}$ en fonction du biais de $\widehat{\boldsymbol{\Sigma}}$ (pour la preuve de la dilution de données binomiale négative, se référer à Neufeld et al. (2024)). Si $\boldsymbol{X}^{(1)}$ et $\boldsymbol{X}^{(2)}$ ne sont plus indépendantes, c'est-à-dire lorsque $\widehat{\boldsymbol{\Sigma}}$ est un estimateur biaisé de $\boldsymbol{\Sigma}$, un découpage d'une seule composante en deux clusters sur les réalisations de $\boldsymbol{X}^{(1)}$ peut facilement être transféré sur $\boldsymbol{X}^{(2)}$ et entraîne donc une inflation de l'erreur de type I lors de l'étape d'inférence, même si cette dernière est réalisée sur les réalisations de $\boldsymbol{X}^{(2)}$.

	Fission de données	**Dilution de données**
$\mathrm{Cov}\left(\boldsymbol{X}^{(1)}, \boldsymbol{X}^{(2)}\right)$	$\boldsymbol{\Sigma} - \widehat{\boldsymbol{\Sigma}}$	$\tau(1-\tau)\left(\boldsymbol{\Sigma} - \widehat{\boldsymbol{\Sigma}}\right)$

TABLE 5.2 – **Le biais dans l'estimation de $\boldsymbol{\Sigma}$ induit des corrélations entre $\boldsymbol{X}^{(1)}$ et $\boldsymbol{X}^{(2)}$.**

Pour décrire davantage les répercussions de l'estimation de la variance dans le contexte de la fission de données, nous avons dérivé l'expression analytique de l'erreur de type I du test t en fonction du biais dans l'estimation de ce paramètre. Soient $X_1, \ldots, X_n$ n variables aléatoires indépendantes et identiquement distribuées telles que, pour tout $i = 1, \ldots, n$,

$X_i \overset{i.i.d}{\sim} \mathcal{N}(\mu, \sigma^2)$. n joue ici le rôle de la taille d'échantillon. Soient $Z_1, \ldots, Z_n \overset{i.i.d}{\sim} \mathcal{N}(0, b^2)$ et $\tau \in]0, +\infty[$. Ici b^2 représente n'importe quelle valeur utilisée en lieu et place de la variance de X, et en particulier b^2 peut-être une estimation obtenue pour σ^2. Pour tout $i = 1, \ldots, n$, le processus de fission de X_i est :

$$X_i^{(1)} = X_i + \tau Z_i \qquad \text{et} \qquad X_i^{(2)} = X_i - \frac{1}{\tau} Z_i$$

On en déduit immédiatement les distributions marginales de $X_i^{(1)}$ et $X_i^{(2)}$ pour tout $i = 1, \ldots, n$, grâce à l'indépendance entre X_i et Z_i :

$$X_i^{(1)} \sim \mathcal{N}\left(\mu, \sigma^2 + \tau^2 b^2\right) \qquad \text{et} \qquad X_i^{(2)} \sim \mathcal{N}\left(\mu, \sigma^2 + \frac{1}{\tau^2} b^2\right) \tag{5.4}$$

On note $\sigma^2_{X^{(1)}} = \sigma^2 + \tau^2 b^2$ et $\sigma^2_{X^{(2)}} = \sigma^2 + \frac{1}{\tau^2} b^2$ les variances respectives de $X_i^{(1)}$ et $X_i^{(2)}$ ci-dessus.

Dans le contexte de la fission de données pour répondre aux défis de l'inférence post-clustering, un algorithme de clustering est appliqué aux observations de $\boldsymbol{X}^{(1)}$. Sans perte de généralité, on suppose que l'algorithme de clustering appliqué aux réalisations sépare $X_1^{(1)}, \ldots, X_n^{(1)}$ en deux clusters C_1 et C_2 autour de μ (ce qui est le cas en pratique avec l'algorithme des k-means ou avec un modèle de mélange gaussien à 2 composantes de variance homogène lorsque n est suffisament grand). On suppose également que ces deux clusters sont de même effectif et de même variance. Ainsi, les clusters C_1 et C_2 peuvent être exprimés comme :

$$C_1 = \left\{i = 1, \ldots, n : X_i^{(1)} > \mu\right\} \qquad \text{et} \qquad C_2 = \left\{i = 1, \ldots, n : X_i^{(1)} \leq \mu\right\}$$

On peut alors expliciter les distributions conditionnelles de $X_i^{(1)}|C_1$ et $X_i^{(1)}|C_2$. En effet, $\mathbb{P}(X_i^{(1)} = x|C_1) = \mathbb{P}(X_i^{(1)} = x|X_i^{(1)} > \mu)$ avec $X_i^{(1)} \sim \mathcal{N}\left(\mu, \sigma^2_{X^{(1)}}\right)$. $X_i^{(1)}|X_i^{(1)} > \mu$ suit donc une distribution demie-gaussienne, et pour tout $i = 1, \ldots, n$:

$$\mathbb{E}\left[X_i^{(1)}|X_i^{(1)} > \mu\right] = \mu + \sqrt{\frac{2\sigma^2_{X^{(1)}}}{\pi}} \quad \text{et} \quad \text{Var}\left(X_i^{(1)}|X_i^{(1)} > \mu\right) = \left(1 - \frac{2}{\pi}\right)\sigma^2_{X^{(1)}}$$

Comme le cluster $C_2 = \left\{i = 1, \ldots, n : X_i^{(1)} \leq \mu\right\}$ représente simplement le cluster de l'autre côté de la moyenne μ, on a de la même façon :

$$\mathbb{E}\left[X_i^{(1)}|X_i^{(1)} \leq \mu\right] = \mu - \sqrt{\frac{2\sigma^2_{X^{(1)}}}{\pi}} \quad \text{et} \quad \text{Var}\left(X_i^{(1)}|X_i^{(1)} \leq \mu\right) = \left(1 - \frac{2}{\pi}\right)\sigma^2_{X^{(1)}}$$

Dans le cadre de l'inférence post-clustering, les tests d'hypothèses sont effectués sur

l'autre partie de l'information, contenue en l'occurrence dans $\boldsymbol{X}^{(2)}$. Nous nous intéressons au test t pour deux échantillons afin de tester une potentielle différence de moyenne sur $\boldsymbol{X}^{(2)}$ selon les groupes définis par les deux clusters $C_1 = \left\{i = 1, \dots, n : X_i^{(1)} > \mu\right\}$ et $C_2 = \left\{i = 1, \dots, n : X_i^{(1)} \leq \mu\right\}$. On s'intéresse donc aux hypothèses suivantes :

$$\mathcal{H}_0 : \mu_{C_1} = \mu_{C_2} \qquad \text{vs} \qquad \mathcal{H}_1 : \mu_{C_1} \neq \mu_{C_2}$$

où $\mu_{C_1} = \mathbb{E}\left[X_i^{(2)} | X_i^{(1)} > \mu\right]$ et $\mu_{C_2} = \mathbb{E}\left[X_i^{(2)} | X_i^{(1)} \leq \mu\right]$ sont respectivement la moyenne des $X_i^{(2)}$ dans le cluster C_1 et dans le cluster C_2.

Puisque nous avons supposé que les deux clusters obtenus étaient de variances égales (et de même effectif $n/2$), on note $s^2(X^{(2)}) = \text{Var}\left(X_i^{(2)} | X_i^{(1)} > \mu\right) = \text{Var}\left(X_i^{(2)} | X_i^{(1)} \leq \mu\right)$ cette variance commune. La statistique de test correspondante s'écrit alors :

$$T = \frac{\overline{X_{C_1}^{(2)}} - \overline{X_{C_2}^{(2)}}}{\sqrt{\frac{4s^2\left(X^{(2)}\right)}{n}}} \quad \text{où} \quad \overline{X_{C_k}^{(2)}} = \frac{1}{n/2} \sum_{i \in C_k} X_i^{(2)} \quad \text{pour} \quad k = 1, 2$$

Bien que chaque $X_i^{(2)}$ soit gaussien, ce n'est plus le cas conditionnellement aux clusters, c'est-à-dire à $X_i^{(1)} > \mu$ pour C_1 (ou à $X_i^{(1)} \leq \mu$ pour C_2). Cependant, lorsque n est suffisamment grand, nous pouvons appliquer le théorème central limite qui nous donne :

$$\overline{X_{C_1}^{(2)}} - \overline{X_{C_2}^{(2)}} \overset{\mathcal{L}}{\sim} \mathcal{N}\left(\mu_{C_1} - \mu_{C_2}, \frac{4}{n} s^2\left(X^{(2)}\right)\right)$$

La distribution asymptotique de notre statistique de test T est donc :

$$T = \frac{\overline{X_{C_1}^{(2)}} - \overline{X_{C_2}^{(2)}}}{\sqrt{\frac{4s^2\left(X^{(2)}\right)}{n}}} \overset{\mathcal{L}}{\sim} \mathcal{N}\left(\frac{\mu_{C_1} - \mu_{C_2}}{\sqrt{\frac{4s^2\left(X^{(2)}\right)}{n}}}, 1\right) \tag{5.5}$$

Cette distribution asymptotique dépend donc de trois quantités : μ_{C_1}, μ_{C_2} et $s^2(X^{(2)})$, que l'on peut calculer. Grâce au théorème de l'espérance totale, on remarque que :

$$\mu_{C_1} = \mathbb{E}\left[X_i^{(2)} | X_i^{(1)} > \mu\right] = \mathbb{E}\left[\mathbb{E}\left[X_i^{(2)} | X_i^{(1)}\right] | X_i^{(1)} > \mu\right] \tag{5.6}$$

Comme $X_i^{(1)}$ et $X_i^{(2)}$ sont deux variables aléatoires gaussiennes, alors, pour tout $i = 1, \dots, n$, on a le vecteur gaussien suivant :

$$\begin{pmatrix} X_i^{(1)} \\ X_i^{(2)} \end{pmatrix} \sim \mathcal{N}_2\left(\begin{pmatrix} \mu \\ \mu \end{pmatrix}, \begin{pmatrix} \sigma^2_{X^{(1)}} & \rho\sigma_{X^{(1)}}\sigma_{X^{(2)}} \\ \rho\sigma_{X^{(1)}}\sigma_{X^{(2)}} & \sigma^2_{X^{(2)}} \end{pmatrix}\right)$$

où $\rho = \text{Cor}\left(X^{(1)}, X^{(2)}\right)$. Grâce aux propriétés des vecteurs gaussiens, on peut en déduire

la distribution de $X_i^{(2)}|X_i^{(1)}$ qui est, pour tout $i = 1, \ldots, n$:

$$X_i^{(2)}|X_i^{(1)} \sim \mathcal{N}\left(\mu + \frac{\rho\sigma_{X^{(1)}}\sigma_{X^{(2)}}}{\sigma^2_{X^{(1)}}}\left(X_i^{(1)} - \mu\right), \sigma^2_{X^{(2)}} - \frac{\rho^2\sigma^2_{X^{(1)}}\sigma^2_{X^{(2)}}}{\sigma^2_{X^{(1)}}}\right) \tag{5.7}$$

qui se simplifie en :

$$X_i^{(2)}|X_i^{(1)} \sim \mathcal{N}\left(\mu + \rho\frac{\sigma_{X^{(2)}}}{\sigma_{X^{(1)}}}\left(X_i^{(1)} - \mu\right), \sigma^2_{X^{(2)}}\left(1 - \rho^2\right)\right) \tag{5.8}$$

En injectant l'ésperance de $X_i^{(2)}|X_i^{(1)}$ ci-dessus dans l'équation (5.6), on obtient :

$$\begin{aligned}
\mu_{C_1} &= \mathbb{E}\left[\mu + \rho\frac{\sigma_{X^{(2)}}}{\sigma_{X^{(1)}}}\left(X_i^{(1)} - \mu\right)|X_i^{(1)} > \mu\right] \\
&= \mu + \rho\frac{\sigma_{X^{(2)}}}{\sigma_{X^{(1)}}}\left(\mathbb{E}\left[X_i^{(1)}|X_i^{(1)} > \mu\right] - \mu\right) \\
&= \mu + \rho\frac{\sigma_{X^{(2)}}}{\sigma_{X^{(1)}}}\left(\mu + \sqrt{\frac{2\sigma^2_{X^{(1)}}}{\pi}} - \mu\right) \\
&= \mu + \rho\sqrt{\frac{2}{\pi}\sigma^2_{X^{(2)}}}
\end{aligned}$$

Par un raisonnement identiqueme on montre que :

$$\mu_{C_2} = \mathbb{E}\left[X_i^{(2)}|X_i^{(1)} \leq \mu\right] = \mu - \rho\sqrt{\frac{2}{\pi}\sigma^2_{X^{(2)}}}$$

Enfin, en utilisant maintenant le théorème de la variance totale, on a :

$$\mathrm{Var}\left(X_i^{(2)}|X_i^{(1)} > \mu\right) = \mathbb{E}\left[\mathrm{Var}\left(X_i^{(2)}|X_i^{(1)}\right)|X_i^{(1)} > \mu\right] + \mathrm{Var}\left(\mathbb{E}\left[X_i^{(2)}|X_i^{(1)}\right]|X_i^{(1)} > \mu\right)$$

Or, toujours grâce à l'équation (5.8), on remarque d'abord que :

$$\begin{aligned}
\mathbb{E}\left[\mathrm{Var}\left(X_i^{(2)}|X_i^{(1)}\right)\middle| X_i^{(1)} > \mu\right] &= \mathbb{E}\left[\sigma^2_{X^{(2)}}\left(1 - \rho^2\right)|X_i^{(1)} > \mu\right] \\
&= \sigma^2_{X^{(2)}}\left(1 - \rho^2\right)
\end{aligned}$$

et ensuite que :

$$\begin{aligned}
\mathrm{Var}\left(\mathbb{E}\left[X_i^{(2)}|X_i^{(1)}\right]|X_i^{(1)} > \mu\right) &= \mathrm{Var}\left(\mu + \rho\frac{\sigma_{X^{(2)}}}{\sigma_{X^{(1)}}}\left(X_i^{(1)} - \mu\right)|X_i^{(1)} > \mu\right) \\
&= \rho^2\frac{\sigma^2_{X^{(2)}}}{\sigma^2_{X^{(1)}}}\mathrm{Var}\left(X_i^{(1)}|X_i^{(1)} > \mu\right) \\
&= \rho^2\frac{\sigma^2_{X^{(2)}}}{\sigma^2_{X^{(1)}}}\left(1 - \frac{2}{\pi}\right)\sigma^2_{X^{(1)}} \\
&= \rho^2\left(1 - \frac{2}{\pi}\right)\sigma^2_{X^{(2)}}
\end{aligned}$$

Donc finalement, on obtient :

$$\begin{aligned}
s^2\left(X^{(2)}\right) &= \mathrm{Var}\left(X_i^{(2)}|X_i^{(1)} > \mu\right) \\
&= \sigma^2_{X^{(2)}}\left(1 - \rho^2\right) + \rho^2\left(1 - \frac{2}{\pi}\right)\sigma^2_{X^{(2)}} \\
&= \sigma^2_{X^{(2)}}\left(1 - \frac{2}{\pi}\rho^2\right)
\end{aligned}$$

Par un raisonnement identiqueme on montre que :

$$\mathrm{Var}\left(X_i^{(2)}|X_i^{(1)} \leq \mu\right) = \sigma^2_{X^{(2)}}\left(1 - \frac{2}{\pi}\rho^2\right),$$

ce qui, au passage, vérifie heureusement notre hypothèse initiale d'égalité des variances intra-clusters. On peut ainsi enfin calculer :

$$\mathbb{E}\left[T\right] = \frac{\mu_{C_1} - \mu_{C_2}}{\sqrt{\frac{4s^2\left(X^{(2)}\right)}{n}}} = \frac{\mu + \rho\sqrt{\frac{2}{\pi}\sigma^2_{X^{(2)}}} - \left(\mu - \rho\sqrt{\frac{2}{\pi}\sigma^2_{X^{(2)}}}\right)}{\sqrt{\frac{4\sigma^2_{X^{(2)}}\left(1-\frac{2}{\pi}\rho^2\right)}{n}}} = \frac{\rho\sqrt{n}\sqrt{\frac{2}{\pi}}}{\sqrt{1 - \frac{2}{\pi}\rho^2}},$$

pour finalement obtenir :

$$T \overset{\mathcal{L}}{\sim} \mathcal{N}\left(\frac{\rho\sqrt{n}}{\sqrt{\frac{\pi}{2} - \rho^2}}, 1\right). \tag{5.9}$$

Rappelons que l'échantillon $X_1, \ldots, X_n$ est normalement distribué : il ne contient aucun véritable cluster, et donc aucune différence de moyennes n'existe réellement entre des sous-groupes d'observations qui ne soit pas la conséquence du clustering. La statistique de test T devrait ainsi être sous $\mathcal{H}_0$, et donc être centrée en 0. Dans notre résultat en (5.9), on observe une déviation de la distribution de T par rapport à 0 quantifiée par $\rho\sqrt{n}/\sqrt{\frac{\pi}{2} - \rho^2}$. L'erreur de type I au niveau α associée à ce test est alors donnée par $1 - F(q_{\alpha/2}) + F(-q_{\alpha/2})$, où F est la fonction de répartition de la loi normale $\mathcal{N}\left(\frac{\rho\sqrt{n}}{\sqrt{\frac{\pi}{2}-\rho^2}}, 1\right)$, et $q_{\alpha/2}$ est le quantile d'ordre $\alpha/2$ de la distribution normale centrée réduite $\mathcal{N}(0, 1)$.

Nous avons considéré ici que toutes les variances étaient connues et que donc $s^2\left(X^{(2)}\right)$ l'était également. Il a donc été possible de calculer la distribution de la statistique de test du test Z pour comparer les moyennes entre deux échantillons. Mais ce résultat s'étend facilement au cas pratique plus envisageable où l'on considère plutôt $s^2\left(X^{(2)}\right)$ inconnue et que l'on utilise alors une estimation $\widehat{s^2}\left(X^{(2)}\right)$. Dans ce cas, en supposant toujours l'égalité des variances entre les deux clusters, la distribution de la statistique de test T est une distribution de Student $\mathcal{T}(n-2)$ (en raison des incertitudes liées à l'estimation de cette variance commune). L'erreur de type I associée devient : $1 - F_{\mathcal{T}}(q_{\alpha/2}) + F_{\mathcal{T}}(-q_{\alpha/2})$, où $F_{\mathcal{T}}$ est la fonction de répartition de la distribution de Student non-centrale de moyenne $\rho\sqrt{n}/\sqrt{\frac{\pi}{2} - \rho^2}$ à $n-2$ degrés de liberté, et $q_{\alpha/2}$ est le quantile d'ordre $\alpha/2$ de la distribution de Student à $n-2$ degrés de liberté.

Ce résultat illustre l'importance cruciale d'une estimation précise de la variance pour l'application de la fission de données. En effet, pour que le test soit valide, c'est-à-dire que l'erreur de type I soit contrôlée au niveau α, il faut que la distribution de la statistique de test soit centrée en 0. Cela implique que :

$$
\begin{aligned}
\rho\sqrt{n} = 0 &\iff \rho = 0 \\
&\iff \operatorname{Cor}\left(X_i^{(1)}, X_i^{(2)}\right) = 0 \qquad \forall i = 1, \ldots, n \\
&\iff \frac{\operatorname{Cov}\left(X_i^{(1)}, X_i^{(2)}\right)}{\sigma_{X^{(1)}}\sigma_{X^{(2)}}} = 0 \qquad \forall i = 1, \ldots, n \\
&\iff \operatorname{Cov}\left(X_i^{(1)}, X_i^{(2)}\right) = 0 \qquad \forall i = 1, \ldots, n \\
&\iff \sigma^2 - b^2 = 0
\end{aligned}
$$

Donc seule une fission de données effectuée avec le vrai paramètre de variance (ou du moins un estimateur sans biais de ce dernier) peut garantir un contrôle effectif de l'erreur de type I.

Nous avons validé ce résultat théorique par des simulations numériques et nous avons également examiné l'influence des valeurs de variance originale ainsi que de la taille de l'échantillon sur l'erreur de type I du test t post-clustering dans le cadre de la fission de données gaussiennes. Nous avons généré n réalisations d'une variable aléatoire gaussienne avec une moyenne $\mu = 0$ et des variances σ^2. Nous avons ensuite considéré une grille de valeurs pour $b^2 = \widehat{\sigma^2}$ qui peuvent représenter différentes estimations de la variance originale σ^2. Nous avons ensuite appliqué la fission de données avec ces valeurs variables de $\widehat{\sigma^2}$, obtenant $X^{(1)}$ pour le clustering par les k-means avec $K = 2$ et $X^{(2)}$ pour tester les différences de moyenne entre les deux clusters estimés par le test t. 1 000 réplicats de ces simulations ont été effectués. Tout d'abord, nous avons examiné l'impact de la vraie variance initiale σ^2 en considérant des valeurs de σ^2 de $\{0.01, 0.25, 1, 4\}$ pour une taille

d'échantillon fixe de $n = 100$. La Figure 5.3A illustre la relation entre le biais dans l'estimation de σ^2 et l'erreur de type I obtenue, démontrant un accord constant avec l'erreur de type I théorique, quelque soit la variance initiale σ^2 ou le biais dans son estimation. Nous avons également étudié le comportement de l'erreur de type I pour une valeur fixe $\sigma^2 = 1$ mais pour des tailles d'échantillon variables $n \in \{50, 100, 200, 500, 1000\}$. La Figure 5.3B montre alors l'impact attendu de la taille de l'échantillon sur l'erreur de type I. Ces résultats soulignent collectivement l'importance critique de l'estimation précise de la variance pour obtenir une bonne calibration de l'erreur de type I.

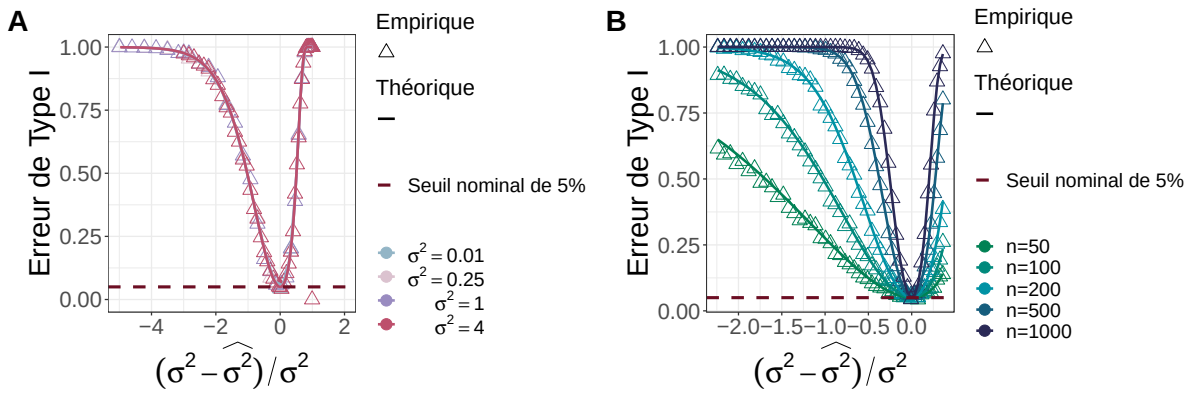


FIGURE 5.3 – **Impact de l'estimation de la variance sur l'erreur de type I dans la fission de données gaussiennes**. Panneau A : Évolution de l'erreur de type I post-fission en fonction du biais relatif et de la variance originale des données. Panneau B : Évolution de l'erreur de type I post-fission en fonction du biais relatif et de la taille de l'échantillon.

5.3.2 Performances de l'estimateur de variance locale

Nous avons également évalué la performance de l'estimateur de variance non paramétrique défini en (5.3) au travers de simulations numériques. Pour ce faire, nous avons repris l'exemple de la Figure 5.2 et nous avons généré $n = 100$ réalisations d'un mélange gaussien univarié à deux composantes : $0.5\mathcal{N}(0, \sigma^2) + 0.5\mathcal{N}(\delta, \sigma^2)$. Nous avons alors exploré une gamme de valeurs du ratio δ/σ allant de 0, représentant l'absence totale de vrais clusters dans les données, à 100, indiquant une séparation extrême entre les vrais clusters. Nous avons également considéré différentes valeurs de $\sigma^2 \in \{0.01, 0.25, 1, 4\}$. Pour chaque paire (δ, σ^2) définissant le mélange et paramétrisant la séparation de ses composantes, nous avons estimé les variances individuelles de chaque observation par notre estimateur non-paramétrique de la variance et nous nous en sommes servis pour appliquer la fission de données à l'échelle individuelle. Nous avons ensuite uniquement considéré les observations provenant de la première composante du mélange pour appliquer les k-means sur les observations de $X^{(1)}$ et construire $K = 2$ clusters. Cela implique qu'une vraie composante est incorrectement partagée en deux clusters. Nous avons ensuite testé une différence de

moyenne entre ces deux clusters en utilisant le test t sur les observations correspondantes de $X^{(2)}$. Ce scénario a été répliqué 1 000 fois, et nous avons calculé l'erreur de type I au niveau $\alpha = 5\%$.

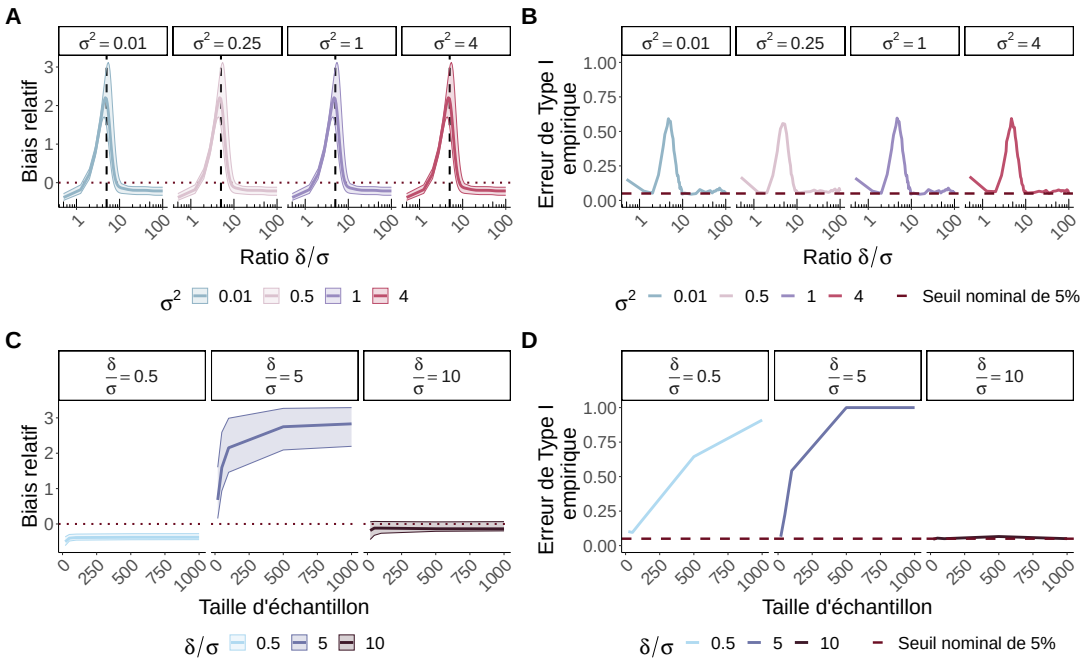


FIGURE 5.4 – **Évaluation des performances de l'estimateur de variance non paramétrique dans un cadre univarié simulé.** Tous les résultats empiriques ont été obtenus à partir de 1000 simulations des données. Panneau A : Biais relatif médian, défini comme $\left(\widehat{\sigma^2} - \sigma^2\right)/\sigma^2$ et sa plage inter-quartile associée, par rapport au ratio signal sur bruit δ/σ. Panneau B : Erreur de type I au niveau $\alpha = 5\%$ par rapport à δ/σ. Panneau C : Biais relatif médian et sa plage inter-quartile associée en fonction de la taille de l'échantillon pour trois degrés de séparation entre les deux composantes (donnés par le rapport δ/σ). Panneau D : Erreur de type I au niveau $\alpha = 5\%$ associée en fonction de la taille de l'échantillon.

La Figure 5.4A montre que les variances locales sont sous-estimées, ce qui se manifeste par un biais relatif négatif, jusqu'à ce que le rapport signal/bruit δ/σ atteigne environ 3, une valeur seuil de séparation dans un modèle de mélange gaussien déjà mentionné par Siffer et al. (2018), Hivert, Agniel, Thiébaut & Hejblum (2024). En effet, avant ce seuil, il n'y a pas de points de changement clairs dans la dispersion des distances entre observations et toutes les observations devraient contribuer à poids égal au calcul de la variance. Cependant, ce n'est pas le cas ici, en raison d'un paramètre de lissage h_i trop petit, ce qui conduit à une sous-estimation des variances locales. À mesure que le ratio augmente dans la plage $3.5 \leq \delta/\sigma \leq 10$, la séparation des composantes devient plus claire mais l'identification des points de changement reste difficile. La détection des points de changement est complexe pour les observations dans les queues entre les composantes, et cela peut conduire à attribuer un poids non-nuls à une observations x_j pourtant issue

d'une composante différente de x_i. En conséquence, une surestimation des variances locales, se traduisant par un biais relatif positif, est observée ainsi qu'une augmentation de leur intervalle inter-quartile associé. Enfin, pour $\delta/\sigma > 10$, la séparation entre les vraies composantes est suffisante pour garantir une bonne estimation des variances locales, ce qui souligne l'importance d'une séparation nette des composantes pour une estimation précise de la variance. Ainsi, la méthodologie décrite dans la Section 5.2.4.2 n'est applicable que lorsque les composantes du mélange sont séparées par un très fort signal. Dans la Figure 5.4B, l'erreur de type I à 5% reste bien calibrée uniquement pour les valeurs de δ/σ qui assurent des estimations de variance robustes (c'est-à-dire lorsque $\delta/\sigma > 10$), ce qui montre à nouveau la nécessité d'une bonne séparation des composantes pour garantir une bonne calibration des tests post-clustering dans le cadre de la fission de données. La Figure 5.4C montre que l'augmentation de la taille de l'échantillon ne parvient pas à améliorer les estimations, pourtant non paramétriques, des variances locales pour trois valeurs particulières du ratio $\delta/\sigma \in \{0.5, 5, 10\}$. Cela indique donc que les performances de l'estimateur de la variance locale dépendent davantage de la séparation des composantes que de la taille de l'échantillon. L'erreur de type I correspondante, illustrée dans la Figure 5.4D, s'aligne donc avec les résultats obtenus en termes de biais relatif, soulignant une fois de plus l'importance d'une estimation précise des variances locales pour assurer un contrôle effectif de l'erreur de type I.

5.3.3 Application à l'analyse de données scRNA-seq

Étant donné leur nature de comptage surdispersée, la distribution binomiale négative est préférable à la distribution gaussienne pour modéliser les données scRNA-seq. Malheureusement, la distribution binomiale négative est soumise aux mêmes problématiques de circularité que la distribution gaussienne pour la dilution de données. Le paramètre de surdispersion θ joue un rôle similaire au paramètre de variance dans la distribution gaussienne, car il est considéré comme connu lors du processus de dilution. Par conséquent, la qualité de son estimation impacte directement $\text{Cov}\left(X^{(1)}, X^{(2)}\right) = \tau(1-\tau)\frac{\mu^2}{\theta}\left(1 - \frac{\theta+1}{\widehat{\theta}+1}\right)$, où $\widehat{\theta}$ est une estimation de θ (Neufeld et al. 2023). De plus, pour les mélanges binomiaux négatifs, la décomposition par dilution de données n'est à nouveau réalisable qu'à l'échelle de la composante. Étant donné que le paramètre de surdispersion est spécifique aux composantes (Li et al. 2018), fournir un estimateur sans biais de ce paramètre est de nouveau très complexe sans la connaissance a priori sur le processus de mélange sous-jacent aux données.

Nous avons illustré l'importance d'appliquer la dilution de données de manière intra-composante (avec la surdispersion intra-composante associée) pour assurer un contrôle effectif de l'erreur de type I dans le cadre de l'inférence post-clustering avec des simulations numériques. Nous avons généré $n = 100$ observations d'un mélange binomial négatif

à deux composantes : $0.5\text{NegBin}(\mu_1, \theta_1) + 0.5\text{NegBin}(\mu_2, \theta_2)$ où les composantes sont paramétrisées par $(\mu_1, \theta_1) = (5, 5)$ et $(\mu_2, \theta_2) = (60, 40)$. Nous avons effectué une inférence post-clustering similaire à celle de la Figure 5.2. Dans la Figure 5.5A, l'application des k-means avec $K = 3$ clusters force la première composante du mélange à être incorrectement divisée en 2 clusters (C_1 et C_3). Nous avons ensuite évalué l'erreur de type I associée au test de Wilcoxon entre ces clusters incorrects lors de l'application de la dilution de données avec divers estimateurs de surdispersion. Tout d'abord, nous avons appliqué la dilution de données intra-composante en utilisant des estimations oracles $\widetilde{\theta}_k$, $k = 1, 2$, représentant les vrais paramètres de surdispersion intra-composante (irréalisable dans une application pratique où la structure réelle des clusters est inconnue). Nous avons comparé ces résultats avec deux alternatives réalisables en pratique : *i*) appliquer la dilution de données intra-cluster basé sur les résultats des k-means de la Figure 5.5A avec leurs $\widehat{\theta}_{\hat{k}}$ associés, $\hat{k} = 1, 2, 3$, et *ii*) la dilution de données globale avec son estimateur global $\widehat{\theta}$ associé. Il est à noter que toutes les estimations de surdispersion ont été réalisées par maximum de vraisemblance. La Figure 5.5B présente le QQ-plot par rapport à la distribution Uniforme des p-valeurs associées au test de Wilcoxon sur 1 000 réplications de l'expérience. De même que dans le cadre gaussien, obtenir des p-valeurs uniformément distribuées n'est possible qu'à travers la dilution de données intra-composante utilisant l'estimateur oracle de surdispersion $\widetilde{\theta}_k$ associé à chaque composante. Toutes les autres approches de dilution de données sont effectuées avec des estimateurs biaisés de la surdispersion, compromettant l'indépendance entre $X^{(1)}$ et $X^{(2)}$, et conduisant ainsi à un échec du contrôle de l'erreur de type I post-clustering.

En élargissant notre investigation des scénarios simulés aux applications réelles, nous avons utilisé un ensemble de données scRNA-seq du consortium Tabula Sapiens (Consortium* et al. 2022) pour explorer les défis pratiques liés à l'estimation de la surdispersion. Notre analyse s'est concentrée sur cinq populations cellulaires distinctes : 2 560 neutrophiles, 105 macrophages, 386 monocytes, 454 granulocytes et 833 cellules T CD4, toutes issues d'un seul donneur. Dans ce cadre contrôlé, où les types de cellules sont connus, nous avons réussi à estimer la surdispersion de 8 333 gènes pour chaque type cellulaire en utilisant la méthode `vst` implémentée dans le package `sctransform` (Choudhary & Satija 2022). La Figure 5.5C illustre la comparaison de la surdispersion des gènes lorsqu'elle est estimée uniquement chez les neutrophiles par rapport à la surdispersion estimée pour les mêmes gènes dans les quatre autres populations cellulaires. Les valeurs de l'Erreur Quadratique Moyenne (EQM) ont été calculées pour quantifier l'accord entre les estimations. Nos résultats révèlent que la surdispersion est spécifique à chaque population cellulaire, comme en témoigne une déviation notable de la diagonale et des valeurs de RMSE relativement élevées. Cela souligne le défi d'estimer précisément ce paramètre sans connaissance préalable du vrai mélange sous-jacent aux données. Combinés à nos études de simulations, ces résultats montrent comment l'erreur de type I peut

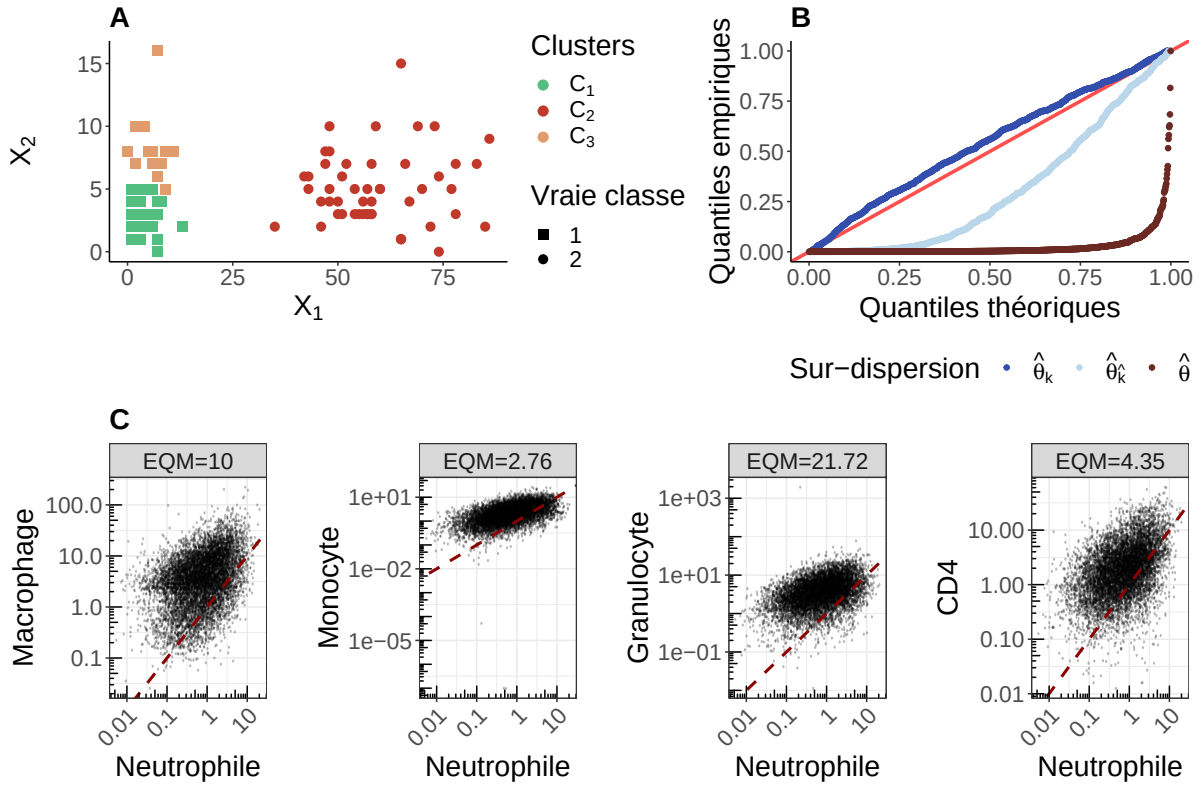


FIGURE 5.5 – **Défis liés à l'estimation du paramètre de surdispersion, un paramètre gène-spécifique, pour la dilution binomiale négative.** Le panneau A illustre les résultats d'un mauvais clustering sur un mélange de binomiales négatives à deux composantes, et le panneau B donne le QQ-plot contre la distribution uniforme des p-valeurs du test de Wilcoxon post-clustering entre le cluster 1 et le cluster 3 lorsque la dilution de données est appliquée avec différentes estimations de la surdispersions pour 1000 réplicats de l'expérience. Le panneau C fournit les estimations gène par gène de la surdispersion dans quatre populations cellulaires différentes (macrophages, monocytes, granulocytes, and CD4) comparées à celle des neutrophiles avec les Erreurs Quadratiques Moyennes (EQM) associées.

facilement être enflée dans des applications réelles de la dilution de données pour l'analyse des données scRNA-seq, en particulier en raison de la difficulté à fournir une estimation non biaisée de la surdispersion des gènes sans connaissances préalables sur le mélange.

5.4 Discussion

Dans ce travail, nous avons souligné les limites pratiques inhérentes à la fission de données et à son extension, la dilution de données, pour répondre aux défis d'inférence post-clustering. Un problème crucial est l'hypothèse d'une distribution homogène des données, une distribution simple, qui ne peut que traduire une absence totale de véritables clusters dans les données. Pour répondre à cette limitation et s'adapter aux scénarios où de réels clusters peuvent être présents, un passage aux modèles de mélange devient impératif.

Cependant, ces modèles manquent d'une décomposition prédéfinie par fission ou dilution de données.

Nous avons proposé une décomposition intra-composante pour la fission et la dilution des données, et démontré sa validité théorique. Elle repose sur la connaissance a priori des paramètres d'échelle des composantes du mélange, tels que les variances dans la distribution gaussienne ou la surdispersion dans la distribution binomiale négative. Cependant, dans les applications réelles, ces paramètres sont inconnus. Estimer adéquatement ces paramètres devient complexe en présence de véritables clusters, étant donné leur nature spécifique aux composantes ; tandis que la qualité de l'estimation de ces paramètres est directement liée à la covariance entre les nouvelles variables aléatoires, $X^{(1)}$ et $X^{(2)}$, issues de la décomposition des données originales. Seule une estimation non biaisée de ces paramètres assure l'indépendance entre $X^{(1)}$ et $X^{(2)}$. Cette indépendance est primordiale pour que l'inférence post-clustering garantisse un control adéquate de l'erreur de type I.

Dans le cadre gaussien, nous avons quantifié théoriquement la relation entre le biais relatif dans l'estimation de la variance et l'erreur de type I associée dans les tests t post-clustering. En pratique, nos résultats de simulations suggèrent qu'un petit biais relatif peut être acceptable tout en permettant un contrôle efficace de l'erreur de type I. Ces premiers résultats ouvrent la voie à la définition d'une approche rationnelle pour ajuster l'hyperparamètre τ dans la fission et la dilution de données pour l'inférence post-clustering afin d'optimiser la puissance statistique.

Comme solution pour éviter la nécessité de connaître a priori les paramètres de variance spécifiques aux composantes dans le cadre du mélange gaussien, nous proposons un modèle hétéroscédastique avec des variances individuelles, pouvant être estimé par un estimateur pondéré non paramétrique. Cette approche s'aligne davantage avec les hypothèses distributionnelles faites par la fission et la dilution de données. Cependant, les performances de cette approche non paramétrique dépendent fortement du choix du paramètre de lissage. La meilleure valeur de ce paramètre serait celle ne capturant que les observations provenant d'une même composante, mais cela nécessiterait à nouveau la connaissance des véritables composantes du mélange, alors que leur estimation fait l'objet du clustering, ayant lieu post-fission.

Enfin, nous avons démontré que les résultats obtenus dans le cadre gaussien sont facilement généralisables à la distribution binomiale négative couramment utilisée pour modéliser les données RNA-seq. En particulier, nous avons illustré sur des données réelles que la surdispersion est également spécifique aux composantes. Par conséquent, sans connaître le vériable processus latent dans les données, la dilution de données ne peut pas garantir l'indépendance nécessaire entre le clustering et les tests d'analyse différentielle.

En pratique, l'application de la fission ou de la dilution de données pour l'inférence post-clustering semble s'apparenter à un cercle vicieux généré par un raisonnement circulaire. Bien que présentées comme une solution pour relever les défis de l'inférence post-

clustering, toutes les stratégies qui pourraient théoriquement assurer l'indépendance entre les deux étapes de l'analyse reposent en fin de compte sur la connaissance des clusters réels, mais inconnus. Malgré son attrait conceptuel, l'utilité pratique de ces méthodes pour l'inférence post-clustering reste limitée aux cas extrêmes avec des rapports signal/bruit extrêmement faibles, soulignant la nécessité de méthodologies alternatives capables de gérer plus efficacement la complexité induite par les structures de classes inconnues. Tous les codes et données nécessaires pour reproduire les résultats présentés ici sont accessibles librement sur Zenodo avec le DOI 10.5281/zenodo.11207777 (Hivert, Agniel, Thiébaut & Hejblum 2024).

Chapitre 6

Conclusion et perspectives

Dans cette thèse, j'ai étudié les enjeux méthodologiques liés au clustering et à l'analyse différentielle de données d'expression génique. J'ai mis en évidence un défi statistique majeur : l'analyse différentielle post-clustering, où les données sont utilisées successivement pour d'abord définir des clusters puis à nouveau ré-utilisées pour tester des différences entre ces clusters. Cette double utilisation des données contredit le cadre de l'inférence statistique classique où les hypothèses de tests doivent être fixées a priori, et conduit à une inflation directe de l'erreur de type I. Cela signifie que les tests statistiques traditionnels, qui sont utilisés sans tenir compte de cette double utilisation des données et des incertitudes résultant du clustering, peuvent générer un nombre excessif de fausses découvertes. Dans le contexte des données RNA-seq, cette problématique est particulièrement critique. En effet, ces analyses en deux étapes sont couramment mises en oeuvre en pratique et les conclusions biologiques qui en sont tirées peuvent être à l'origine d'expériences supplémentaires, souvent coûteuses en temps et en ressources. Ce travail contribue à la génération de résultats plus rigoureux et statistiquement valides d'une part en évaluant l'impact de la standardisation sur les performances du clustering et d'autre part en proposant et en étudiant des méthodes répondant aux enjeux de l'inférence post-clustering.

En amont du clustering et de l'analyse différentielle, les données RNA-seq subissent un nombre important de prétraitements, chacun influençant les performances du clustering. Je me suis alors intéressé à l'impact de la standardisation par les z-scores sur la qualité du clustering. Traditionnellement, cette étape se justifie notamment par la volonté de rendre les variables comparables entre elles, afin qu'elles aient toutes le même poids dans la construction des clusters. Cependant, dans le cadre des données d'expression génique, toutes les variables sont déjà mesurées dans une même échelle puisqu'elles décrivent toutes un même processus biologique : l'expression génique. Bien que peu étudiée jusqu'à présent, j'ai montré à l'aide de simulations numériques et de 4 jeux de données RNA-seq en masse que cette standardisation pouvait détériorer la qualité du clustering. Standardiser les données devient alors seulement nécessaire dans ce contexte lorsque la variance du bruit est plus forte que celle des variables informatives pour la définition des clusters. Néanmoins, dissocier les variables informatives des variables de bruit, afin de déterminer a priori si la standardisation sera bénéfique ou non, n'est possible que si les clusters sont connus en amont, ce qui est irréalisable en pratique puisque tout l'objectif du

clustering est de les identifier. Par conséquent, nous recommandons une approche pragmatique : comparer systématiquement les résultats du clustering obtenus avec les données originales et les données standardisées. Cette comparaison permet d'évaluer directement l'impact de la standardisation et de choisir l'option qui conduit aux clusters ayant le plus de sens biologiquement parlant.

Pour répondre efficacement aux enjeux de l'inférence post-clustering, deux approches basées sur les concepts d'inférence sélective sont actuellement utilisées : le conditionnement sur l'événement de clustering et la décomposition explicite de l'information. Ces méthodes ont pour objectif commun de dissocier l'information utilisée pour constuire les clusters de celle utilisée pour l'analyse différentielle. Cette dissociation permet de préserver les propriétés statistiques essentielles des tests appliqués après le clustering. Au début de cette thèse, en 2020, les options pour l'inférence post-clustering étaient limitées. Deux méthodes principales existaient alors : l'une, qui combine les deux approches mentionnées (conditionnement et décomposition de l'information) (Zhang et al. 2019), ce qui la rend souvent trop complexe et difficile à mettre en œuvre en pratique. L'autre méthode, basée sur un conditionnement sur le clustering (Gao et al. 2024), se concentre uniquement sur le test des différences de barycentres entre deux clusters, empêchant ainsi son application directe pour l'analyse différentielle où l'intérêt réside à l'échelle de la variable. Ces limitations ont souligné la nécessité de développer des méthodes plus robustes et applicables pour résoudre les problèmes d'inférence post-clustering dans les études d'expression génique en faisant un domaine de recherche maintenant particulièrement actif. Notre travail s'inscrit dans cette perspective, en cherchant à adapter les approches existantes et à proposer des solutions adaptées aux besoins spécifiques des analyses génomiques.

Nous avons proposé une adaptation du test d'inférence sélective de Gao et al. (2024) pour tester les différences de moyennes entre deux clusters, non plus à l'échelle des barycentres, mais à l'échelle des variables, conformément aux exigences de l'analyse différentielle. Cette adaptation, qui repose sur un conditionnement sur l'événement de clustering, garantit un contrôle effectif de l'erreur de type I après le clustering. Cependant, ce conditionnement présente plusieurs défis. Pour commencer, bien que nous puissions dériver une p-valeur explicite pour ce test, son calcul est complexe. Il nécessite de décrire l'ensemble de toutes les perturbations des données qui permettent de conserver les clusters lorsque la même méthode de clustering est ré-appliquée. Pour surmonter cette difficulté, nous utilisons une approche de Monte-Carlo avec échantillonnage préférentiel pour estimer les p-valeurs. Deuxièmement, dans le domaine de l'inférence sélective, il est établi que le conditionnement sur l'événement de sélection (ici le clustering) peut réduire la puissance statistique. Dans notre cas, cette réduction de puissance est renforcée par la présence de clusters intercalés entre les deux clusters testés. Pour atténuer cette perte, nous avons développé une extension du test qui intègre l'information de tous les clusters intercalés pour mieux inférer leur séparation. Cette version agrégée du test maintient le contrôle

de l'erreur de type I tout en augmentant la puissance par rapport au test d'inférence sélective original. Néanmoins, cette amélioration augmente le coût computationnel, car la procédure de Monte-Carlo doit être appliquée à toutes les paires de clusters adjacents intercalés. En réponse à ces limitations, nous avons également proposé d'utiliser un test non-paramétrique de multimodalité, le test Dip (Hartigan et al. 1985), pour détecter la présence d'un continuum entre deux clusters. Contrairement à la multimodalité, la présence d'un tel continuum ne peut pas être causée par un mauvais clustering des données. Ce test, bien qu'il n'offre aucune garantie théorique quant au contrôle de l'erreur de type I après le clustering, peut être considéré comme une extension du test d'inférence sélective, où la notion de "perturbations préservant le clustering" est étendue à celle de "continuum entre les clusters". Les performances et les limites de ces trois tests ont été évaluées par des simulations numériques et des applications à des jeux de données réelles provenant de l'écologie et de la médecine. S'ils n'ont pas été appliqués pour tester l'expression différentielle sur des données RNA-seq, c'est parce que ces tests ne sont pas directement adaptés à la nature de ces données d'expression génique. D'abord, la grande dimension des données augmente considérablement le temps de calcul pour le test d'inférence sélective, car l'approche de Monte-Carlo nécessite de répéter le clustering sur de nombreuses versions perturbées des données. Ensuite, le test de multimodalité repose sur l'hypothèse d'unimodalité au sein des clusters, ce qui n'est pas toujours vérifié dans les données d'expression génique, notamment en raison de l'inflation en zéro. Enfin, tous ces tests ignorent les corrélations entre variables, ce qui peut poser problème. Bien que le test de multimodalité soit robuste face à ces corrélations, les tests d'inférence sélective, qui sont basés sur des perturbations des données, échouent lorsque les corrélations sont trop fortes.

Pour alors relever le défi de l'analyse différentielle post-clustering, nous nous sommes focalisés sur les méthodes de décomposition. L'approche naïve consistant à diviser l'échantillon en deux parties distinctes (apprentissage et test, comme en apprentissage automatique) est impraticable dans ce contexte en raison de la difficulté à transférer les labels de clustering du jeu d'apprentissage au jeu de test. Par conséquent, nous avons exploré des stratégies plus sophistiquées de décomposition de l'information au niveau de chaque observation : la fission de données et la dilution de données. Ces deux méthodes ont pour but de diviser l'information de chaque observation en deux parties indépendantes, généralement en ajoutant du bruit aux données originales. Cela permet d'utiliser une partie pour le clustering et l'autre pour l'analyse différentielle. Étant donné que chaque observation est présente dans les deux ensembles de données, le problème de transfert des labels disparaît. Enfin, leur indépendance garantit un contrôle effectif de l'erreur de type I pour tout test statistique bien calibré. De plus, ces méthodes présentent deux avantages majeurs par rapport aux approches basées sur le conditionnement. Premièrement, elles offrent une grande flexibilité : n'importe quelle méthode de clustering et d'analyse différentielle peut être employée. À l'inverse, les méthodes conditionnelles nécessitent souvent

le développement de tests spécifiques, parfois utilisable uniquement pour une méthode de clustering donnée. Deuxièmement, ces méthodes de découpage de l'information sont compatibles avec plusieurs distributions de probabilité, notamment les distributions de Poisson, gaussienne et binomiale négative. Cela est particulièrement pertinent pour les données d'expression génique, qui sont souvent des comptes (modélisables par les distributions de Poisson et binomiale négative) ou présentent des corrélations entre variables (capturées par la distribution gaussienne multivariée). Bien que les méthodes de fission et de dilution de données aient été proposées en réponse aux défis de l'analyse différentielle post-clustering et soient théoriquement valides, j'ai montré qu'elles sont difficiles à appliquer en pratique. En effet, de nombreuses décompositions nécessitent la connaissance d'un paramètre d'échelle, souvent inconnu et qui doit être estimé au préalable. Dans le cas des données gaussiennes, nous avons démontré que la qualité de cette estimation affecte directement l'indépendance des deux parties générées. Nous avons aussi quantifié, de manière théorique, l'inflation de l'erreur de type I qui en résulte. Estimer correctement ces paramètres d'échelle est particulièrement complexe car leurs valeurs sont spécifiques à chaque cluster, qui sont eux-mêmes inconnus et à estimer. Mêmes des méthodes d'estimation non-paramétriques, faisant l'hypothèse que ces paramètres sont spécifiques à chaque individu permettant ainsi de s'affranchir de la connaissance a priori des clusters, ne fonctionnent que pour des cas exagérés de séparation entre ces clusters. Cela révèle toute la complexité inhérente à ce problème. En conséquence, l'application de la fission et de la dilution de données dans le cadre de l'inférence post-clustering ne résout pas la circularité dans le processus d'analyse : pour utiliser ces méthodes de manière rigoureuse afin d'identifier des clusters dans les données et tester des différences entre eux, il faut déjà connaître la structure de partition des données. Toutefois, ces méthodes restent applicables dans d'autres contextes d'inférence sélective, comme la sélection Lasso, où une hypothèse d'homogénéité des données sans mélange est plus raisonnable.

Cette thèse met en lumière le rôle essentiel de la variance des variables dans les problématiques de clustering et d'analyse différentielle, en particulier pour les données RNA-seq. Nous avons démontré que la variance peut informer sur la présence ou l'absence de clusters et guider efficacement les résultats du clustering. Dans le contexte des données RNA-seq, où les variables sont mesurées sur une même échelle, la variance devient un indicateur clé pour déceler la structure en groupes sous-jacente des données, comme le montre notre analyse de l'impact de la standardisation par les z-scores. La connaissance de la variance est également cruciale pour les méthodes d'inférence post-clustering, où l'estimation précise de ce paramètre pose un défi majeur, étant donné que les clusters eux-mêmes sont à identifier. En effet, cette variance encode le bruit qui est ajouté aux données permettant de garantir que la même information n'est pas utilisée deux fois. Elle définit alors les perturbations qui sont appliquées dans les tests d'inférence sélective et permet de garantir l'indépendance nécessaire dans les approches de fission et de la dilution de données. Une

estimation incorrecte de la variance peut entraîner des biais dans le contrôle de l'erreur de type I et réduire la puissance statistique des tests. Ce problème a déjà été abordé par Yun & Foygel Barber (2023) dans le cadre de l'inférence post-clustering pour des méthodes basées sur le conditionnement, où ils ont développé un test valide même lorsque la variance est inconnue. Notre travail, complémentaire au leur, a mis en évidence cette même problématique pour les approches basées sur la décomposition de l'information, montrant que la précision de l'estimation de la variance est tout aussi critique pour garantir l'efficacité et la validité de ces méthodes sans pour autant apporter de solutions.

Ce travail de thèse a mis en évidence la circularité inhérente qui se manifeste dans la problématique de l'inférence post-clustering. Pour commencer, les analyses basées sur l'inférence post-clustering sont de fait circulaires en raison de la double utilisation des données. Cette circularité reste présente dans les méthodes développées spécifiquement pour tenter de résoudre ce problème d'inférence post-clustering. En effet, pour garantir les performances de ces méthodes, il faudrait connaître les clusters a priori. Enfin, elle fait écho aux problématiques de standardisation : des recherches antérieures ont montré que la meilleure standardisation est celle effectuée au sein de chaque cluster (Cormack 1971, Hartigan 1975, Everitt 1980). Toutefois, cette connaissance des clusters est précisément ce que les méthodes de clustering visent à découvrir. Ainsi, à la fois pour l'inférence post-clustering et pour la standardisation des données, les approches les plus efficaces reposent sur des informations que ces mêmes approches cherchent à estimer. Cette circularité met en lumière toute la complexité du problème d'inférence post-clustering particulièrement causée par la nature non-supervisée du clustering lui-même.

Pour l'avenir, plusieurs axes de recherche s'ouvrent afin de surmonter cette circularité pour rendre possible l'application des méthodes étudiées. Une priorité serait de perfectionner les techniques d'estimation de la variance, en explorant notamment des approches non paramétriques capables de prendre en compte la proximité entre les observations pour imiter la structure en clusters inconnus. Une autre piste, inspirée des travaux de Yun & Foygel Barber (2023), consisterait à contourner la nécessité préalable de connaître la variance avant les décompositions. Ce concept pourrait être réalisé par le développement de méthodes de décomposition inspirées des tests de permutations, générant de manière non paramétrique des données synthétiques sous l'hypothèse nulle d'absence de clusters, auxquelles les données observées pourraient être comparées. Ce genre d'approches a commencé à être exploré par Song et al. (2023) mais uniquement dans le cas particulier où seulement deux clusters sont estimés. Or, dans ce cas particulier, les méthodes basées sur l'inférence sélective se révèlent déjà assez efficaces. En effet, lorsque seuls deux clusters sont en jeu, l'hypothèse nulle se ramène à une situation binaire : soit ces deux clusters n'existent pas vraiment et les données sont homogènes, soit ils existent bien et sont distincts. Cette simplification réduit la complexité de l'estimation de la variance puisqu'il n'y a pas de clusters supplémentaires susceptibles d'influencer cette estimation. L'estimateur

global de la variance (utilisant toutes les observations) tend alors à surestimer la variance intra-cluster, ce qui conduit à un test d'inférence sélective moins puissant mais contrôlant l'erreur de type I (Gao et al. 2024). Pour les méthodes basées sur la décomposition, ce cas où seulement deux clusters sont présents constitue également une situation spécifique : si ces deux clusters sont effectivement séparés, l'estimateur global de la variance présentera un biais par rapport à la vraie variance intra-cluster, introduisant donc des corrélations entre les deux parties de l'information. Cependant, lorsque les $K = 2$ clusters sont correctement estimés dans la première partie des données, ces corrélations peuvent faciliter leur transfert à la seconde partie. À l'inverse, si les deux clusters ne sont pas réellement séparés, l'estimateur global de la variance deviendra non biaisé par rapport à la vraie variance, garantissant ainsi l'indépendance entre les deux parties et le contrôle de l'erreur de type I. Ce scénario à $K = 2$ pourrait ainsi servir de fondement à une méthode d'inférence post-clustering itérative, en écho à notre test d'inférence sélective par agrégation des p-valeurs adjacentes, restreignant l'analyse aux situations où seuls deux clusters sont estimés pour lesquels les méthodes étudiées restent valides bien que peu puissantes.

En conclusion, cette thèse a mis en évidence les défis complexes et les enjeux cruciaux associés à l'analyse différentielle de données d'expression génique en présence d'un clustering. Nos contributions visent à améliorer la fiabilité et la rigueur des résultats en développant des approches innovantes tout en identifiant les limitations fondamentales des méthodes existantes. Pour l'avenir, il est essentiel de poursuivre la recherche dans la standardisation des données, l'estimation précise de la variance en grande dimension, et l'exploration de nouvelles approches d'inférence post-clustering pour répondre aux défis croissants posés par ces données biologiques complexes. Ces efforts continus sont essentiels pour garantir des découvertes biologiques robustes et fiables, permettant ainsi de mieux comprendre les mécanismes biologiques sous-jacents et d'orienter efficacement la recherche future dans le domaine des sciences médicales et particulièrement de l'immunologie.

Bibliographie

Agresti, A. (2015), *Foundations of linear and generalized linear models*, John Wiley & Sons. (Cité en page 16.)

Ameijeiras-Alonso, J., Crujeiras, R. M. & Rodriguez-Casal, A. (2021), 'multimode : An r package for mode assessment', *Journal of Statistical Software* **97**(1), 1–32. (Cité en pages 73 et 143.)

Amezquita, R. A., Lun, A. T., Becht, E., Carey, V. J., Carpp, L. N., Geistlinger, L., Marini, F., Rue-Albrecht, K., Risso, D., Soneson, C. et al. (2020), 'Orchestrating single-cell analysis with bioconductor', *Nature methods* **17**(2), 137–145. (Cité en page 37.)

Anderberg, M. R. (1973), *Cluster analysis for applications*, Probability and mathematical statistics, 19, Academic Press, New York. (Cité en page 46.)

Anders, S. & Huber, W. (2010), 'Differential expression analysis for sequence count data', *Nature Precedings* pp. 1–1. (Cité en page 16.)

Argelaguet, R., Arnol, D., Bredikhin, D., Deloro, Y., Velten, B., Marioni, J. C. & Stegle, O. (2020), 'Mofa+ : a statistical framework for comprehensive integration of multi-modal single-cell data', *Genome biology* **21**, 1–17. (Cité en page 18.)

Aschenbrenner, A. C., Mouktaroudi, M., Krämer, B., Oestreich, M., Antonakos, N., Nuesch-Germano, M., Gkizeli, K., Bonaguro, L., Reusch, N., Baßler, K. et al. (2021), 'Disease severity-specific neutrophil signatures in blood transcriptomes stratify covid-19 patients', *Genome medicine* **13**, 1–25. (Cité en page 19.)

Azzalini, A. (2013), *The skew-normal and related families*, Vol. 3, Cambridge University Press. (Cité en page 154.)

Bachoc, F., Maugis-Rabusseau, C. & Neuvial, P. (2023), 'Selective inference after convex clustering with ℓ_1 penalization', *arXiv preprint arXiv :2309.01492* . (Cité en pages 64 et 72.)

Bakk, Z., Oberski, D. L. & Vermunt, J. K. (2014), 'Relating latent class assignments to external variables : Standard errors for correct inference.', *Political analysis* **22**(4). (Cité en page 40.)

Balakrishnan, P., Cooper, M. C., Jacob, V. S. & Lewis, P. A. (1994), 'A study of the classification capabilities of neural networks using unsupervised learning : A comparison with k-means clustering', *Psychometrika* **59**, 509–525. (Cité en page 58.)

Bellman, R., Bellman, R. & Corporation, R. (1957), *Dynamic Programming*, Rand Corporation research study, Princeton University Press.
URL: *https ://books.google.fr/books ?id=rZW4ugAACAAJ* (Cité en page 21.)

Benjamini, Y. & 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.
URL: *http ://www.jstor.org/stable/2346101* (Cité en page 35.)

Bernardes, J. P., Mishra, N., Tran, F., Bahmer, T., Best, L., Blase, J. I., Bordoni, D., Franzenburg, J., Geisen, U., Josephs-Spaulding, J. et al. (2020), 'Longitudinal multi-omics analyses identify responses of megakaryocytes, erythroid cells, and plasmablasts as hallmarks of severe covid-19', *Immunity* **53**(6), 1296–1314. (Cité en page 18.)

Bishop, C. M. (2006), 'Pattern recognition and machine learning', *Springer google schola* **2**, 5–43. (Cité en page 30.)

Blanco-Melo, D., Nilsson-Payant, B. E., Liu, W.-C., Uhl, S., Hoagland, D., Møller, R., Jordan, T. X., Oishi, K., Panis, M., Sachs, D. et al. (2020), 'Imbalanced host response to sars-cov-2 drives development of covid-19', *Cell* **181**(5), 1036–1045. (Cité en page 19.)

Blondel, V. D., Guillaume, J.-L., Lambiotte, R. & Lefebvre, E. (2008), 'Fast unfolding of communities in large networks', *Journal of statistical mechanics : theory and experiment* **2008**(10), P10008. (Cité en page 31.)

Bonafede, E., Picard, F., Robin, S. & Viroli, C. (2016), 'Modeling overdispersion heterogeneity in differential expression analysis using mixtures', *Biometrics* **72**(3), 804–814. (Cité en page 16.)

Bouveyron, C., Celeux, G., Murphy, T. B. & Raftery, A. E. (2019), *Model-based clustering and classification for data science : with applications in R*, Vol. 50, Cambridge University Press. (Cité en page 30.)

Brusco, M. J. & Cradit, J. D. (2001), 'A variable-selection heuristic for K-means clustering', *Psychometrika* **66**(2), 249–270.
URL: *http ://link.springer.com/10.1007/BF02294838* (Cité en page 58.)

Carmone Jr, F. J., Kara, A. & Maxwell, S. (1999), 'Hinov : A new model to improve market segment definition by identifying noisy variables', *Journal of Marketing Research* **36**(4), 501–509. (Cité en page 58.)

Chacón, J. E. & Duong, T. (2020), *Multivariate Kernel Smoothing and Its Applications SMOOTHING AND ITS APPLICATIONS*, CRC PRESS. (Cité en page 105.)

Chari, T. & Pachter, L. (2023), 'The specious art of single-cell genomics', *PLOS Computational Biology* **19**(8), e1011288. (Cité en page 38.)

Chen, Y. T. & Gao, L. L. (2023), 'Testing for a difference in means of a single feature after clustering', *arXiv preprint arXiv :2311.16375* . (Cité en page 93.)

Chen, Y. T. & Witten, D. M. (2023), 'Selective inference for k-means clustering', *Journal of Machine Learning Research* **24**(152), 1–41. (Cité en page 63.)

Chiquet, J., Mariadassou, M. & Robin, S. (2018), 'Variational inference for probabilistic Poisson PCA', *The Annals of Applied Statistics* **12**(4), 2674 – 2698.
URL: *https ://doi.org/10.1214/18-AOAS1177* (Cité en page 23.)

Choudhary, S. & Satija, R. (2022), 'Comparison and evaluation of statistical error models for scrna-seq', *Genome Biology* **23**, 20.
URL: *https ://doi.org/10.1186/s13059-021-02584-9* (Cité en page 114.)

Cole, M. B., Risso, D., Wagner, A., DeTomaso, D., Ngai, J., Purdom, E., Dudoit, S. & Yosef, N. (2019), 'Performance assessment and selection of normalization procedures for single-cell rna-seq', *Cell systems* **8**(4), 315–328. (Cité en page 44.)

Consortium*, T. T. S., Jones, R. C., Karkanias, J., Krasnow, M. A., Pisco, A. O., Quake, S. R., Salzman, J., Yosef, N., Bulthaup, B., Brown, P. et al. (2022), 'The tabula sapiens : A multiple-organ, single-cell transcriptomic atlas of humans', *Science* **376**(6594), eabl4896. (Cité en page 114.)

Cormack, R. M. (1971), 'A review of classification', *Journal of the Royal Statistical Society : Series A (General)* **134**(3), 321–353. (Cité en page 123.)

Cotugno, N., Ruggiero, A., Santilli, V., Manno, E. C., Rocca, S., Zicari, S., Amodio, D., Colucci, M., Rossi, P., Levy, O. et al. (2019), 'Omic technologies and vaccine development : from the identification of vulnerable individuals to the formulation of invulnerable vaccines', *Journal of immunology research* **2019**. (Cité en page 14.)

Dahl, D. B. (2006), 'Model-based clustering for expression data via a dirichlet process mixture model', *Bayesian inference for gene expression and proteomics* **4**, 201–218. (Cité en page 29.)

Defays, D. (1977), 'An efficient algorithm for a complete link method', *The computer journal* **20**(4), 364–366. (Cité en page 26.)

Dempster, A. P., Laird, N. M. & 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. (Cité en page 30.)

Donoho, D. L. et al. (2000), 'High-dimensional data analysis : The curses and blessings of dimensionality', *AMS math challenges lecture* **1**(2000), 32. (Cité en page 20.)

Duò, A., Robinson, M. D. & Soneson, C. (2018), 'A systematic performance evaluation of clustering methods for single-cell rna-seq data', *F1000Research* **7**. (Cité en pages 43 et 45.)

Eppstein, D., Paterson, M. S. & Yao, F. F. (1997), 'On nearest-neighbor graphs', *Discrete & Computational Geometry* **17**, 263–282. (Cité en page 31.)

Everitt, B. (1980), 'Cluster analysis.', *Quality & Quantity* **14**(1). (Cité en page 123.)

Everitt, B. S. & Hothorn, T. (2006), *A handbook of statistical analyses using R*, Chapman & Hall, Boca Raton, FL. (Cité en pages 23, 24 et 47.)

Finak, G., Langweiler, M., Jaimes, M., Malek, M., Taghiyar, J., Korin, Y., Raddassi, K., Devine, L., Obermoser, G., Pekalski, M. L. et al. (2016), 'Standardizing flow cytometry immunophenotyping analysis from the human immunophenotyping consortium', *Scientific reports* **6**(1), 1–11. (Cité en page 91.)

Finak, G., McDavid, A., Yajima, M., Deng, J., Gersuk, V., Shalek, A. K., Slichter, C. K., Miller, H. W., McElrath, M. J., Prlic, M. et al. (2015), 'Mast : a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell rna sequencing data', *Genome biology* **16**, 1–13. (Cité en page 37.)

Fithian, W., Sun, D. & Taylor, J. (2014), 'Optimal inference after model selection', *arXiv preprint arXiv :1410.2597* . (Cité en pages 59, 61 et 65.)

Gao, L. L., Bien, J. & Witten, D. (2024), 'Selective inference for hierarchical clustering', *Journal of the American Statistical Association* **119**(545), 332–342. (Cité en pages 61, 62, 63, 64, 73, 74, 75, 76, 77, 94, 120 et 124.)

Gauthier, M., Agniel, D., Thiébaut, R. & Hejblum, B. P. (2020), 'dearseq : a variance component score test for rna-seq differential analysis that effectively controls the false discovery rate', *NAR genomics and bioinformatics* **2**(4), lqaa093. (Cité en pages 37 et 52.)

Giraud, C. (2021), *Introduction to high-dimensional statistics*, Chapman and Hall/CRC. (Cité en pages 21 et 22.)

Goeman, J. J. & Solari, A. (2023), 'On selection and conditioning in multiple testing and selective inference', *Biometrika* p. asad078. (Cité en page 60.)

Guo, J., Levina, E., Michailidis, G. & Zhu, J. (2010), 'Pairwise variable selection for high-dimensional model-based clustering', *Biometrics* **66**(3), 793–804. (Cité en page 72.)

Hao, Y., Stuart, T., Kowalski, M. H., Choudhary, S., Hoffman, P., Hartman, A., Srivastava, A., Molla, G., Madad, S., Fernandez-Granda, C. & Satija, R. (2023), 'Dictionary learning for integrative, multimodal and scalable single-cell analysis', *Nature Biotechnology* .
URL: *https ://doi.org/10.1038/s41587-023-01767-y* (Cité en page 32.)

Hartigan, J. A. (1975), *Clustering algorithms*, John Wiley & Sons, Inc. (Cité en page 123.)

Hartigan, J. A., Hartigan, P. M. et al. (1985), 'The dip test of unimodality', *Annals of statistics* **13**(1), 70–84. (Cité en pages 73, 79, 80 et 121.)

Heidenreich, N.-B., Schindler, A. & Sperlich, S. (2013), 'Bandwidth selection for kernel density estimation : a review of fully automatic selectors', *AStA Advances in Statistical Analysis* **97**, 403–433. (Cité en page 104.)

Hejblum, B. P., Alkhassim, C., Gottardo, R., Caron, F. & Thiébaut, R. (2019), 'Sequential Dirichlet process mixtures of multivariate skew t-distributions for model-based clustering of flow cytometry data', *The Annals of Applied Statistics* **13**(1), 638 – 660.
URL: *https ://doi.org/10.1214/18-AOAS1209* (Cité en page 29.)

Hennig, C., Meila, M., Murtagh, F. & Rocci, R. (2015), *Handbook of cluster analysis*, CRC Press. (Cité en page 23.)

Hicks, S. C., Townes, F. W., Teng, M. & Irizarry, R. A. (2018), 'Missing data and technical variability in single-cell rna-sequencing experiments', *Biostatistics* **19**(4), 562–578. (Cité en page 17.)

Hinneburg, A., Aggarwal, C. C. & Keim, D. A. (2000), 'What is the nearest neighbor in high dimensional spaces ?'. (Cité en page 21.)

Hivert, B., Agniel, D., Thiébaut, R. & Hejblum, B. P. (2024), 'Post-clustering difference testing : Valid inference and practical considerations with applications to ecological and biological data', *Computational Statistics & Data Analysis* **193**, 107916. (Cité en page 112.)

Hivert, B., Agniel, D., Thiébaut, R. & Hejblum, B. P. (2023), 'Reproducible codes and results for post-clustering difference testing valid inference (Version v2) [Data set]', *Zenodo* . DOI : 10.5281/zenodo.7660128.
URL: *https ://doi.org/10.5281/zenodo.7660128* (Cité en page 93.)

Hivert, B., Agniel, D., Thiébaut, R. & Hejblum, B. P. (2024), 'Reproducible codes and results for Running in circles : is practical application feasible for data fission and data thinning in post-clustering differential analysis ? (Version v1) [Data set]', *Zenodo* . DOI :

10.5281/zenodo.11207777.
URL: *https ://doi.org/10.5281/zenodo.11207777* (Cité en page 117.)

Holm, S. (1979), 'A simple sequentially rejective multiple test procedure', *Scandinavian journal of statistics* pp. 65–70. (Cité en page 22.)

Horst, A. M., Hill, A. P. & Gorman, K. B. (2020), *palmerpenguins : Palmer Archipelago (Antarctica) penguin data*. R package version 0.1.0.
URL: *https ://allisonhorst.github.io/palmerpenguins/* (Cité en page 87.)

Hubert, L. & Arabie, P. (1985), 'Comparing partitions', *Journal of classification* **2**, 193–218. (Cité en page 52.)

Hwang, B., Lee, J. H. & Bang, D. (2018), 'Single-cell rna sequencing technologies and bioinformatics pipelines', *Experimental & molecular medicine* **50**(8), 1–14. (Cité en page 37.)

Hyun, S., Lin, K. Z., G'Sell, M. & Tibshirani, R. J. (2021), 'Post-selection inference for changepoint detection algorithms with application to copy number variation data', *Biometrics* **77**(3), 1037–1049. (Cité en page 61.)

Islam, S., Zeisel, A., Joost, S., La Manno, G., Zajac, P., Kasper, M., Lönnerberg, P. & Linnarsson, S. (2014), 'Quantitative single-cell RNA-seq with unique molecular identifiers', *Nature Methods* **11**(2), 163–166.
URL: *https ://doi.org/10.1038/nmeth.2772* (Cité en page 15.)

Jackson, H., Rivero Calle, I., Broderick, C., Habgood-Coote, D., d'Souza, G., Nichols, S., Vito, O., Gómez-Rial, J., Rivero-Velasco, C., Rodríguez-Núñez, N. et al. (2022), 'Characterisation of the blood rna host response underpinning severity in covid-19 patients', *Scientific reports* **12**(1), 12216. (Cité en page 19.)

Jaskowiak, P. A., Costa, I. G. & Campello, R. J. (2018), 'Clustering of rna-seq samples : Comparison study on cancer data', *Methods* **132**, 42–49. (Cité en pages 44 et 45.)

Jewell, S., Fearnhead, P. & Witten, D. (2022), 'Testing for a change in mean after changepoint detection', *Journal of the Royal Statistical Society Series B : Statistical Methodology* **84**(4), 1082–1104. (Cité en pages 61, 63, 74 et 75.)

Källberg, D., Vidman, L. & Rydén, P. (2021), 'Comparison of methods for feature selection in clustering of high-dimensional rna-sequencing data to identify cancer subtypes', *Frontiers in Genetics* **12**, 632620. (Cité en page 45.)

Kalogeratos, A. & Likas, A. (2012), 'Dip-means : an incremental clustering method for estimating the number of clusters', *Advances in neural information processing systems* **25**, 2393–2401. (Cité en pages 72 et 73.)

Kim, C., Lee, H., Jung, J., Jung, K. & Han, B. (2021), 'Marcopolo : a clustering-free approach to the exploration of differentially expressed genes along with group information in single-cell rna-seq data', *bioRxiv* pp. 2020–11. (Cité en page 72.)

Kiselev, V. Y., Andrews, T. S. & Hemberg, M. (2019), 'Challenges in unsupervised clustering of single-cell rna-seq data', *Nature Reviews Genetics* **20**(5), 273–282. (Cité en page 45.)

Korthauer, K. D., Chu, L.-F., Newton, M. A., Li, Y., Thomson, J., Stewart, R. & Kendziorski, C. (2016), 'A statistical approach for identifying differential distributions in single-cell rna-seq experiments', *Genome biology* **17**, 1–15. (Cité en pages 17, 36 et 37.)

Kotliar, D., Lin, A. E., Logue, J., Hughes, T. K., Khoury, N. M., Raju, S. S., Wadsworth, M. H., Chen, H., Kurtz, J. R., Dighero-Kemp, B., Bjornson, Z. B., Mukherjee, N., Sellers, B. A., Tran, N., Bauer, M. R., Adams, G. C., Adams, R., Rinn, J. L., Melé, M., Schaffner, S. F., Nolan, G. P., Barnes, K. G., Hensley, L. E., McIlwain, D. R., Shalek, A. K., Sabeti, P. C. & Bennett, R. S. (2020), 'Single-cell profiling of ebola virus disease in vivo reveals viral and host dynamics', *Cell* **183**(5), 1383–1401.e19.
URL: *https ://www.sciencedirect.com/science/article/pii/S0092867420313088* (Cité en page 13.)

Kriegeskorte, N., Simmons, W. K., Bellgowan, P. S. & Baker, C. I. (2009), 'Circular analysis in systems neuroscience : the dangers of double dipping', *Nature neuroscience* **12**(5), 535–540. (Cité en page 39.)

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**, 1–35. (Cité en page 59.)

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., Fitzhugh, W. et al. (2001), 'Initial sequencing and analysis of the human genome', *Nature* **412**(6846), 565–566. (Cité en page 14.)

Laurent, B., Marteau, C. & Maugis-Rabusseau, C. (2018), Multidimensional two-component gaussian mixtures detection, *in* 'Annales de l'Institut Henri Poincaré, Probabilités et Statistiques', Vol. 54, Institut Henri Poincaré, pp. 842–865. (Cité en page 82.)

Law, C. W., Chen, Y., Shi, W. & Smyth, G. K. (2014), 'voom : Precision weights unlock linear model analysis tools for rna-seq read counts', *Genome biology* **15**, 1–17. (Cité en pages 17 et 37.)

Lee, J. D., Sun, D. L., Sun, Y. & Taylor, J. E. (2016), 'Exact post-selection inference, with application to the lasso', *The Annals of Statistics* **44**(3), 907 – 927.
URL: *https ://doi.org/10.1214/15-AOS1371* (Cité en pages 60, 61 et 64.)

Lehmann, E. L. (2012), Some history of optimality, *in* 'Selected Works of EL Lehmann', Springer, pp. 1033–1039. (Cité en page 89.)

Leiner, J., Duan, B., Wasserman, L. & Ramdas, A. (2023), 'Data fission : splitting a single data point', *Journal of the American Statistical Association* pp. 1–12. (Cité en pages 66, 67 et 97.)

Lévy, Y., Wiedemann, A., Hejblum, B. P., Durand, M., Lefebvre, C., Surénaud, M., Lacabaratz, C., Perreau, M., Foucat, E., Déchenaud, M. et al. (2021), 'Cd177, a specific marker of neutrophil activation, is associated with coronavirus disease 2019 severity and death', *Iscience* **24**(7). (Cité en pages 18, 19 et 51.)

Li, Q., Noel-MacDonnell, J. R., Koestler, D. C., Goode, E. L. & Fridley, B. L. (2018), 'Subject level clustering using a negative binomial model for small transcriptomic studies', *BMC bioinformatics* **19**, 1–10. (Cité en pages 29 et 113.)

Li, Y., Ge, X., Peng, F., Li, W. & Li, J. J. (2022), 'Exaggerated false positives by popular differential expression methods when analyzing human population samples', *Genome biology* **23**(1), 79. (Cité en page 37.)

Li, Y., Rahman, T., Ma, T., Tang, L. & Tseng, G. C. (2023), 'A sparse negative binomial mixture model for clustering rna-seq count data', *Biostatistics* **24**(1), 68–84. (Cité en page 29.)

Liu, H. & Yu, L. (2005), 'Toward integrating feature selection algorithms for classification and clustering', *IEEE Transactions on knowledge and data engineering* **17**(4), 491–502. (Cité en page 23.)

Liu, J., Wang, J., Xu, J., Xia, H., Wang, Y., Zhang, C., Chen, W., Zhang, H., Liu, Q., Zhu, R. et al. (2021), 'Comprehensive investigations revealed consistent pathophysiological alterations after vaccination with covid-19 vaccines', *Cell Discovery* **7**(1), 1–15. (Cité en page 13.)

Liu, K., Markovic, J. & Tibshirani, R. (2018), 'More powerful post-selection inference, with application to the lasso', *arXiv preprint arXiv :1801.09037* . (Cité en page 61.)

Liu, Y., Li, Z., Xiong, H., Gao, X. & Wu, J. (2010), Understanding of internal clustering validation measures, *in* '2010 IEEE international conference on data mining', IEEE, pp. 911–916. (Cité en page 77.)

Love, M. I., Huber, W. & Anders, S. (2014), 'Moderated estimation of fold change and dispersion for rna-seq data with deseq2', *Genome biology* **15**, 1–21. (Cité en page 37.)

Luecken, M. D. & Theis, F. J. (2019), 'Current best practices in single-cell rna-seq analysis : a tutorial', *Molecular systems biology* **15**(6), e8746. (Cité en pages 44, 45, 49 et 57.)

MacQueen, J. et al. (1967), Some methods for classification and analysis of multivariate observations, *in* 'Proceedings of the fifth Berkeley symposium on mathematical statistics and probability', Vol. 1, Oakland, CA, USA, pp. 281–297. (Cité en pages 27 et 28.)

Marioni, J. C., Mason, C. E., Mane, S. M., Stephens, M. & Gilad, Y. (2008), 'Rna-seq : an assessment of technical reproducibility and comparison with gene expression arrays', *Genome research* **18**(9), 1509–1517. (Cité en page 16.)

Maugis, C., Celeux, G. & Martin-Magniette, M.-L. (2009), 'Variable selection for clustering with gaussian mixture models', *Biometrics* **65**(3), 701–709. (Cité en pages 72, 84 et 85.)

McInnes, L., Healy, J. & Melville, J. (2018), 'Umap : Uniform manifold approximation and projection for dimension reduction', *arXiv preprint arXiv :1802.03426* . (Cité en page 38.)

Miao, Z., Deng, K., Wang, X. & Zhang, X. (2018), 'Desingle for detecting three types of differential expression in single-cell rna-seq data', *Bioinformatics* **34**(18), 3223–3224. (Cité en page 37.)

Milligan, G. W. (1985), 'An algorithm for generating artificial test clusters', *Psychometrika* **50**(1), 123–127.
URL: *http ://link.springer.com/10.1007/BF02294153* (Cité en page 58.)

Milligan, G. W. & Cooper, M. C. (1988), 'A study of standardization of variables in cluster analysis', *Journal of classification* **5**, 181–204. (Cité en pages 47, 48, 49, 50 et 58.)

Molania, R., Foroutan, M., Gagnon-Bartsch, J. A., Gandolfo, L. C., Jain, A., Sinha, A., Olshansky, G., Dobrovic, A., Papenfuss, A. T. & Speed, T. P. (2023), 'Removing unwanted variation from large-scale rna sequencing data with prps', *Nature Biotechnology* **41**(1), 82–95. (Cité en page 44.)

Mounir, Mohamed, Lucchetta, Marta, Silva, C, T., Olsen, Catharina, Bontempi, Gianluca, Chen, Xi, Noushmehr, Houtan, Colaprico, Antonio, Papaleo & Elena (2019), 'New functionalities in the tcgabiolinks package for the study and integration of cancer data from gdc and gtex', *PLoS computational biology* **15**(3), e1006701. (Cité en page 51.)

Müller, D. W. & Sawitzki, G. (1991), 'Excess mass estimates and tests for multimodality', *Journal of the American Statistical Association* **86**(415), 738–746. (Cité en page 73.)

Nagalakshmi, U., Wang, Z., Waern, K., Shou, C., Raha, D., Gerstein, M. & Snyder, M. (2008), 'The transcriptional landscape of the yeast genome defined by rna sequencing', *Science* **320**(5881), 1344–1349. (Cité en page 14.)

Nakagawa, S. (2004), 'A farewell to bonferroni : the problems of low statistical power and publication bias', *Behavioral ecology* **15**(6), 1044–1045. (Cité en page 35.)

Nawy, T. (2014), 'Single-cell sequencing', *Nature methods* **11**(1), 18–18. (Cité en page 15.)

Neufeld, A. C., Gao, L. L. & Witten, D. M. (2022), 'Tree-values : selective inference for regression trees', *Journal of Machine Learning Research* **23**(305), 1–43. (Cité en page 61.)

Neufeld, A., Dharamshi, A., Gao, L. L. & Witten, D. (2024), 'Data thinning for convolution-closed distributions', *Journal of Machine Learning Research* **25**(57), 1–35. (Cité en pages 67, 97, 99 et 105.)

Neufeld, A., Popp, J., Gao, L. L., Battle, A. & Witten, D. (2023), 'Negative binomial count splitting for single-cell rna sequencing data', *arXiv preprint arXiv :2307.12985* . (Cité en pages 99 et 113.)

Obermoser, G., Presnell, S., Domico, K., Xu, H., Wang, Y., Anguiano, E., Thompson-Snipes, L., Ranganathan, R., Zeitner, B., Bjork, A. et al. (2013), 'Systems scale interactive exploration reveals quantitative and qualitative differences in response to influenza and pneumococcal vaccines', *Immunity* **38**(4), 831–844. (Cité en page 13.)

O'Connor, D. & Pollard, A. J. (2013), 'Characterizing vaccine responses using host genomic and transcriptomic analysis', *Clinical infectious diseases* **57**(6), 860–869. (Cité en page 13.)

Oshlack, A., Robinson, M. D. & Young, M. D. (2010), 'From rna-seq reads to differential expression results', *Genome biology* **11**, 1–10. (Cité en page 37.)

Pasquini, G., Arias, J. E. R., Schäfer, P. & Busskamp, V. (2021), 'Automated methods for cell type annotation on scrna-seq data', *Computational and Structural Biotechnology Journal* **19**, 961–969. (Cité en page 38.)

Pasquini, G., Cora, V., Swiersy, A., Achberger, K., Antkowiak, L., Müller, B., Wimmer, T., Fraschka, S. A.-K., Casadei, N., Ueffing, M. et al. (2020), 'Using transcriptomic analysis to assess double-strand break repair activity : towards precise in vivo genome editing', *International journal of molecular sciences* **21**(4), 1380. (Cité en page 37.)

Patro, R., Duggal, G., Love, M. I., Irizarry, R. A. & Kingsford, C. (2017), 'Salmon provides fast and bias-aware quantification of transcript expression', *Nature methods* **14**(4), 417–419. (Cité en page 15.)

Phipson, B. & Smyth, G. K. (2010), 'Permutation p-values should never be zero : calculating exact p-values when permutations are randomly drawn', *Statistical applications in genetics and molecular biology* **9**(1). (Cité en pages 77, 141 et 142.)

Qiu, P. (2020), 'Embracing the dropouts in single-cell rna-seq analysis', *Nature communications* **11**(1), 1169. (Cité en page 17.)

Qiu, W. & Joe, H. (2006), 'Generation of Random Clusters with Specified Degree of Separation', *Journal of Classification* **23**(2), 315–334.
URL: *http ://link.springer.com/10.1007/s00357-006-0018-y* (Cité en page 58.)

Raftery, A. E. & Dean, N. (2006), 'Variable selection for model-based clustering', *Journal of the American Statistical Association* **101**(473), 168–178. (Cité en pages 72 et 149.)

Raghavan, U. N., Albert, R. & Kumara, S. (2007), 'Near linear time algorithm to detect community structures in large-scale networks', *Physical review E* **76**(3), 036106. (Cité en page 30.)

Raymaekers, J. & Zamar, R. H. (2020), 'Pooled variable scaling for cluster analysis', *Bioinformatics* **36**(12), 3849–3855. (Cité en pages 49, 50 et 58.)

Rechtien, A., Richert, L., Lorenzo, H., Martrus, G., Hejblum, B., Dahlke, C., Kasonta, R., Zinser, M., Stubbe, H., Matschl, U. et al. (2017), 'Systems vaccinology identifies an early innate immune signature as a correlate of antibody responses to the ebola vaccine rvsv-zebov', *Cell reports* **20**(9), 2251–2261. (Cité en page 13.)

Reust, M. J., Lee, M. H., Xiang, J., Zhang, W., Xu, D., Batson, T., Zhang, T., Downs, J. A. & Dupnik, K. M. (2018), 'Dried blood spot rna transcriptomes correlate with transcriptomes derived from whole blood rna', *The American Journal of Tropical Medicine and Hygiene* **98**(5), 1541. (Cité en page 13.)

Reuter, J. A., Spacek, D. V. & Snyder, M. P. (2015), 'High-throughput sequencing technologies', *Molecular cell* **58**(4), 586–597. (Cité en page 15.)

Rinchai, D., Deola, S., Zoppoli, G., Kabeer, B. S. A., Taleb, S., Pavlovski, I., Maacha, S., Gentilcore, G., Toufiq, M., Mathew, L. et al. (2022), 'High–temporal resolution profiling reveals distinct immune trajectories following the first and second doses of covid-19 mrna vaccines', *Science advances* **8**(45), eabp9961. (Cité en page 13.)

Risso, D. & Pagnotta, S. M. (2021), 'Per-sample standardization and asymmetric winsorization lead to accurate clustering of rna-seq expression profiles', *Bioinformatics* **37**(16), 2356–2364. (Cité en pages 44 et 47.)

Risso, D., Perraudeau, F., Gribkova, S., Dudoit, S. & Vert, J.-P. (2018), 'A general and flexible method for signal extraction from single-cell rna-seq data', *Nature communications* **9**(1), 284. (Cité en page 17.)

Robinson, M. D., McCarthy, D. J. & Smyth, G. K. (2010), 'edger : a bioconductor package for differential expression analysis of digital gene expression data', *bioinformatics* **26**(1), 139–140. (Cité en pages 17 et 36.)

Rouanet, A., Johnson, R., Strauss, M., Richardson, S., Tom, B. D., White, S. R. & Kirk, P. D. (2023), 'Bayesian profile regression for clustering analysis involving a longitudinal response and explanatory variables', *Journal of the Royal Statistical Society Series C : Applied Statistics* p. qlad097. (Cité en page 29.)

Sammut, C. & Webb, G. I. (2011), *Encyclopedia of machine learning*, Springer Science & Business Media. (Cité en page 23.)

Saxena, A., Dagur, P. K. & Biancotto, A. (2019), Multiparametric flow cytometry analysis of naïve, memory, and effector t cells, *in* 'Immunophenotyping', Springer, pp. 129–140. (Cité en page 92.)

Schaffer, C. M. & Green, P. E. (1996), 'An empirical comparison of variable standardization methods in cluster analysis', *Multivariate Behavioral Research* **31**(2), 149–167. (Cité en pages 48 et 50.)

Schelling, B. & Plant, C. (2020), 'Dataset-transformation : improving clustering by enhancing the structure with dipscaling and diptransformation', *Knowledge and Information Systems* **62**(2), 457–484. (Cité en page 73.)

Scrucca, L., Fop, M., Murphy, T. B. & Raftery, A. E. (2016), 'mclust 5 : clustering, classification and density estimation using Gaussian finite mixture models', *The R Journal* **8**(1), 289–317.
URL: *https ://doi.org/10.32614/RJ-2016-021* (Cité en page 84.)

Shaffer, J. P. (1995), 'Multiple hypothesis testing', *Annual review of psychology* **46**(1), 561–584. (Cité en page 36.)

Siffer, A., Fouque, P.-A., Termier, A. & Largouët, C. (2018), Are your data gathered ?, *in* 'Proceedings of the 24th acm sigkdd international conference on knowledge discovery & data mining', pp. 2210–2218. (Cité en pages 50, 73, 83 et 112.)

Silverman, B. W. (1981), 'Using kernel density estimates to investigate multimodality', *Journal of the Royal Statistical Society : Series B (Methodological)* **43**(1), 97–99. (Cité en page 73.)

Sneath, P. H. (1957), 'The application of computers to taxonomy', *Microbiology* **17**(1), 201–226. (Cité en page 26.)

Sokal, A., Chappert, P., Barba-Spaeth, G., Roeser, A., Fourati, S., Azzaoui, I., Vandenberghe, A., Fernandez, I., Meola, A., Bouvier-Alias, M. et al. (2021), 'Maturation and persistence of the anti-sars-cov-2 memory b cell response', *Cell* **184**(5), 1201–1213. (Cité en pages 13, 20 et 37.)

Song, D., Li, K., Ge, X. & Li, J. J. (2023), 'Clusterde : a post-clustering differential expression (de) method robust to false-positive inflation caused by double dipping', *Research Square* . (Cité en page 123.)

Steinley, D. (2004*a*), 'Properties of the hubert-arable adjusted rand index.', *Psychological methods* **9**(3), 386. (Cité en page 52.)

Steinley, D. (2004*b*), Standardizing variables in k-means clustering, *in* 'Classification, Clustering, and Data Mining Applications : Proceedings of the Meeting of the International Federation of Classification Societies (IFCS), Illinois Institute of Technology, Chicago, 15–18 July 2004', Springer, pp. 53–60. (Cité en pages 48, 49, 50, 51, 54 et 57.)

Suzuki, R. & Shimodaira, H. (2006), 'Pvclust : an r package for assessing the uncertainty in hierarchical clustering', *Bioinformatics* **22**(12), 1540–1542. (Cité en page 40.)

Tibshirani, R. J., Taylor, J., Lockhart, R. & Tibshirani, R. (2016), 'Exact post-selection inference for sequential regression procedures', *Journal of the American Statistical Association* **111**(514), 600–620. (Cité en pages 60 et 61.)

Tipping, M. E. & Bishop, C. M. (1999), 'Probabilistic principal component analysis', *Journal of the Royal Statistical Society Series B : Statistical Methodology* **61**(3), 611–622. (Cité en page 23.)

Townes, F. W. (2020), 'Review of probability distributions for modeling count data', *arXiv preprint arXiv :2001.04343* . (Cité en page 16.)

Townes, F. W., Hicks, S. C., Aryee, M. J. & Irizarry, R. A. (2019), 'Feature selection and dimension reduction for single-cell rna-seq based on a multinomial model', *Genome biology* **20**, 1–16. (Cité en pages 17, 23 et 37.)

Van der Maaten, L. & Hinton, G. (2008), 'Visualizing data using t-sne.', *Journal of machine learning research* **9**(11). (Cité en page 38.)

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A. et al. (2001), 'The sequence of the human genome', *science* **291**(5507), 1304–1351. (Cité en page 14.)

Vermunt, J. K. (2010), 'Latent class modeling with covariates : Two improved three-step approaches', *Political analysis* **18**(4), 450–469. (Cité en page 40.)

Vidman, L., Källberg, D. & Rydén, P. (2019), 'Cluster analysis on high dimensional rna-seq data with applications to cancer research-an evaluation study', *PLoS One* **14**(12), e0219102. (Cité en pages 45 et 58.)

Villani, A.-C., Satija, R., Reynolds, G., Sarkizova, S., Shekhar, K., Fletcher, J., Griesbeck, M., Butler, A., Zheng, S., Lazo, S. et al. (2017), 'Single-cell rna-seq reveals new types of human blood dendritic cells, monocytes, and progenitors', *Science* **356**(6335), eaah4573. (Cité en page 37.)

Vovk, V. & Wang, R. (2020), 'Combining p-values via averaging', *Biometrika* **107**(4), 791–808. (Cité en pages 78 et 146.)

Wagner, G. P., Kin, K. & Lynch, V. J. (2012), 'Measurement of mrna abundance using rna-seq data : Rpkm measure is inconsistent among samples', *Theory in biosciences* **131**, 281–285. (Cité en page 37.)

Waller, N. G., Kaiser, H. A., Illian, J. B. & Manry, M. (1998), 'A comparison of the classification capabilities of the 1-dimensional kohonen neural network with two pratitioning and three hierarchical cluster analysis algorithms', *Psychometrika* **63**, 5–22. (Cité en page 58.)

Wang, C., Gao, X. & Liu, J. (2020), 'Impact of data preprocessing on cell-type clustering based on single-cell rna-seq data', *BMC bioinformatics* **21**, 1–13. (Cité en pages 43 et 44.)

Ward Jr, J. H. (1963), 'Hierarchical grouping to optimize an objective function', *Journal of the American statistical association* **58**(301), 236–244. (Cité en pages 25 et 26.)

Wasserman, L., Azizyan, M. & Singh, A. (2014), 'Feature selection for high-dimensional clustering', *arXiv preprint arXiv :1406.2240* . (Cité en page 73.)

Weber, A. P. (2015), 'Discovering new biology through sequencing of rna', *Plant physiology* **169**(3), 1524–1531. (Cité en page 14.)

Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. R., Ozenberger, B. A., Ellrott, K., Shmulevich, I., Sander, C. & Stuart, J. M. (2013), 'The cancer genome atlas pan-cancer analysis project', *Nature genetics* **45**(10), 1113–1120. (Cité en page 51.)

Welch, B. L. (1947), 'The generalization of 'student's'problem when several different population varlances are involved', *Biometrika* **34**(1-2), 28–35. (Cité en pages 34 et 97.)

Wilk, A. J., Rustagi, A., Zhao, N. Q., Roque, J., Martínez-Colón, G. J., McKechnie, J. L., Ivison, G. T., Ranganath, T., Vergara, R., Hollis, T. et al. (2020), 'A single-cell atlas of the peripheral immune response in patients with severe covid-19', *Nature medicine* **26**(7), 1070–1076. (Cité en page 19.)

Wolfe, J. (1963), *Object Cluster Analysis of Social Areas*, University of California. **URL:** *https ://books.google.fr/books ?id=RFUdHwAACAAJ* (Cité en page 29.)

Yang, F., Foygel Barber, R., Jain, P. & Lafferty, J. (2016), 'Selective inference for group-sparse linear models', *Advances in Neural Information Processing Systems* **29**, 2469–2477. (Cité en page 76.)

Yun, Y.-J. & Foygel Barber, R. (2023), 'Selective inference for clustering with unknown variance', *Electronic Journal of Statistics* **17**(2), 1923–1946. (Cité en pages 100 et 123.)

Zhang, J. M., Kamath, G. M. & David, N. T. (2019), 'Valid post-clustering differential analysis for single-cell rna-seq', *Cell systems* **9**(4), 383–392. (Cité en pages 61, 62, 65, 66, 72 et 120.)

Zhang, J.-Y., Wang, X.-M., Xing, X., Xu, Z., Zhang, C., Song, J.-W., Fan, X., Xia, P., Fu, J.-L., Wang, S.-Y. et al. (2020), 'Single-cell landscape of immunological responses in patients with covid-19', *Nature immunology* **21**(9), 1107–1118. (Cité en page 20.)

Annexe A

Matériels Supplémentaires associés au Chapitre 4

A.1 Détails sur le caclul de la p-valeur du test d'inférence sélective

Pour calculer numériquement la p-valeur du test d'inférence sélective donnée en (4.4), nous utilisons en pratique une procédure de Monte-Carlo avec échantillonnage préférentiel, ce qui conduit à la p-valeur décrite en (4.7). Cependant, Phipson & Smyth (2010) ont montré que l'estimateur de Monte-Carlo classique d'une p-valeur pouvait être biaisé pour des p-valeurs proches de zéro. En effet, de très petites p-valeurs peuvent tomber exactement à 0 en utilisant l'approche de Monte-Carlo, ce qui conduit à des tests d'hypothèses statistiques qui ne contrôlent pas efficacement l'erreur de type I. Pour surmonter ce problème, Phipson & Smyth (2010) proposent alors de corriger la p-valeur de Monte-Carlo en ajoutant 1 à la fois au numérateur et au dénominateur de cette p-valeur estimée. Avec cette correction, au lieu d'avoir des p-valeurs à exactement 0, elles sont approchées par $\frac{1}{N+1}$ où N est le nombre d'échantillons de Monte-Carlo.

Malheureusement, nous avons montré à l'aide de simulations numériques que cette correction ne pouvait pas fonctionner avec la p-valeur du test d'inférence sélective décrite en (4.7) pour deux raisons. La première est que cette p-valeur provient à l'origine d'une probabilité conditionnelle, donc par définition, il y a en fait deux probabilités à calculer et à corriger. Le deuxième problème survient en raison de l'utilisation de l'échantillonnage préférentiel. En effet, parce que sous $\mathcal{H}_1$ les π^i en (4.7) sont très petits, ajouter 1 changera radicalement l'échelle de la $p-$valeur.

Ainsi, en raison de l'échantillonnage préférentiel, nous devons corriger notre p-valeur de Monte-Carlo en ajoutant une constante de l'ordre de π. Nous proposons alors d'ajouter $\overline{\pi} = \frac{1}{N}\sum_{i=1}^{N} \pi^i \mathbb{1}\left\{\widehat{C}_k, \widehat{C}_l \in c(\boldsymbol{X}(\omega_g^i))\right\}$ à la fois dans le numérateur et le dénominateur de (4.7). Cette correction est raisonnable car pour une petite p-valeur, c'est-à-dire sous $\mathcal{H}_1$, nous avons :

i) $|\boldsymbol{x}_g^t \boldsymbol{\eta}|$ est grand car $\widehat{C}_k$ et $\widehat{C}_l$ sont véritablement séparés sur $\boldsymbol{X}_g$

ii) $\sum_{i=1}^{N} \pi^i \mathbb{1}\{|\omega_g^i| > |\boldsymbol{x}_g^t \boldsymbol{\eta}|, \widehat{C}_k, \widehat{C}_l \in c(\boldsymbol{X}(\omega_g^i)\} \simeq 0$ car chaque $\pi^i = \frac{f_1(\omega_g^i)}{f_2(\omega_g^i)}$ où f_1 est la

densité d'une distribution gaussienne avec une moyenne de 0. Ainsi, parce que $\omega_g^i \sim \mathcal{N}(\boldsymbol{x}_g^t\boldsymbol{\eta}, \sigma^2\|\boldsymbol{\eta}\|_2)$ où $|\boldsymbol{x}_g^t\boldsymbol{\eta}|$ est grand, f_1 est évaluée en un point qui est loin de la moyenne, et c'est pourquoi $f_1(\omega_g^i) \simeq 0$.

Ainsi, en utilisant i) et ii) :

$$\begin{aligned}
\frac{\sum_{i=1}^{N} \pi^i \mathbb{1}\left\{|\omega_g^i| \geq |\boldsymbol{x}_g^t\boldsymbol{\eta}|, \widehat{C}_k, \widehat{C}_l \in c(\boldsymbol{X}(\omega_g^i))\right\} + \overline{\pi}}{\sum_{i=1}^{N} \pi^i \mathbb{1}\left\{\widehat{C}_k, \widehat{C}_l \in c(\boldsymbol{X}(\omega_g^i))\right\} + \overline{\pi}} &= \frac{0 + \overline{\pi}}{\sum_{i=1}^{N} \pi^i \mathbb{1}\left\{\widehat{C}_k, \widehat{C}_l \in c(\boldsymbol{X}(\omega_g^i))\right\} + \overline{\pi}} \\
&\simeq \frac{\frac{1}{N}\sum_{i=1}^{N} \pi^i \mathbb{1}\left\{\widehat{C}_k, \widehat{C}_l \in c(\boldsymbol{X}(\omega_g^i))\right\}}{\sum_{i=1}^{N} \pi^i \mathbb{1}\left\{\widehat{C}_k, \widehat{C}_l \in c(\boldsymbol{X}(\omega_g^i))\right\}\left[1 + \frac{1}{N}\right]} \\
&= \frac{\frac{1}{N}}{1 + \frac{1}{N}} \\
&= \frac{1}{N+1}
\end{aligned}$$

Ainsi, en corrigeant notre p-valeur de Monte-Carlo en ajoutant $\overline{\pi}$, nous obtenons l'estimateur proposé par Phipson & Smyth (2010) pour les petites p-valeurs.

A.2 Figure Supplémentaire 1

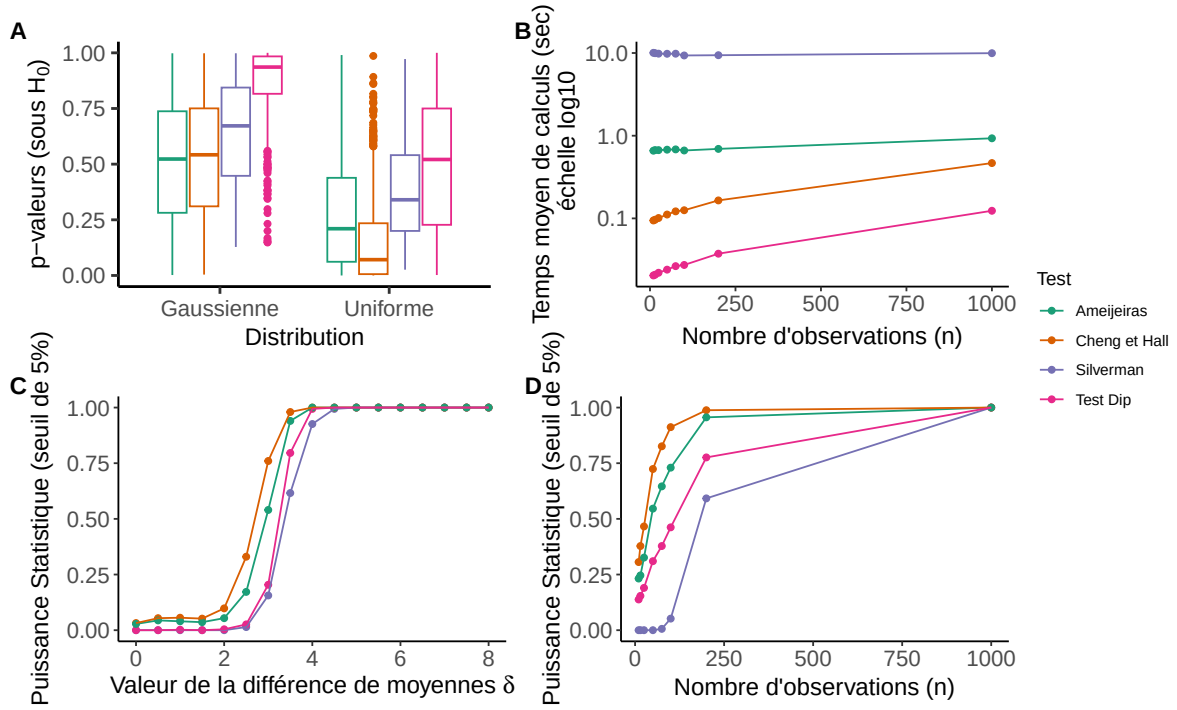


FIGURE A.1 – **Comparaison de différents tests de multimodalité implémentés dans le package** `R multimode`**(Ameijeiras-Alonso et al. 2021).** Panneau A : p-valeurs de chaque test de multimodalité sous l'hypothèse nulle pour 500 simulations de 200 réalisations d'une distribution gaussienne ou uniforme. Panneau B : Temps moyen de calcul requis par chaque test en fonction du nombre d'observations n (moyenne sur les 500 simulations). Panneau C : Puissance statistique (au seuil de significativité $\alpha = 5\%$) de chaque test de multimodalité en fonction de δ, la différence de moyennes entre deux modes d'un mélange gaussien à deux composants ($n = 200$ observations). Panneau D : Puissance statistique (au seuil de significativité $\alpha = 5\%$) de chaque test de multimodalité en fonction du nombre d'observations pour $\delta = 3.5$ fixé.

A.3 Figure Supplémentaire 2

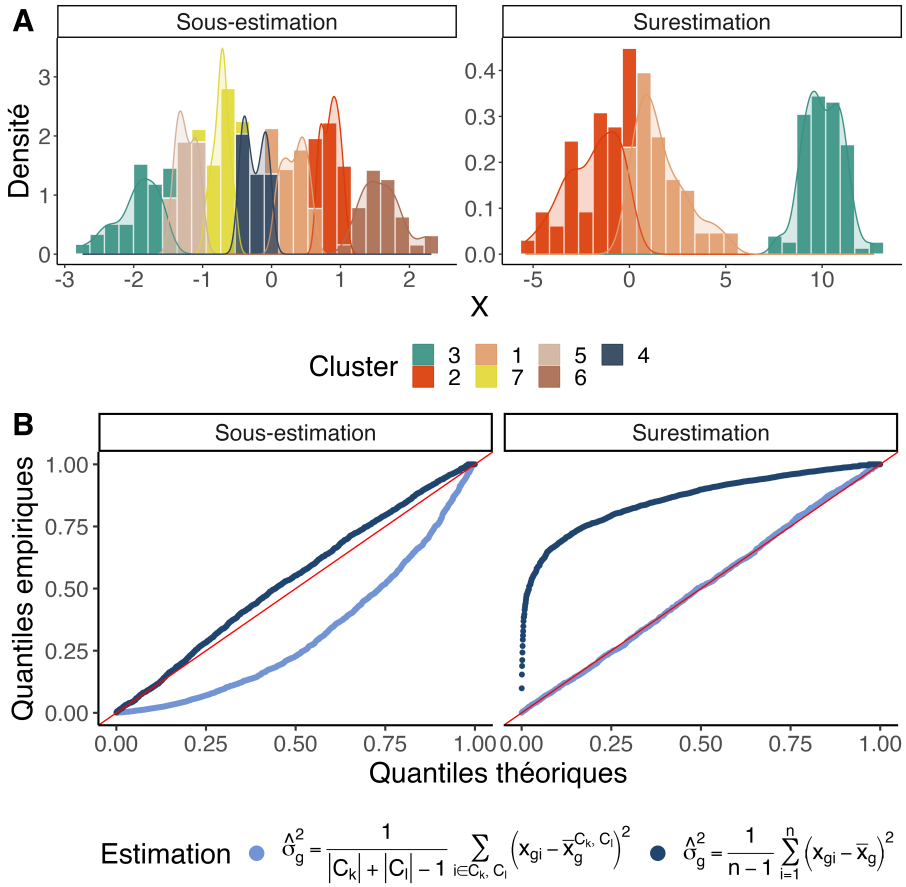


FIGURE A.2 – **Impact de l'estimation de la variance sur les p-valeurs du test d'inférence sélective.** Panneau A : Les données sont simulées selon un mélange gaussien univarié à deux composantes : $X \sim 0.5\mathcal{N}(0,4) + 0.5\mathcal{N}(10,1)$ pour le panneau de surestimation et selon une distribution gaussienne univariée $\mathcal{N}(0,1)$ pour le panneau de sous-estimation. Panneau B : QQ-plot des p-valeurs du test d'inférence sélective par rapport à la distribution uniforme selon l'estimateur de variance pour le test Cluster 1 vs. Cluster 2 pour 2000 simulations des données. La surestimation de la variance conduit à des p-valeurs conservatrices, tandis que la sous-estimation conduit à des faux positifs.

A.4 Figure Supplémentaire 3

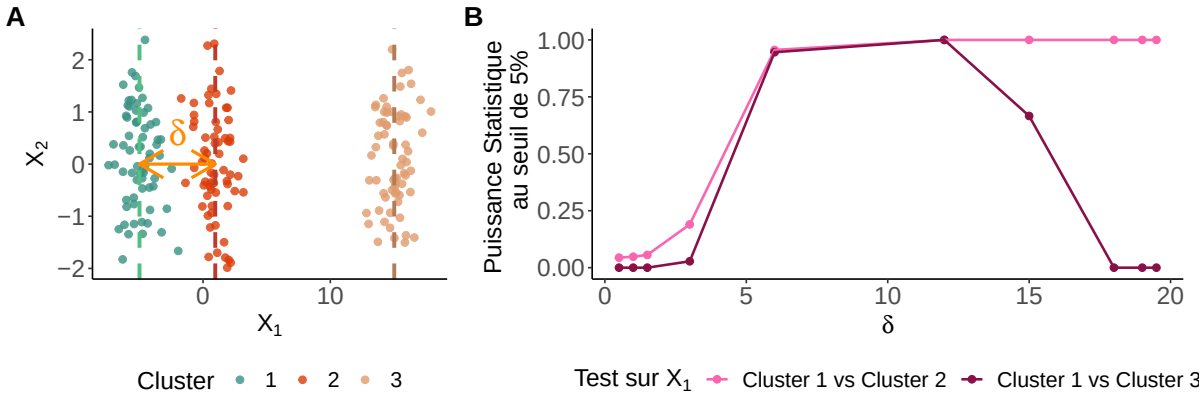


FIGURE A.3 – **Illustration de la possible perte de puissance statistique du test d'inférence sélective dans les cas où plus de deux clusters sont estimés.** Panneau A : Illustration du processus de génération des données. Un ensemble de données bivarié est simulé de telle sorte que trois clusters soient tous séparés uniquement sur X_1. Le Cluster 1 et le Cluster 2 sont séparés selon une différence de moyenne $\delta \in 0.5, 1, 1.5, 3, 6, 12, 15, 18, 19, 19.5$. Panneau B : Puissance statistique au seuil $\alpha = 5\%$ du test d'inférence sélective calculée à l'aide de 500 simulations des données telles que décrites dans le panneau A en fonction de δ. Pour chaque simulation, le test d'inférence sélective est appliqué pour tester la séparation du Cluster 1 vs le Cluster 2 et du Cluster 1 vs le Cluster 3 uniquement sur X_1.

A.5 Figure Supplémentaire 4

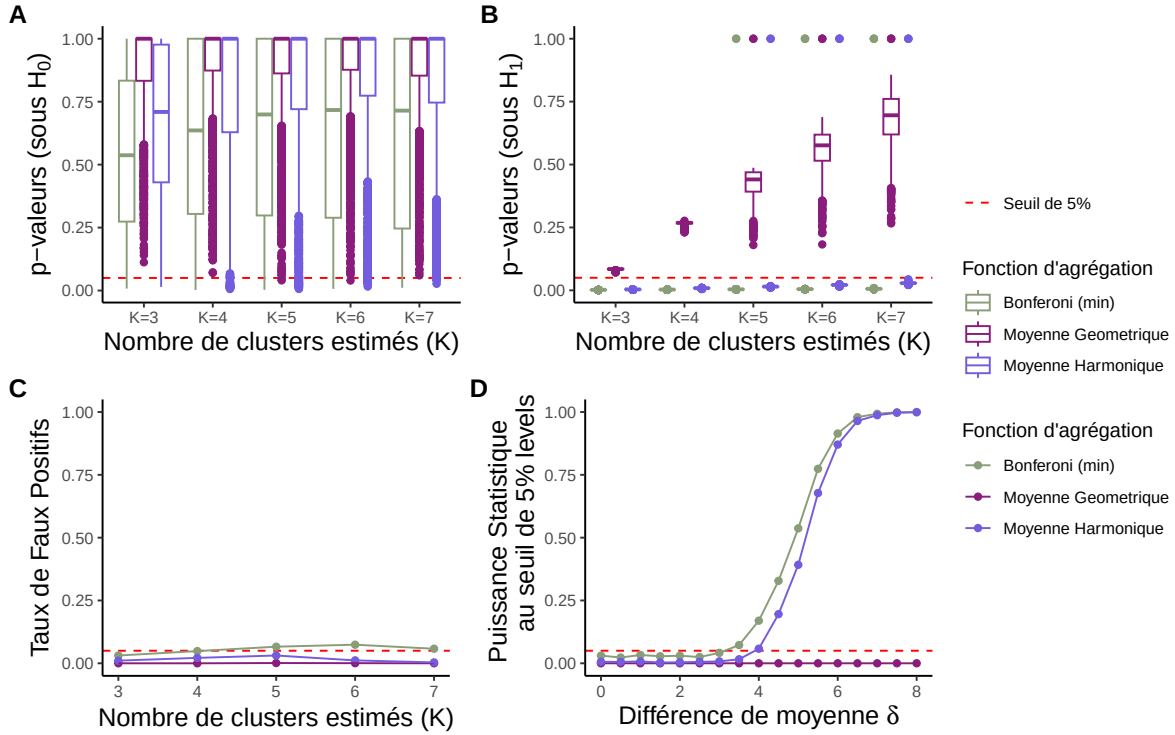


FIGURE A.4 – **Comparaison de trois différentes fonctions de d'agrégation de p-valeurs présentées dans Vovk & Wang (2020)**. Panneau A : Distribution des p-valeurs agrégées sous $\mathcal{H}_0$ en fonction du nombre de clusters estimés. Panneau B : Distribution des p-valeurs agrégées sous $\mathcal{H}_1$ (mélange gaussien univarié avec seulement deux composantes de proportion et de variances égales) en fonction du nombre de clusters estimés. Panneau C : Taux de faux positifs en fonction du nombre de clusters estimés. Panneau D : Puissance statistique en fonction de la différence de moyennes δ entre les deux modes du mélange où $K = 4$ clusters sont estimés (la même simulation que dans la figure 4.2). Le test d'inférence sélective est toujours appliqué aux clusters les plus extrêmes, de sorte que le nombre maximum de p-valeurs adjacentes soient agrégées. 2 000 simulations des données ont été utilisées.

A.6 Figure Supplémentaire 5

Nous avons étudié l'effet de la taille de l'échantillon à la fois sur l'erreur de type I et sur la puissance statistique de tous nos tests. Pour ce faire, nous avons utilisé le même schéma de simulation que celui présenté dans la figure 4.2. Nous avons simulé des données selon $n \in \{10, 15, 20, 25, 30, 50, 75, 100, 500\}$ réalisations d'un mélange gaussien univarié à deux composantes séparées par une différence moyenne $\delta = 0$ pour le cas sans cluster, et $\delta = 5$ pour le cas à deux clusters. Ensuite, nous avons appliqué la classification ascendante hiérarchique de Ward pour construire $K = 2$ ou $K = 4$ clusters et nous avons testé une séparation entre les deux clusters les plus extrêmes à l'aide de nos tests. Les résultats en termes d'erreur de type I ($\delta = 0$) et de puissance statistique ($\delta = 5$) estimés à l'aide de 1000 simulations des données sont présentés dans la figure A.5

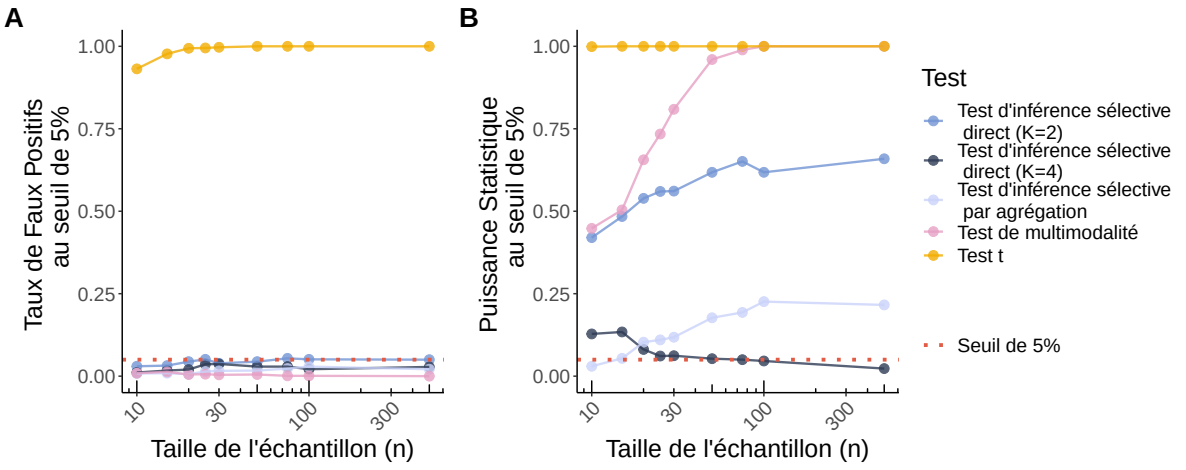


FIGURE A.5 – **Effets de la taille de l'échantillon sur le comportement de nos tests.** Panneau A : Erreur de type I ($\delta = 0$) de nos tests en fonction de la taille de l'échantillon. Panneau B : Puissance statistique ($\delta = 5$) de nos tests en fonction de la taille de l'échantillon.

Tous les tests contrôlent correctement l'erreur de type I au seuil désiré de 5% quel que soit la taille de l'échantillon, à l'exception du test t qui ne tient pas compte de l'étape de clustering (panneau A). La puissance statistique des tests (panneau B) est une fonction croissante de la taille de l'échantillon. Seul le test de multimodalité atteint une puissance statistique de 1 (pour $\delta = 5$) pour une taille d'échantillon modérée ($n = 75$). La puissance statistique du test d'inférence sélective (dans le cas $K = 2$) est limitée en raison du conditionnement sur l'événement de clustering dans sa définition : c'est le prix à payer pour un bon contrôle de l'erreur de type I sous l'hypothèse nulle. De plus, elle est plus impactée par la force de la séparation δ entre les clusters que par la taille de l'échantillon. Enfin, lorsque $K = 4$ clusters sont estimés, le test d'inférence sélective a une puissance statistique très faible (voir Figure supplémentaire A.3 pour des explications), mais l'agrégation des p-valeurs adjacentes permet d'augmenter significativement la puissance statistique dans

ce cas, mais au prix d'être moins puissant que le test d'inférence sélective direct dans le cas à $K = 2$ clusters.

A.7 Figure Supplémentaire 6

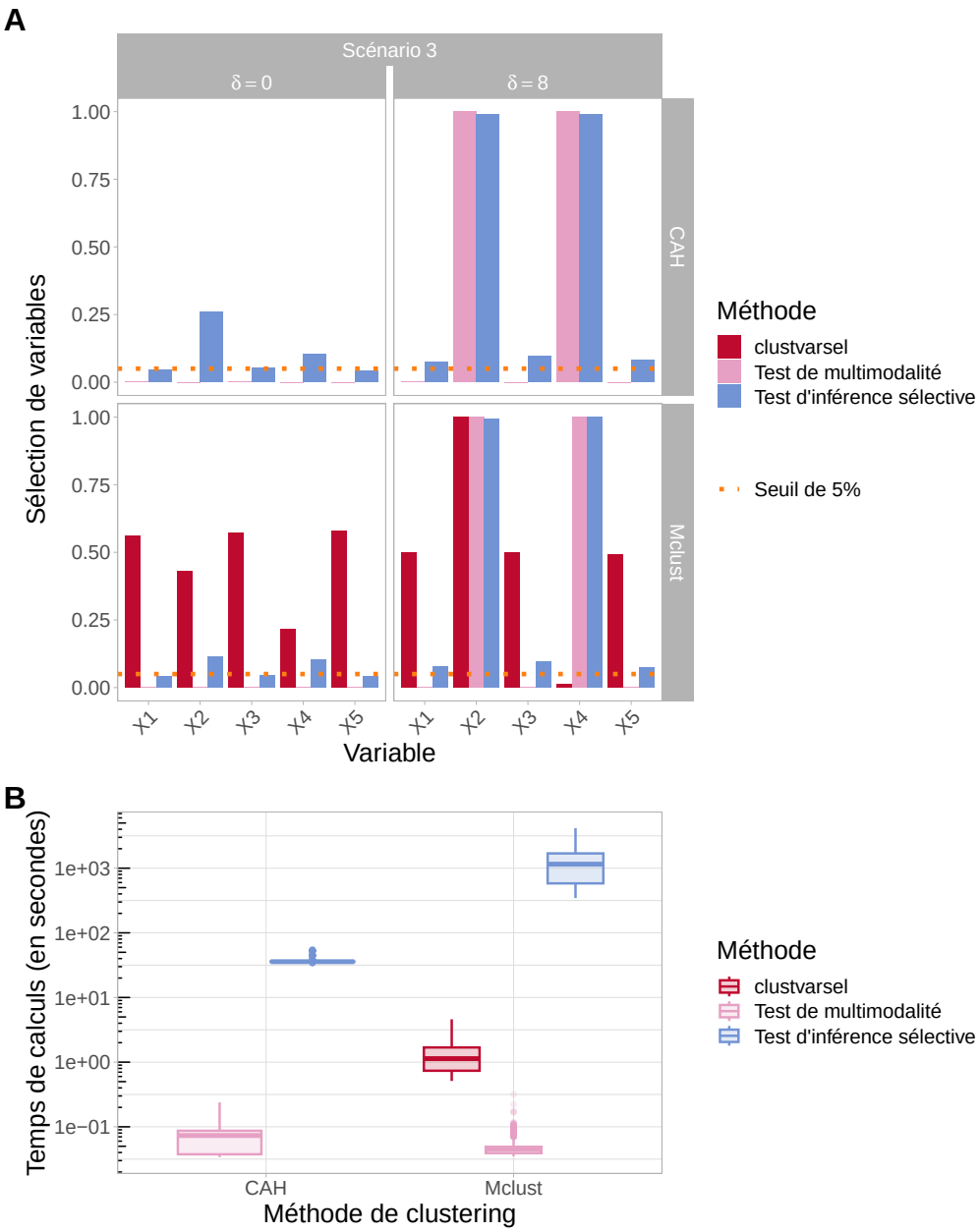


FIGURE A.6 – **Comparaison de nos tests post-clustering avec `clustvarsel` (Raftery & Dean 2006) pour la sélection de variables dans le cadre du clustering.** Le processus de génération de données reflète le Scénario 3 de la figure 4.3, mais sans corrélation entre X^1 et X^2 ($\rho = 0$). L'évaluation des performances est basée sur 1000 simulations des données. Panneau A : Comparaison des variables identifiées par toutes les méthodes. Nos tests présentent une sélection plus conservatrice, identifiant uniquement X^2 lorsque $\delta = 0$ en raison de sa corrélation avec X^4, et à la fois X^2 et X^4 lorsqu'il y a une séparation entre les deux composantes ($\delta = 8$). Panneau B : Distribution des temps de calculs pour chaque méthode de sélection de variables. Les temps de calculs des deux tests post-clustering dépendent de l'algorithme de clustering lui-même. Le test de multimodalité est le plus rapide et rivalise avec `clustvarsel` lorsqu'il est utilisé avec le clustering basé sur modèle gaussien.

A.8 Tableau Supplémentaire 1

Paire de clusters testée Variable testée	**Test d'inférence sélective (direct)**	**Test d'inférence sélective par agrégation**	**Test de multimodalité**	**Test** t
Cluster 1 vs Cluster 2				
longueur du bec	0.4082	0.4110	0.4899	0.0759
profondeur du bec	0.6478	0.6400	0.1478	0.4802
longueur de la nageoire	0.1160	0.1154	0.0992	0.0017*
masse corporelle	0.3321	0.3425	0.8320	0.0000*
Cluster 1 vs Cluster 3				
longueur du bec	0.1748	0.4995	0.6345	0.0001*
profondeur du bec	0.2914	0.3025	0.5242	0.0000*
longueur de la nageoire	0.3361	0.3206	0.6146	0.0005*
masse corporelle	0.3404	0.3868	0.2918	0.1190
Cluster 2 vs Cluster 3				
longueur du bec	0.2096	0.2120	0.9140	0.0041*
profondeur du bec	0.1867	0.6618	0.2376	0.0000*
longueur de la nageoire	0.2101	0.4322	0.1337	0.0000*
masse corporelle	0.1573	0.7967	0.6759	0.0000*

TABLE A.1 – **p-valeurs pour tous les tests entre une paire de clusters le long de chacune des 4 variables à partir des données réelles du contrôle négatif des manchots femmelle de l'espèce *Gentoo*.**
* met en évidence les p-valeurs significatives au seuil $\alpha = 5\%$

A.9 Tableau Supplémentaire 2

Paire de clusters	longueur du bec	profondeur du bec	longueur de la nageoire	masse corporelle
Cluster 1 vs Cluster 2	1,67	1,53	1,94	1,75
Cluster 1 vs Cluster 3	0,16	1,93	0,50	0,16
Cluster 2 vs Cluster 3	1,83	0,40	1,44	1,59

TABLE A.2 – **Valeurs de la différence de moyennes (δ) sur chaque variable ($z-$scores) entre chaque paire de clusters estimés.**

A.10 Figure Supplémentaire 7

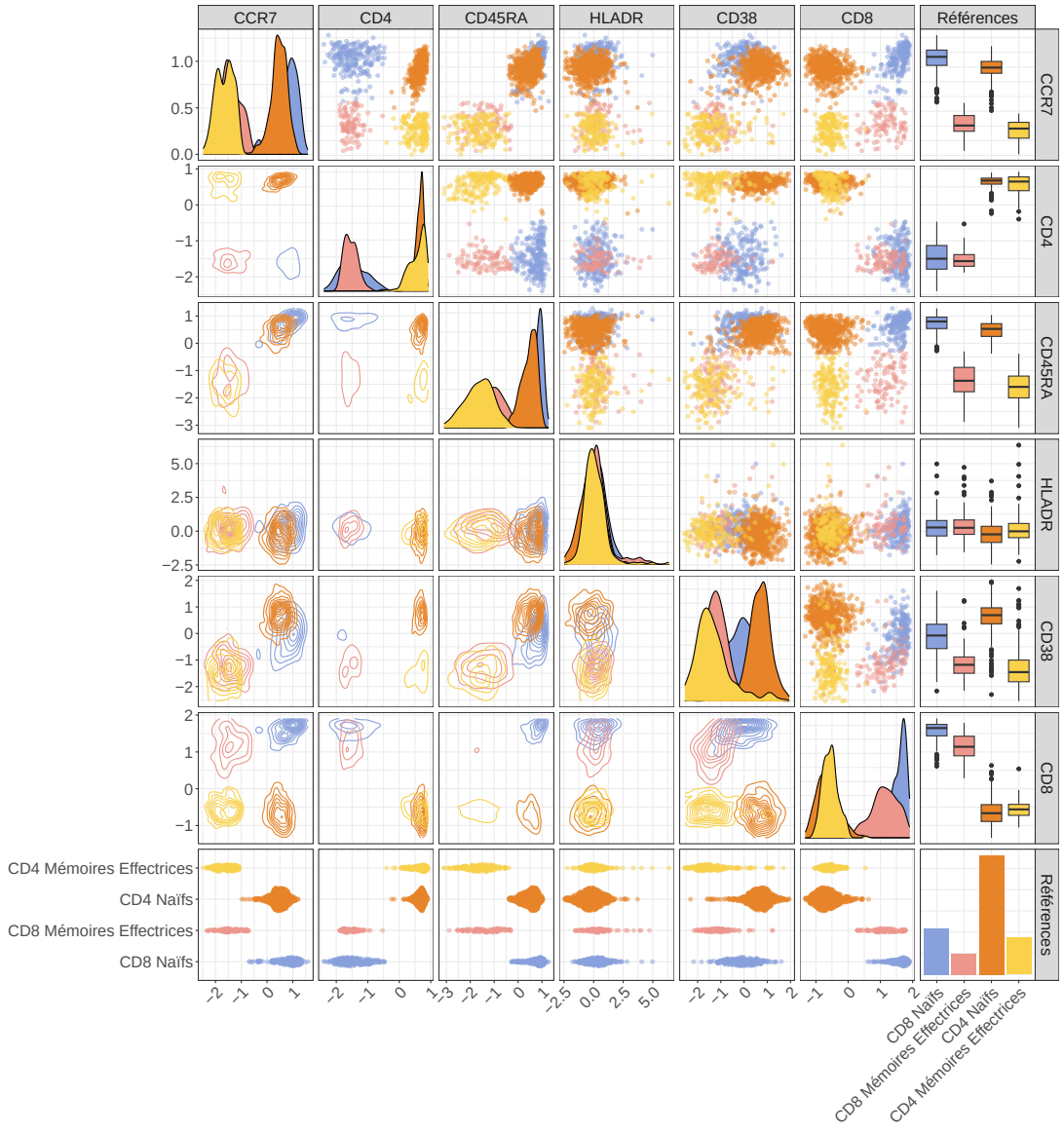


FIGURE A.7 – **Distributions univariées et bivariées des marqueurs cellulaires ($z-$scores) au sein des 4 sous-populations de lymphocytes T étudiées dans le jeu de données HIPC.**

A.11 Tableau Supplémentaire 3

Marqueurs	Test d'inférence sélective direct	Test d'inférence sélective par agrégation	Test de multimodalité
Cluster1 (100% **des CD8 Naïfs) vs Cluster2** (90% **des CD8 Mémoires Effectrices**)			
CCR7	5e-04*	0,001*	0,9048
CD4	0,2364	0,2364	0,6692
CD45RA	5e-04*	0,001*	0,9696
HLADR	0,1171	0,1171	0,9976
CD38	5e-04*	5e-04*	0,8524
CD8	5e-04*	5e-04*	0,9438
Cluster1 (100% des CD8 Naïfs) vs Cluster3 (100% des CD4 Naïfs)			
CCR7	5e-04*	5e-04*	0,9603
CD4	5e-04*	0,0045*	0*
CD45RA	5e-04*	5e-04*	0,9337
HLADR	5e-04*	0,004*	0,994
CD38	5e-04*	5e-04*	0,9777
CD8	5e-04*	0,003*	0*
Cluster1 (100% **des CD8 Naïfs) vs Cluster4** (99% **CD4 des Mémoires Effectrices**)			
CCR7	5e-04*	0,0026*	0*
CD4	5e-04*	0,0019*	0*
CD45RA	5e-04*	0,0033*	0,0869
HLADR	0,1006	0,1006	0,9596
CD38	5e-04*	0,0019*	0,9765
CD8	5e-04*	0,0013*	0*
Cluster2 (90% **des CD8 Mémoires Effectrices) vs Cluster3 (100% des CD4 Naïfs)**			
CCR7	5e-04*	5e-04*	0,6398
CD4	5e-04*	0,0019*	0*
CD45RA	5e-04*	5e-04*	0,8104
HLADR	0,014*	0,0173*	0,9945
CD38	5e-04*	0,001*	0,9915
CD8	5e-04*	0,0019*	0,7741
Cluster2 (90% **des CD4 Mémoires Effectrices) vs Cluster4** (99% **CD4 des Mémoires Effectrices**)			
CCR7	5e-04*	5e-04*	0,8611
CD4	5e-04*	5e-04*	0*
CD45RA	5e-04*	5e-04*	0,9973
HLADR	0,2042	0,2231	0,996
CD38	0,0013*	0,0013*	0,7312
CD8	5e-04*	5e-04*	0*
Cluster3 (100% **des CD4 Naïfs) vs Cluster4** (99% **des CD4 Mémoires Effectrices**)			
CCR7	5e-04*	0,0018*	0*
CD4	5e-04*	5e-04*	0,9991
CD45RA	5e-04*	0,0017*	0,0224*
HLADR	0,0011*	0,0011*	0,9939
CD38	5e-04*	0,0025*	0,9684
CD8	0,0032*	0,0032*	0,9574

TABLE A.3: **Toutes les comparaisons entre les clusters estimés à partir des données HIPC.**
$*$ met en évidence les p-valeurs significatives au seuil $\alpha = 5\%$

A.12 Figure Supplémentaire 8

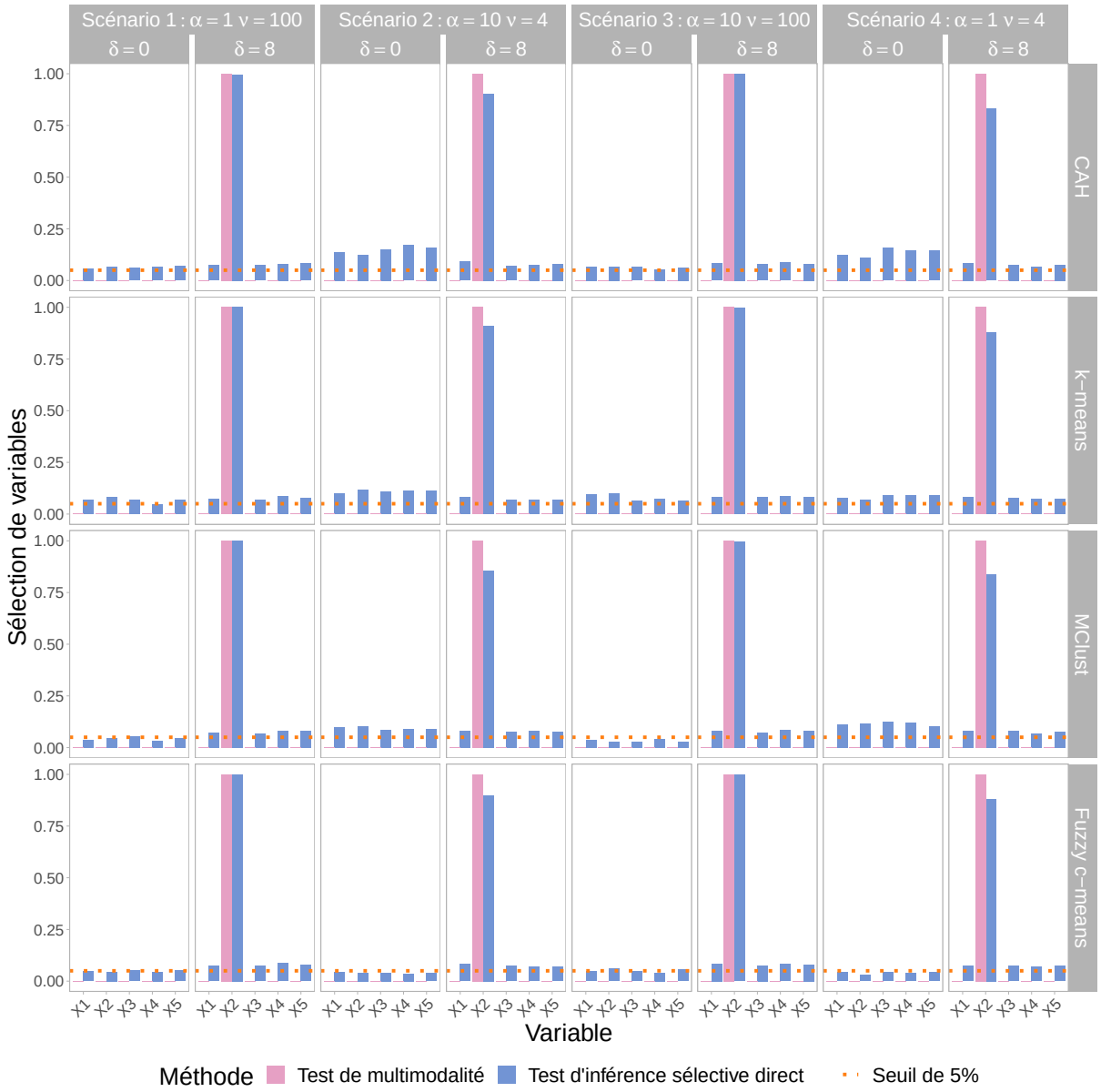


FIGURE A.8 – **Une forte asymétrie dans la distribution des données conduit à une inflation de l'erreur de type I.** Le processus de génération de données reflète le Scénario 1 dans la figure 4.3, où la distribution gaussienne a été remplacée par la distribution t asymétrique (Azzalini 2013). Différents degrés d'asymétrie (α) et différentes valeurs de degrés de liberté (ν) ont été explorés. Les performances sont moyennées sur 1 000 simulations des données sous ce scénario. S'éloigner de l'hypothèse gaussienne (grandes valeurs de α ou de faibles valeurs de ν) sous l'hypothèse nulle globale d'absence de clusters dans les données ($\delta = 0$) entraîne un mauvais contrôle de l'erreur de type I pour le test d'inférence sélective. Cependant, lorsque l'algorithme de clustering découvre avec succès la véritable structure des données ($\delta = 8$), l'erreur de type I pour les variables sous l'hypothèse nulle est bien calibrée. La performance du test de multimodalité reste inchangée en raison de sa nature non paramétrique.

A.13 Comportement des tests proposés dans le cadre multivarié

Pour conseiller le lecteur sur le choix du test d'inférence post-clustering en fonction de la nature des données, nous avons réalisé des études de simulation supplémentaires afin d'évaluer le comportement des trois tests proposés dans un cadre multivarié plus réaliste.

A.13.1 Les corrélations entre les variables peuvent empêcher un bon contrôle de l'erreur de type I

Nous avons d'abord étudié l'impact des corrélations entre les variables sur le comportement de nos tests. Nous avons simulé $n = 100$ réalisations d'une distribution normale multivariée $\mathcal{N}_p(\mathbf{0}_p, \mathbf{\Sigma}_{p\times p})$, telles que, pour $i, j \in \{1, \dots, p\}$:

$$\Sigma_{ij} = \begin{cases} 1 & \text{si } i = j \\ 0.5 & \text{si } i, j \leq \kappa \times p \text{ et } i \neq j \\ 0 & \text{sinon} \end{cases}$$

où $p = 10$ est le nombre de variables et $\kappa \in \{0, 0.5, 1\}$ est la proportion de variables corrélées dans les données. Sur de telles données, où aucun processus n'existe pour partitionner les observations en sous-groupes distincts, nous avons appliqué la classification ascendante hiérarchique de Ward pour estimer $K = 2$ et $K = 4$ clusters (le cas $K = 4$ étant nécessaires pour évaluer les performances du test d'inférence sélective par agrégation des p-valeurs adjacentes). Nous avons ensuite testé la séparation des deux clusters les plus extrêmes sur toutes les $p = 10$ variables à l'aide de nos 3 tests. Comme nous sommes sous l'hypothèse nulle pour chacune des variables, nous avons calculé pour chaque test le taux de faux positifs global au seuil nominal $\alpha = 5\%$. Les résultats pour 1 000 simulations des données sont donnés dans la figure A.9.

La figure A.9 met en évidence un bon contrôle du taux de faux positifs au seuil de 5% de tous les tests proposés dans le cadre de l'inférence post-clustering (contrairement au test t) en présence de données multivariées non corrélées. Le test de multimodalité est le seul test qui n'est pas affecté par la corrélation entre les variables, testant uniquement la séparation entre les clusters au niveau de la variable, indépendamment de la nature multivariée des données, et ignorant alors leurs corrélations. Cependant, le taux de faux positifs des tests d'inférence sélective est une fonction croissante de κ, la proportion de variables corrélées. En raison des corrélations entre les variables, les approches basées sur les perturbations peuvent échouer. Dans notre cas, cela est amplifié par le fait que les perturbations sont uniquement univariées : en raison de la corrélation, il est très peu probable que l'une des perturbations sur une dimension quelconque altère le clustering –

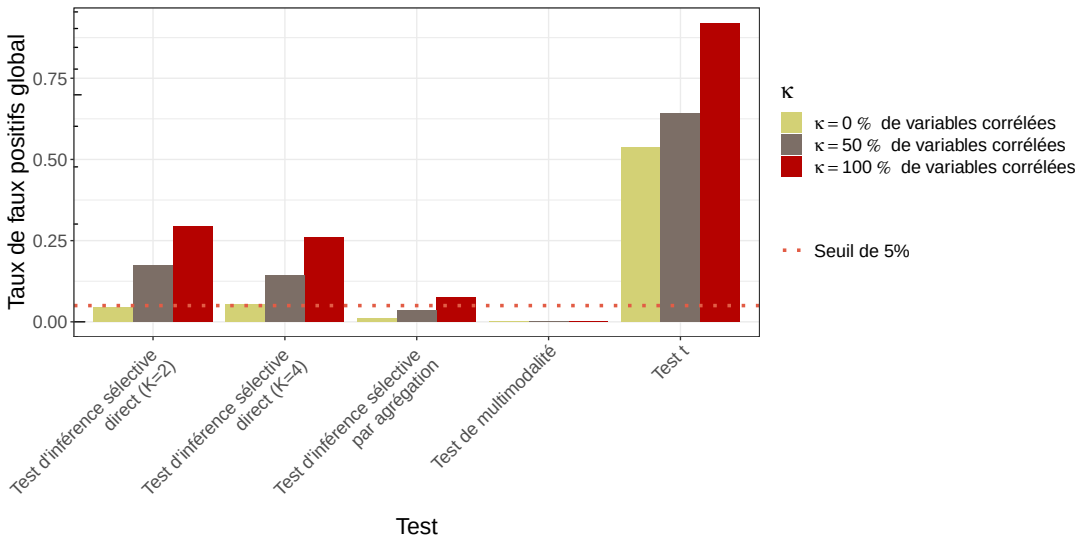


FIGURE A.9 – **Impact des variables corrélées sur le contrôle du taux de faux positifs des tests proposés.**

quelle que soit la qualité de cette structure de clustering par rapport aux données. Ainsi, les tests d'inférence sélective présentent de mauvaises performances dans ce cadre (étant exclusivement basés sur les perturbations).

A.13.2 Le nombre de variables a un impact relativement faible sur les performances des tests

Nous avons également étudié les effets potentiels du nombre de variables ainsi que de la force de la séparation entre les clusters δ sur les performances (à la fois en termes de contrôle du taux de faux positifs, de puissance statistique et de temps de calcul) des tests proposés. Nous avons simulé $n = 100$ réalisations d'un modèle de mélange gaussien à deux composantes $0.5\mathcal{N}_p\left(\mathbf{0}_p, \mathbf{\Sigma}_{p\times p}\right) + 0.5\mathcal{N}_p\left((\delta, \ldots, \delta), \mathbf{\Sigma}_{p\times p}\right)$ où $p \in \{4, 10\}$, $\delta \in \{0, 3, 5\}$ et pour éviter les corrélations non liées à la structure de groupe des données, nous avons fixé $\Sigma = \boldsymbol{I}_{p\times p}$.

La figure A.10 montre les résultats en termes de faux positifs ($\delta = 0$, panneau A) et de puissance statistique ($\delta \in \{3, 5\}$, panneau B) des 3 tests d'inférence post-clustering proposés, basés sur 1 000 simulations des données. Tous les tests post-clustering contrôlent efficacement le taux de faux positifs au seuil de 5%, quel que soit le nombre de variables. La puissance statistique du test d'inférence sélective augmente lorsque le nombre de variables augmente également, étant plus puissant que le test de multimodalité pour identifier des signaux faibles mais répétés ($\delta = 3$). La simplicité du test de multimodalité se fait au détriment d'une hypothèse nulle trop stricte : il nécessite une séparation nette entre les clusters sur la variable pour bien fonctionner, car il n'utilise que l'information au niveau de la variable et ne prend pas en compte l'ensemble de la structure des données.

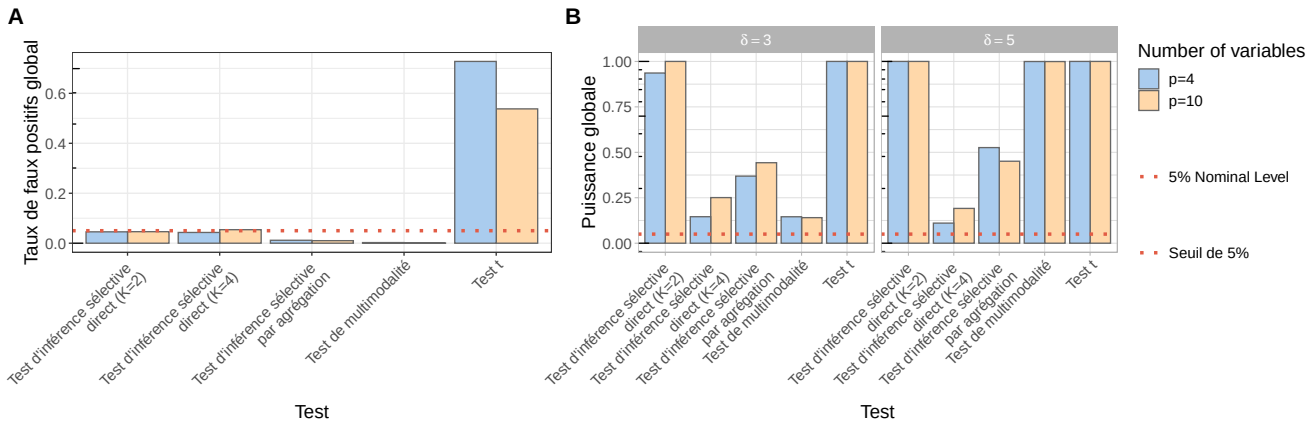


FIGURE A.10 – **Performances des tests proposés dans un cadre multivarié.** Panneau A : Taux de faux positifs des tests post-clustering. Panneau B : Puissance statistique des tests post-clustering en fonction de δ.

Au contraire, le test d'inférence sélective utilise l'information apportée par toutes les variables puisque le clustering est ré-appliqué à chaque version pertrubée des données en considérant l'ensemble des variables, ce qui facilite l'identification des signaux faibles au niveau de la variable (en augmentant le nombre de variables, la structure de regroupement est plus facile à préserver lors des étapes de perturbations au niveau des variables). Le test d'inférence sélective et le test de multimodalité restent tous deux efficaces pour identifier des signaux forts ($\delta = 5$) quel que soit le nombre de variables. Enfin, lorsque $K = 4$ clusters sont estimés, l'approche par agrégation des p-valeurs adjacentes parvient à corriger la perte de puissance statistique observée pour le test d'inférence sélective lorsque les deux clusters les plus extrêmes sont considérés.

Augmenter le nombre de variables semble aider le test d'inférence sélective à gagner en puissance, mais au prix d'une augmentation des temps de calcul. Cela est dû à la nécessité de ré-appliquer la méthode de clustering sur chaque version perturbée de l'ensemble des données, les temps de calcul nécessaires pour un test augmentent donc avec la dimension des données. Les temps de calcul du test par agrégation sont encore plus impactés par la dimensionnalité des données car il nécessite l'application du test d'inférence sélective sur chaque paire de clusters adjacents entre les deux clusters d'intérêt. Seul le test de multimodalité reste efficace sur le plan computationnel quel que soit le nombre de variables, car il n'utilise que l'information au niveau de la variable (figure A.11).

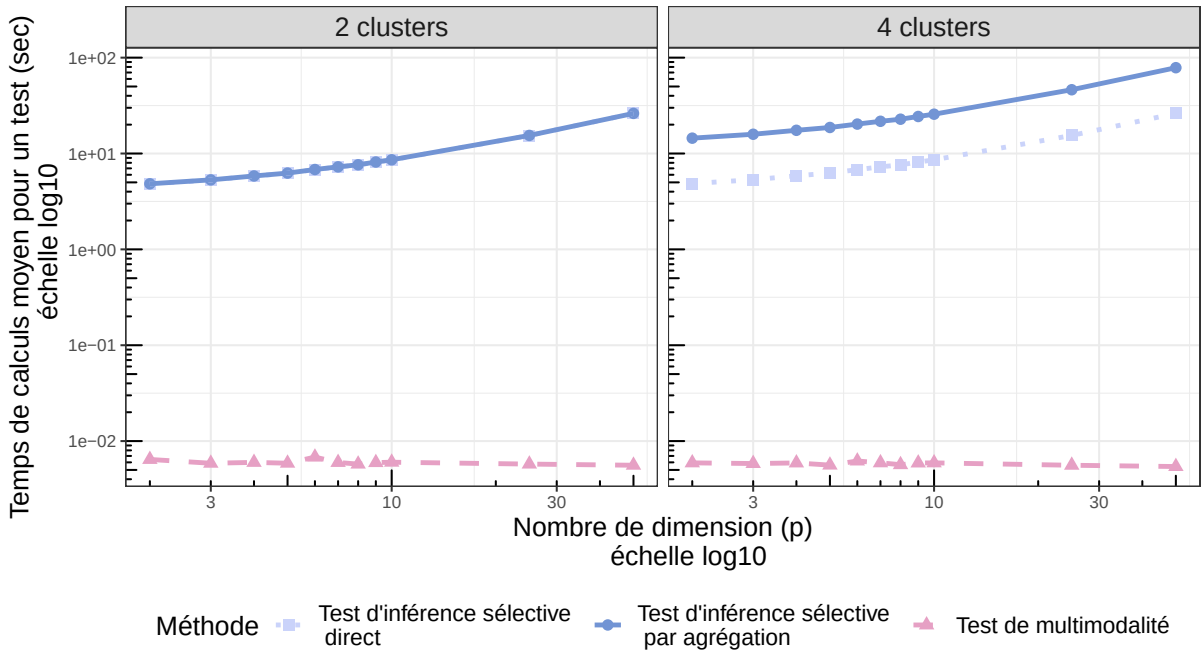


FIGURE A.11 – **Temps de calcul moyen des trois tests proposés (basé sur** 500 **simulations des données) en fonction du nombre de dimensions** (p) **des données**. Les tests sont effectués uniquement sur la première variable, donc la dimensionnalité des données n'affecte que les temps de calcul des tests d'inférence sélective puisque la méthode de clustering doit être ré-appliquée sur les données pour chaque perturbation en utilisant toutes les dimensions de $\boldsymbol{X}$.

Partenaires

Cette thèse a été soutenue en partie par le *Digital Public Health Graduate's school*, financé par le PIA 3 (Investissements d'Avenir – Référence du projet : 17-EURE-0019). Elle a également été soutenue par l'équipe associée DESTRIER d'Inria, dans le cadre du programme Inria@SiliconValley (code analytique : DRI-012215), et par le projet CARE financé par le *Innovative Medicines Initiative 2 Joint Undertaking* (JU) en vertu de l'accord de subvention n° IMI2-101005077. Le JU reçoit le soutien du programme de recherche et d'innovation Horizon 2020 de l'Union européenne, de l'EFPIA, de la Fondation Bill & Melinda Gates, du *Global Health Drug Discovery Institute* et de l'Université de Dundee.

Cette thèse a reçu un financement du programme de recherche et d'innovation Horizon 2020 de l'Union européenne dans le cadre de l'accord de subvention EHVA n° H2020-681032. Cette thèse a été réalisée dans le cadre du programme France 2030 / RRI PHDS de l'Université de Bordeaux. Le temps de calcul pour cette thèse a été fourni par les installations de calcul du MCIA (Mésocentre de Calcul Intensif Aquitain) de l'Université de Bordeaux et de l'Université de Pau et des Pays de l'Adour.

Résumé/Abstract

Clustering et analyse différentielle de données d'expression génique

Résumé : Les analyses des données d'expression génique issues du séquençage de l'ARN (RNA-seq) en masse (bulk RNA-seq) ou en cellule unique (scRNA-seq) sont devenues courantes dans les études immunologiques. Elles permettent entre autres une meilleure compréhension de l'hétérogénéité présente dans les réponses immunitaires, qu'elles soient en réponse à la vaccination ou face à des maladies. Les analyses de ces données se font souvent selon deux étapes : i) d'abord une classification non supervisée, ou clustering, utilisant l'ensemble des gènes pour regrouper les échantillons en sous-groupes distincts et homogènes ; ensuite ii) l'analyse différentielle se faisant à l'aide de tests d'hypothèse visant à identifier les gènes qui sont différentiellement exprimés entre ces sous-groupes. Cependant, ces deux étapes successives soulèvent un problème méthodologique actuellement souvent ignoré dans la littérature appliquée. En effet, les méthodes traditionnelles d'inférence nécessitent des hypothèses de tests fixées a priori, sans dépendre des données, pour garantir un contrôle effectif de l'erreur de type I. Dans le contexte de ces analyses en deux étapes, les hypothèses de tests sont basées sur les résultats du clustering ce qui compromet le contrôle de l'erreur de type I des méthodes traditionnelles qui peuvent alors conduire à de fausses découvertes. Nous proposons alors de nouvelles méthodes statistiques qui permettent de tenir compte de cette double utilisation des données, garantissant un contrôle effectif du nombre de fausses découvertes.

Mots clés : Clustering, analyse différentielle, inférence sélective, RNA-seq

---

Clustering and differential analysis of gene expression data

Abstract : Analyses of gene expression data obtained from bulk RNA sequencing (bulk RNA-seq) or single-cell RNA sequencing (scRNA-seq) have become commonplace in immunological studies. They allow for a better understanding of the heterogeneity present in immune responses, whether in reaction to vaccination or disease. Typically, the analysis of these data is conducted in two steps : i) first, an unsupervised classification, or clustering, is performed using all the genes to group samples into distinct and homogeneous subgroups ; ii) then, differential analysis is conducted using hypothesis tests to identify genes that are differentially expressed between these subgroups. However, these two successive steps lead to methodological challenge that is often overlooked in the applied literature. Traditional inference methods require hypothesis to be fixed a priori and independent of the data to ensure effective control of type I error. In the context of these two-steps analyses, the hypothesis tests are based on the results of the clustering, which compromises the control of type I error by traditional methods and can lead to false discoveries. We propose new statistical methods that account for this double use of the data and ensure an effective control of the number of false discoveries.

Key words : Clustering, differential analysis, selective inference, RNA-seq

---

Discipline : Santé Publique - option Biostatistiques
Laboratoire : INSERM U1219, Bordeaux Population Health Center - Inria - Université de Bordeaux, 146 rue Léo Saignat 33076 Bordeaux