Impact de la dépendance dans les procédures de tests multiples en grande dimension

Chloé Friguet

N° ordre : 2010-21
N° Série : G-7

THESE / AGROCAMPUS OUEST

Sous le sceau de l'Université Européenne de Bretagne

pour obtenir le diplôme de :

DOCTEUR DE L'INSTITUT SUPERIEUR DES SCIENCES AGRONOMIQUES, AGRO-ALIMENTAIRES, HORTICOLES ET DU PAYSAGE

Spécialité : Mathématiques appliquées

École Doctorale : MATISSE

présentée par :

Chloé FRIGUET

IMPACT DE LA DÉPENDANCE DANS LES PROCÉDURES DE TESTS MULTIPLES EN GRANDE DIMENSION

soutenue le 24 septembre 2010 devant la commission d'Examen

Composition du jury :

Christophe AMBROISE	Président
Stéphane ROBIN	Rapporteur
John STOREY	Rapporteur
David CAUSEUR	Directeur de thèse
Anne SIEGEL	Examinateur

IRMAR

Résumé Motivé par des applications dans le domaine de l'analyse de données génomiques, ce travail de thèse porte sur l'étude de l'impact de la dépendance sur les propriétés des procédures de tests multiples en grande dimension. Notre proposition consiste à considérer un modèle d'Analyse en Facteurs pour la structure de dépendance entre les variables. Un algorithme de type EM est présenté pour estimer les paramètres du modèle ainsi qu'une méthode *ad hoc* pour déterminer le nombre optimal de facteurs à inclure dans le modèle.

De plus, ce modèle définit un cadre général pour la prise en compte de la dépendance dans les procédures de tests multiples. L'estimation du taux de faux-positifs (FDR) et de la proportion d'hypothèses nulles (π_0), paramètre important qui intervient dans le contrôle des taux d'erreurs, sont étudiés plus particulièrement. Ainsi, on montre que la dépendance entre tests entraîne une instabilité des procédures d'inférence simultanée. Une nouvelle approche est présentée : l'objectif est de réduire cette dépendance, procurant à la fois une augmentation de la puissance des tests et une diminution de la variabilité des taux d'erreurs.

Enfin, ces résultats méthodologiques sont illustrés à partir de données génomiques et la procédure est implémentée dans le logiciel libre `R` au sein du package `FAMT`.

Mots clés *Tests multiples, Dépendance, Analyse en Facteurs, Proportion d'hypothèses nulles, FDR, Package* ***R FAMT***

Abstract Motivated by issues raised by the analysis of gene expressions data, this thesis focuses on the impact of dependence on the properties of multiple testing procedures for high-dimensional data. We propose a methodology based on a Factor Analysis model for the correlation structure. Model parameters are estimated thanks to an EM algorithm and an *ad hoc* methodology allowing to determine the model that fits best the covariance structure is defined.

Moreover, the factor structure provides a general framework to deal with dependence in multiple testing. Two main issues are more particularly considered : the estimation of π_0, the proportion of true null hypotheses, and the control of error rates. The proposed framework leads to less variability in the estimation of both π_0 and the number of false-positives. Consequently, it shows large improvements of power and stability of simultaneous inference with respect to existing multiple testing procedures.

These results are illustrated by real data from microarray experiments and the proposed methodology is implemented in a `R` package called `FAMT`.

Key words *Multiple testing, Dependence, Factor Analysis, Proportion of null hypotheses, FDR,* ***R*** *package* ***FAMT***

REMERCIEMENTS

Comme le veut la tradition, je vais ici me satisfaire à l'exercice de la page des remerciements, une des premières du manuscrit dans la numérotation, mais néanmoins la dernière écrite. Entreprendre la rédaction de cette partie est donc un moment émouvant... Cela signifie que la fin de quelque chose est proche, mais surtout que le début de quelque chose d'autre approche !

Mes premiers remerciements sont naturellement pour David Causeur, mon cher directeur de thèse. Je le remercie très sincèrement pour la confiance qu'il m'a accordée au cours de ces années. J'ai beaucoup apprécié travailler sous sa direction, pendant ces trois années de thèse et dans la continuité de mon stage de M2. J'espère sincèrement que à l'avenir nos collaborations seront nombreuses. Ses qualités humaines et scientifiques et ses précieux conseils m'ont permis de travailler dans de bonnes conditions. Nos nombreuses discussions m'ont permis de progresser, et d'une manière plus générale de mieux appréhender les différentes facettes du métier d'enseignant-chercheur. David, merci pour tout cela, et pour tout le reste.

Je remercie ensuite les rapporteurs de cette thèse Stéphane Robin et John Storey pour l'intérêt qu'ils ont porté à mon travail en acceptant de faire partie du jury, mais surtout pour leurs conseils avisés et suggestions pertinentes qui ont permis l'amélioration de ce manuscrit. Merci également aux autres membres du jury, Anne Siegel et Christophe Ambroise, qui ont accepté de juger ce travail. Je suis très touchée de l'honneur que me font l'ensemble des membres du jury d'y participer.

Je ne peux écrire cette page sans mentionner le cadre dans lequel cette thèse a vu le jour. Travailler à Agrocampus, et en particulier au sein du laboratoire de Mathématiques Appliquées, est réellement agréable. L'équipe y est pour beaucoup : David, bien sûr, Jérôme Pagès, directeur du laboratoire, François Husson et Sébastien Lê, maîtres de conférence, Julie Josse, Magalie Houée-Bigot et Gwenaelle Fournier, ingénieures, et enfin Elisabeth Lenauld, Aline Legrand et Karine Bagory, secrétaires. Un entourage scientifique de qualité, et une ambiance très conviviale : je souhaite sincèrement à tout thésard de pouvoir bénéficier d'un tel environnement !

J'ai (volontairement) omis trois personnes de cette équipe que je souhaite maintenant remercier plus particulièrement. Tout d'abord, merci à Maela Kloareg, maître de conférence, d'avoir conforter mon goût pour la statistique appliquée tout au long de ma thèse et pour son efficace *coaching* lors de mes

premiers pas dans le monde de l'enseignement. *Mersi bras* ! Merci également à Yuna Blum, doctorante en statistiques et en génétique, avec qui j'apprécie travailler tout particulièrement. Et puis Marine Cadoret, ma fort sympathique collègue de bureau ! On a commencé nos thèses respectives ensemble, on les termine ensemble : il y a pas mal de souvenirs pour ces trois années, d'autant plus que les journées sont parfois très longues en thèse !

Je remercie également Sandrine Lagarrigue, professeur au laboratoire de Génétique Animale d'Agrocampus/INRA, pour les données et la validation biologique de notre approche, avec Yuna. Je poursuis ces remerciements par un petit mot pour les collègues statisticiens de Rennes2 : Mathieu Emily, Arnaud Guyader (pour tous les bons moments de la SFdS en particulier), Eric Matzner, Laurent Rouvière, Nicolas Jégou et Magalie Fromont pour la bonne expérience des TD à l'ENSAI ;

Ma thèse a été l'occasion de nombreuses rencontres avec des doctorants (certains sont devenus docteurs depuis !) de divers disciplines, en particulier à travers l'association DocAIR : Marina, Marie-Laure, Luc, Hélène B., Francine, Lucie, Marion, Séb, Hélène E., Didier, Bertrand et Thierry ; l'équipe d'organisation du festival *Sciences en Cour[t]s* : Marine, Hélène, Véro, Didier, Benjamin & Aymeric, et tous les doctorants-réalisateurs des Très Courts-Métrages 2010, qui nous ont permis de faire cette année encore un joli festival (tcm-rennes.org) ; les doctorants de Nicomaque : Valentin, Laurent, Aymeric ; et finalement les doctorants de l'IRMAR que j'ai croisé parfois : Ludo et Victor, parce que quelque part, on a fait partie de la même équipe ! !

Par ailleurs, j'ai pu présenter mes travaux de thèse dans de nombreux congrès, d'Ottawa à Ascona, en passant par St Petersbourg, Bordeaux, Paris ou encore Brest. Merci à David et Jérôme de m'avoir permis de saisir ces opportunités de valoriser mes travaux de recherche et d'y faire de nombreuses rencontres, car je sais que ce n'est pas offert à tous les doctorants ailleurs. Une thèse c'est un projet scientifique mené au sein d'un laboratoire de recherche, mais qui comporte également quelques aspects logistiques et administratifs : j'ajoute ici un remerciement à Hervé LeBris et Françoise Pringent, de la Coordination des Écoles Doctorales à Agrocampus, et à Olivier Bonnaud et Élodie Cottrel de l'École Doctorale MATISSE. Je suis également reconnaissante envers la Région Bretagne qui a financé cette thèse pendant 3ans.

Je termine ces remerciements par une note plus personnelle pour ma famille, en Bretagne bien-sûr, sans oublier la Drôme, la Vendée, Paris, Berlin et la Nouvelle-Calédonie ; et pour mes amis, Bobo, Dédé, Fofie, Nanard et l'ACM en général pour leur amitié depuis des années ; Lolo, Pierre M., Antoine, Jojo, Delphine, Alex, Julien, Geoffroy, Jérem', Pierre T., Fafa, Thomas M., Wilfried, Bénouze, Clervie, Romain mon fillot et surtout ma Marie chérie, pour les moments très *drillant* à l'IUP GIS de Vannes ; Delphine, Thierry, Pierre T., Thomas L. qu'on n'oublie pas, Raymond, Cléo et Philippe de la spé stat à Agrocampus ; Alain, Giz, Pascal, Sandy, Sylvain, Laurent, Blandine et Clairette pour tous ces jours heureux à la MG.

Et enfin, Laurent, pour m'avoir soutenue avec patience (euh..) et amour (là, oui, toujours !) car j'en avais bien besoin ces derniers mois, mais surtout pour tout ce qu'il y a de meilleur dans l'avenir...

TABLE DES MATIÈRES

TABLE DES FIGURES

LISTE DES TABLEAUX

INTRODUCTION

Dans le prolongement de la très éprouvée théorie des tests d'hypothèses, les problèmes posés par le test simultané de plusieurs hypothèses, ou plus généralement l'inférence simultanée, font l'objet de discussions récurrentes dans la littérature statistique depuis très longtemps. Dès les années 1930, les premières procédures de tests multiples sont proposées par Fisher pour les tests simultanés de plusieurs contrastes dans le cadre du modèle linéaire d'analyse de la variance. Elles sont fortement imprégnées des grands principes de la théorie des tests d'hypothèses, formalisée au début du XXème siècle par les contributions de Fischer, Student, Neymann, K. et E. Pearson, mais déjà initiée, dès la fin du IXXème siècle, par les travaux de Laplace, DeMoivre et Bernoulli sur la maîtrise des erreurs et de l'aléatoire, en particulier en astronomie (pour plus de détails historiques, voir Salsburg [2002]).

Cette théorie statistique de la décision introduit une dissymétrie dans les deux issues possibles du test, l'hypothèse nulle H_0 et l'hypothèse dite alternative H_1. L'objectif d'une procédure de test est alors le contrôle du risque de rejeter à tort H_0 (risque de type-I). L'extension au cas de tests multiples vise alors naturellement au contrôle du risque de rejeter H_0 au moins une fois à tort. Dans le cadre des tests paramétriques univariés, cette approche établit d'ailleurs une unité entre la théorie des tests et celle de l'estimation puisque le calcul de la région de rejet de H_0 apparaît comme transposable à celui de l'intervalle de confiance sur le paramètre testé. La transposition aux tests multiples établit, là aussi, une équivalence avec la construction d'un intervalle de confiance simultané des paramètres testés.

La théorie des tests d'hypothèses individuels identifie des solutions optimales, au sens de procédures de puissance maximale parmi celles contrôlant le risque de première espèce. Néanmoins, aucune notion universelle d'optimalité d'une procédure de tests multiples ne s'impose, laissant ainsi ouvertes les questions relatives à la recherche de la meilleure procédure [Shaffer, 1995, Dudoit et al., 2003]. La multiplicité est étudiée effectivement comme un problème à part entière dans les tests simultanés en particulier par Duncan, Dunnett, Scheffé ou encore Tukey, qui ont laissé leurs noms à des méthodes de

tests post-hoc en analyse de variance. L'utilisation préférentielle de l'une ou l'autre de ces méthodes dépend essentiellement de la problématique et donc du contexte d'application : en biologie, où l'on étudie l'effet d'un traitement sur plusieurs variables d'intérêt, en épidémiologie, où l'on étudie l'effet de la dose de médicament sur des mesures réalisées à différentes étapes des essais cliniques, en agro-alimentaire, où des analyses sensorielles de différents produits sont réalisées à partir de plusieurs descripteurs.

Les procédures de tests multiples reposent toutes sur un même principe de choix d'un seuil sur les probabilités critiques associées aux tests individuels. Elles se différencient par le mode de calcul de ce seuil. Les études comparatives se concentrent alors principalement sur le contrôle d'un risque de type-I établi à l'échelle de l'ensemble des tests, en l'occurrence le risque de rejeter au moins une fois l'hypothèse nulle à tort (appelé Family-Wise Error Rate, FWER). Dans le contexte où le nombre de tests est modéré, le contrôle du risque de type-I focalise toutes les attentions. Les questions relatives à la puissance des procédures de tests multiples sont peu abordées. Or, de manière générale, les nombreuses procédures proposées [Bonferroni, 1936, Sidak, 1967, Holm, 1979] pour assurer le contrôle du FWER conduisent, lorsque le nombre de tests augmente, à des procédures très conservatrices. Par conséquent, le risque de non-détection de l'hypothèse H_1 est alors important.

Par ailleurs, l'hypothèse d'indépendance sur laquelle repose la plupart des procédures de test multiples fait rarement l'objet de discussions. Cela peut s'expliquer par le fait que dans le cadre des tests post-hoc en analyse de la variance, cette dépendance résulte exclusivement du dispositif expérimental et des contrastes testés. On peut penser que son impact est le même sur chacun des tests, si ce dispositif est équilibré. D'autre part, les approches générales de prise en compte de la dépendance induisent des problèmes numériques qui limitent voire rendent impossibles leur implémentation pratique.

Tests multiples en grande dimension Des développements méthodologiques innovants ont vu le jour au cours des deux dernières décennies, pour faire face à de nouveaux enjeux dans des domaines scientifiques. Les évolutions technologiques ont conduit à la production de grands volumes de données. De manière générale, ces technologies dites "à haut-débit" se sont développées pour tendre vers une analyse aussi globale que possible d'un système complexe, tel le cerveau humain exploré à l'aide de l'imagerie médicale par résonance magnétique fonctionnelle (IRMf), un système de particules élémentaires en astrophysique, les mouvements de marchés par les flux de transactions commerciales ou encore en biologie fonctionnelle à travers le séquençage de l'ADN.

Comprendre, analyser et prévoir le fonctionnement de systèmes complexes nécessitent de prendre en compte l'hétérogénéité et le grand volume des données résultant de ces technologies. Le plus souvent, ce grand volume des données se traduit par un nombre potentiel de variables de plusieurs milliers, observées sur un petit nombre d'individus statistiques. On parle généralement de données de grande dimension pour caractériser cette situation de grand déséquilibre des dimensions en défaveur de la taille de l'échantillon. La problématique statistique étudiée dans cette thèse est ainsi essentiellement illustrée et motivée par des exemples de données issues d'expériences par biopuces. C'est une biotechnologie permettant de mesurer simultanément le niveau d'expression de chacun des

gènes composant le génome d'un organisme. Elle propose une vision d'ensemble du génome et donne accès à une information essentielle en vue de mieux comprendre le rôle et la fonction de chaque petit morceau d'ADN, chez l'homme, chez l'animal ou encore chez le végétal.

Les données recueillies par biopuces sont utilisées notamment à des fins de diagnostics médicaux ou pour mesurer l'effet d'un traitement par exemple. Ce contexte biologique a profondément contribué au renouveau de la méthodologie statistique des tests multiples en grande dimension [Efron et al., 2001, Storey, 2002, Dudoit et al., 2003]. En effet, une question récurrente lors de l'étude de données issues de biopuces est l'identification de gènes différentiellement exprimés : il s'agit de détecter des gènes dont le niveau d'expression est lié à une covariable, indifféremment qualitative (groupe traitement/témoin) ou quantitative (dose de médicament). On va donc considérer simultanément pour chaque gène le test de l'hypothèse nulle selon laquelle il n'y a pas d'effet de la covariable sur le niveau d'expression génique.

La problématique est donc posée dans le cadre de l'analyse des données génomiques, mais les méthodes statistiques impliquées sont générales.

Les situations évoquées précédemment induisent un nombre de tests simultanés pouvant atteindre plusieurs milliers. La transposition des procédures de tests multiples contrôlant le FWER s'est vite avérée inadaptée pour des données de grande dimension, ces mêmes procédures devenant trop conservatrices dans ce contexte. Une prise en compte alternative de la multiplicité des tests dans la définition de la règle de décision et du risque d'erreur à contrôler est alors apparue comme cruciale. L'introduction par Benjamini and Hochberg [1995] d'une procédure (procédure BH) contrôlant le taux de faux-positifs (False Discovery Rate, FDR) à savoir l'espérance de la proportion d'erreurs parmi les rejets de l'hypothèse nulle, a ouvert de nouvelles perspectives à la mise en œuvre des tests multiples en grande dimension. Cette stratégie est rapidement devenue populaire, notamment lors des phases expérimentales exploratoires. Par ailleurs, sur le plan de la méthodologie statistique, l'introduction de la procédure BH a suscité de très nombreux développements. En effet, le cadre général repose initialement sur l'hypothèse de probabilités critiques indépendantes et identiquement distribuées selon un modèle de mélange à deux composantes [Efron et al., 2001], caractérisant respectivement la distribution sous l'hypothèse nulle et sous l'hypothèse alternative. Le paramètre de pondération de ce mélange, à savoir la vraie proportion d'hypothèses nulles, notée π_0, intervient directement dans le niveau effectif auquel le FDR est contrôlé. L'estimation de ce paramètre constitue une voie d'amélioration des procédures, en terme de puissance en particulier [Black, 2004, Kim and Van de Wiel, 2008]. Parmi les extensions les plus remarquables de la procédure BH, des méthodes adaptatives [Storey et al., 2004, Benjamini et al., 2006] ont donc été proposées, incluant divers estimateurs de π_0.

Si les problèmes liés à la grande dimension ont suscité beaucoup de développements récents dans la littérature statistique, celles relatives à l'hétérogénéité des données restent encore peu abordées.

Tests multiples et dépendance Depuis quelques années, une nouvelle orientation est donnée à l'amélioration des procédures d'inférence simultanée en grande dimension par la prise en compte

de l'impact de la dépendance sur la stabilité des procédures. En effet, la dépendance entre les tests dans l'analyse des systèmes complexes comme ceux évoqués précédemment est directement liée aux liens entre les variables décrivant ce système. Le signal étudié est observé en même temps qu'un certain nombre de facteurs de confusion. Le lien déduit entre les variables d'intérêt et la condition expérimentale peut alors s'avérer erroné, et les effets observés devraient être attribués à ces facteurs. Ils ne sont pas toujours contrôlés par le plan d'expérience et sont parfois non-observables. Ainsi, dans le traitement de données d'expressions géniques, la dépendance entre les tests résulte par exemple de processus biologiques non directement liés à celui étudié mais intervenant sur les variations d'expressions par le jeu des interactions entre les gènes. Les biais technologiques, qui sont pourtant partiellement corrigés par une étape de préparation des données d'expression appelée normalisation, ont aussi un impact sur la dépendance entre les mesures d'expressions. Cette dépendance est donc par nature complexe, difficile à modéliser, et le plus souvent non-négligeable. Néanmoins, sa prise en compte dans l'analyse statistique est un gage de validité des procédures d'inférence.

Méthodologie pour les tests multiples en grande dimension et en situation de dépendance Le FDR a particulièrement été au centre des discussions à ce sujet. Les approches envisagées dans la littérature pour prendre en compte la dépdedance proposent des corrections portant sur une des étapes des procédures de test. Ces modification peuvent avoir lieu au niveau de la statistique de test [Storey et al., 2007], de la définition de la loi de la statistique de test sous l'hypothèse nulle [Efron, 2004] ou encore au niveau de la définition du seuil de rejet des hypothèses [Benjamini and Yekutieli, 2001]. On cherche aussi à améliorer les performances des procédures existantes en apportant des arguments mathématiques démontrant le contrôle du FDR dans des conditions plus larges que celles envisagées initialement [Storey et al., 2004, Blanchard and Roquain, 2008, Sarkar, 2008]. Cela aboutit à des procédures adaptatives pour des données présentant certaines formes de corrélation. De manière générale, il résulte de ces discussions que l'application des méthodes classiques de test multiples en situation de dépendance aboutit au contrôle du FDR à un niveau moins élevé qu'attendu sous l'hypothèse d'indépendance, et donc à une détérioration de la puissance. Toutefois, des travaux récents [Efron, 2007] ont montré que la forte dépendance entre les statistiques de test avaient aussi, et probablement surtout, un impact sur la stabilité des procédures de tests multiples. Il s'avère en effet que la variabilité de la proportion d'erreurs parmi les rejets de H_0, dont le FDR est l'espérance, est d'autant plus grande que la dépendance entre les tests est forte. Dans ces situations de grande dépendance, le contrôle du FDR peut donc s'avérer inefficace.

Contrairement aux approches visant à s'adapter à la situation de grande dimension, la prise en compte de la dépendance remet en cause l'ensemble de la procédure de tests multiples, jusqu'aux statistiques de tests individuelles. En effet, le modèle des relations entre les variables réponses et la variable explicative d'intérêt n'est pas vu ici comme le produit de modèles individuels indépendants mais indépendants conditionnellement aux différentes composantes de l'hétérogénéité des données. Ainsi, la modélisation de la dépendance par des composantes de variance communes portées par l'ensemble des variables est apparue récemment, dans l'optique d'identifier dans la variation des expressions de gènes une part liée à cette hétérogénéité. Leek and Storey [2008] sont les premiers à

établir le lien entre cette dépendance et l'hétérogénéité des données d'expression. La modélisation de la structure de dépendance est alors une étape clé de l'analyse de ce type de données. Les travaux de recherche de cette thèse s'inscrivent dans ce contexte : nous nous intéressons aux procédures et à leurs propriétés en présence de dépendance.

Plus particulièrement, l'approche développée permet de tirer profit de cette information partagée par l'ensemble des variables pour stabiliser les procédures, et ce, par l'intermédiaire d'un modèle d'Analyse en Facteurs. Plus connue des psychométriciens et des sociologues comme une technique de réduction de la dimension, cette méthode a aussi été utilisée récemment comme une technique exploratoire d'analyse de la dépendance des données en grande dimension issues des expérimentations à "haut-débit" [Kustra et al., 2006, Hsu and Chang, 2006, Pournara and Wernisch, 2007]. Elle s'appuie sur l'existence d'un noyau linéaire de vecteurs aléatoires, appelés facteurs ou variables latentes, qui capture la structure de dépendance dans un espace de dimension réduite. Par ailleurs, l'analogie entre le modèle d'Analyse en Facteurs et les modèles à variables latentes permet l'utilisation d'algorithmes d'estimation de type *EM* [Rubin and Thayer, 1982].

Le premier objectif des travaux de recherche présentés ici est d'étudier les propriétés statistiques des procédures d'inférence simultanée en situation de dépendance. En particulier, nous nous intéressons à l'impact de la dépendance sur le contrôle des taux d'erreurs et sur la puissance des tests. Dans un second temps, cette étude analytique aboutit à la construction d'une stratégie adaptative permettant d'améliorer la puissance des tests multiples.

Le premier chapitre est consacré à un état des lieux des méthodes présentées dans la littérature dédiée aux tests multiples en grande dimension. Cette première partie décrit essentiellement les hypothèses sous-jacentes à la construction des procédures classiques de tests multiples, introduisant les différents taux d'erreurs, les méthodes de construction de règles de décision et l'estimation d'un paramètre clé de la plupart des procédures, la proportion d'hypothèses nulles (π_0).

Dans le second chapitre, différentes stratégies mises en œuvre pour prendre en compte la dépendance dans les procédures de tests multiples sont présentées et un cadre général d'étude de la dépendance est proposé. Dans ce cadre, l'impact d'un écart à l'indépendance est étudié, notamment sur la distribution des statistiques de test, le contrôle des taux d'erreur et l'estimation de la proportion d'hypothèses nulles. Le modèle proposé pour la structure de corrélation nous permet d'obtenir une expression exacte de la variance pour le nombre de faux-positifs, ainsi que pour l'estimateur empirique de π_0.

Le troisième chapitre présente l'introduction du modèle d'Analyse en Facteurs de la dépendance entre les tests dans les procédures d'inférence simultanée. Plus précisément, une approche basée sur des statistiques de tests indépendantes conditionnellement aux facteurs est proposée pour réduire l'impact de la dépendance. On montre en particulier l'intérêt de cette méthode pour l'estimation de π_0, des taux d'erreurs et leur contrôle en situation de forte dépendance.

Le quatrième chapitre est centré sur l'estimation des paramètres du modèle par Maximum de Vraisemblance, à travers un algorithme de type EM. En particulier, on montre l'intérêt de cette méthode

pour des données de grande dimension et on propose une méthode adaptée à la problématique des tests multiples pour le choix du nombre optimal de facteurs à inclure dans le modèle.

Enfin, le dernier chapitre propose l'application de notre méthode à des données génomiques. Ce chapitre est présenté comme une étude de cas, dans le double objectif de présenter l'utilisation du package `R` intégrant l'ensemble des outils développés dans cette thèse et de montrer l'intérêt de la démarche proposée pour des biologistes, en termes d'identification et interprétation de facteurs de confusion, de recherche de QTL, ou encore d'inférence sur des réseaux géniques.

Les travaux de recherche menés au cours de cette thèse ont été diffusés sous forme d'articles, de communications orales et de posters. D'autre part la méthode décrite est actuellement disponible pour les utilisateurs sous forme de module informatique dans le logiciel libre `R`. La liste des références et les quatre articles sont proposés à la fin de ce document.

CHAPITRE 1

TESTS MULTIPLES EN GRANDE DIMENSION

Résumé La méthodologie des tests multiples connaît ces dernières années un intérêt croissant, notamment depuis l'avènement de biotechnologies haut-débit, génératrices de tests simultanés en très grand nombre. Dans un premier temps, ce chapitre présente le cadre classique pour l'étude des procédures de tests multiples en grande dimension, notamment les hypothèses usuelles sur la distribution des probabilités critiques. Les différentes stratégies d'estimation des taux d'erreurs et de la proportion d'hypothèses nulles sont ensuite décrites.

Sommaire

Introduction : contexte statistique

Modélisation statistique Nous nous intéressons au lien entre m variables d'intérêt, notées $Y = [Y_1, \ldots, Y_m]$, et p variables explicatives, notées $x_1, \ldots, x_p]$, caractérisant des conditions expérimentales. Ce lien se formalise pour chaque variable Y_k par le modèle de régression suivant :

$$Y_k = m_k(X) + \epsilon_k \quad \forall k \in [1; m] \equiv \mathcal{M} \tag{1.1}$$

où $m_k(X)$ est la fonction de régression et ϵ_k l'erreur résiduelle, de densité φ_k. Par la suite, les erreurs du modèle sont supposées suivre une même loi de densité, définie à une constante d'échelle σ_k près : $\forall k, \varphi_k(\epsilon) = \varphi(\epsilon/\sigma_k)/\sigma_k$. En pratique cela traduit une forme d'homogénéité de la distribution des résidus.

Nous disposons d'échantillons de n observations indépendantes de Y. Par ailleurs, le nombre de variables d'intérêt peut être largement supérieur à la taille d'échantillon ($n << m$).

Tests multiples On parle de tests multiples lorsqu'on réalise simultanément les tests de plusieurs hypothèses. L'objectif est ici d'identifier le sous-ensemble $\mathcal{M}_0 \subset \mathcal{M}$ de taille m_0 pour lequel le modèle m_k prend une forme d'intérêt $m_k^{(0)}$:

$$\begin{cases} H_0^k : & m_k(X) = m_k^{(0)}(X) \\ H_1^k : & m_k(X) \neq m_k^{(0)}(X) \end{cases}$$

La statistique de test, notée $T_k = s_k(Y_k)$, est calculée à partir des observations de la variable correspondante. Sa fonction de répartition est connue sous l'hypothèse nulle, et notée $F_0^k(T)$. On définit alors la probabilité critique pour chaque test :

$$p_k = 1 - F_0^k(T_k) \tag{1.2}$$

Le principe des procédures de tests multiples repose sur deux étapes :

1. Calcul pour chaque variable Y_k de la statistique de test T_k, et de la probabilité critique associée p_k,
2. Définition d'un seuil t sur les probabilités critiques pour conclure sur le rejet ou non des hypothèses nulles.

La première étape des procédures ne dépend que de la nature de la variable d'intérêt Y_k et sa relation aux conditions expérimentales étudiées x. On suppose ici que l'on sait choisir une statistique de test pertinente, en adéquation avec la problématique, ce qui garantit la meilleure puissance.

La deuxième étape quant à elle nécessite une approche globale de la gestion des risques d'erreurs [Dudoit et al., 2002]. En particulier, le choix du seuil sur les probabilités critiques pour le rejet des hypothèses nulles va influencer le nombre d'erreurs commises (faux-positifs et faux-négatifs).

Le TABLEAU 1.1 [Benjamini and Hochberg, 1995] donne les notations utilisées classiquement pour comptabiliser les erreurs d'une procédure de tests multiples.

conclusion du test / réalité	non-rejet de H_0	rejet de H_0	Total
H_0	U_t	V_t	m_0
H_1	T_t	S_t	m_1
Total	$m - R_t$	R_t	m

TAB. 1.1 – Nombre d'erreurs d'une procédure de tests multiples, pour un seuil de rejet t pour les probabilités critiques.

Le nombre m d'hypothèses testées est connu. A l'inverse, m_0 et m_1, respectivement le nombre de vraies hypothèses nulles et de vraies hypothèses alternatives, sont des paramètres inconnus. Considérant un seuil t sur les probabilités critiques pour le rejet des hypothèses, $R_t = \sum_{k \in \mathcal{M}} \mathbb{1}_{p_k \leq t}$ est le nombre total d'hypothèses rejetées. C'est une variable aléatoire observable. En revanche, U_t et S_t d'une part, et T_t et V_t d'autre part, correspondant respectivement aux nombres de bonnes et de mauvaises décisions, sont des variables aléatoires non-observables.

L'objectif d'une procédure de tests multiples est de minimiser le nombre T_t de faux-négatifs (erreurs de type-II), tout en contrôlant le nombre V_t de faux-positifs (erreurs de type-I).

Or, le risque de commettre une erreur de type-I augmente très fortement avec le nombre d'hypothèses testées. Un petit exemple permet d'illustrer ce propos : si on considère les tests simultanés de $m = 10000$ hypothèses, en choisissant $t = 0,05$, $R_{0,05} = 500$ hypothèses seront rejetées par le simple fait du hasard. Cela n'est pas acceptable lorsqu'on souhaite mettre en évidence un phénomène qui concerne quelques centaines de variables.

Ainsi, la multiplicité des tests induit la nécessité de définir tout d'abord un taux d'erreurs de type-I global pour l'ensemble des tests réalisés, et non plus individuellement au niveau de chaque test. Ensuite, il s'agit de définir le seuil t pour les probabilités critiques qui permet d'assurer que ce taux d'erreurs soit inférieur à un seuil α fixé.

De nombreux articles et ouvrages récents décrivent le cadre général des tests multiples en grande dimension, notamment depuis l'avènement de biotechnologies haut-débit où l'analyse des données amène à considérer des problématiques de tests simultanés en très grand nombre [Efron et al., 2001, Storey, 2003, Dudoit et al., 2002, Dudoit and VanDerLaan, 2008]. Ce chapitre est consacré tout

d'abord à la présentation du cadre général des tests multiples en grande dimension. Il définit ensuite les taux d'erreurs les plus utilisés et décrit le rôle de la proportion d'hypothèses nulles dans le contrôle de ces taux d'erreurs. Enfin, la dernière partie de ce chapitre présente les principales procédures de tests multiples.

1.1. Modèle de mélange pour la densité des probabilités critiques

1.1.1 Cadre général

L'hypothèse sur la distribution des erreurs du modèle (1.1) suppose une densité identique à un coefficient d'échelle près. La distribution des statistiques de test $F^k(T)$ est alors caractérisée par un paramètre de non centralité $\tau_k(X) = m_k(X)/\sigma_k$. Dasn la suite, la fonction de répartition des probabilités critiques est notée $G^k(t) = \mathbb{P}(p_k \leq t)$.

Un grand nombre de procédures de tests multiples s'appuient sur le cadre général suivant :

HYPOTHÈSE 1. (Distribution des probabilités critiques)

1. *Sous H_0, la distribution des statistiques de test est la même pour l'ensemble des tests : $\forall k \in \mathcal{M}_0, F_0^k(T) = F_0(T)$.*
   Les probabilités critiques des tests sont donc définies par $p_k = 1 - F_0(T_k)$. Sous H_0, les probabilités critiques sont indépendantes, identiquement distribuées et réparties de façon uniformes sur $[0;1]$: $\forall k \in \mathcal{M}_0, G_0^k(t) = G_0(t) = \begin{cases} t \text{ si } p \in [0;1] \\ 0 \text{ sinon} \end{cases}$
2. *Sous H_1, les probabilités critiques sont indépendantes et identiquement distribuées selon une même loi : $\forall k \in \mathcal{M}_1, G_1^k(t) = G_1(t)$. Le paramètre de non-centralité de la distribution est donc supposé être le même : $\forall k \in \mathcal{M}_1$, $\tau_k = \tau$.*

Sous l'HYPOTHÈSE 1, la distribution des probabilités critiques peut donc s'exprimer sous la forme d'un modèle de mélange [Efron et al., 2001, Storey, 2002] :

$$G(t) = \pi_0 G_0(t) + (1 - \pi_0) G_1(t) \tag{1.3}$$

où $\pi_0 = \frac{m_0}{m}$ est la proportion d'hypothèses nulles (voir SECTION 1.2).

Formellement, Efron [2004] introduit le modèle de mélange d'un point de vue bayésien où m_0 est aléatoire. En pratique, la plupart des auteurs travaillent conditionnellement à m_0. Ce sera également le cas ici par la suite.

Ainsi, de la même manière, si g désigne la densité des probabilités critiques, alors :

$$g(p) = \pi_0 g_0(p) + (1 - \pi_0) g_1(p) \tag{1.4}$$

où $g_0(p) = 1, \forall p$. Le plus souvent, pour des raisons d'identifiabilité du modèle, d'autres hypothèses peuvent compléter le modèle, par exemple la décroissance de g_1 sur $[0;1]$ et $g_1(1) = 0$ [Genovese and Wasserman, 2002].

Nous présentons deux exemples pour illustrer ce modèle.

1.1.2 Cas du modèle linéaire

Le modèle linéaire est un cas particulier très usuel du modèle (1.1). C'est un des principaux outils statistiques mis en oeuvre lors de l'analyse de données expérimentales qui permet de formaliser une relation linéaire entre Y_k et x. Dans ce cas, $m_k(X) = X\theta_k$ (X de dimension $n \times (p+1)$ et θ_k de dimension $(p+1) \times 1$) et les erreurs sont supposées indépendantes et distribuées selon une loi Normale : $\varphi_k(\epsilon) \equiv \mathcal{N}(0; \sigma_k^2)$. Le modèle sous l'hypothèse nulle est alors défini comme un sous-modèle de m_k qui vérifie $c'\theta_k = 0$, où c est un $(p+1)$-vecteur de contrastes linéaires donné. La statistique de test pour tester la nullité d'un tel contraste est celle d'un test de Student :

DÉFINITION 1.1.1 (Statistique de test de Student)**.**

$$T_k = \frac{c'\hat{\theta}_k}{\sqrt{\hat{\mathbb{V}}(c'\hat{\theta}_k)}} = \frac{c'\hat{\theta}_k}{\sqrt{\hat{\sigma}_k^2 c'(X'X)^{-1}c}} \sim \mathcal{T}_{\tau_k}(ddl = n-p-1)$$

où τ_k le paramètre de non-centralité de la loi de Student à ddl degrés de liberté.

On note que $\tau_k = 0$ si $k \in \mathcal{M}_0$ et $\tau_k = \tau \neq 0$, si $k \in \mathcal{M}_1$.

EXEMPLE 1. Dans cet exemple, on considère $m = 500$ variables indépendantes distribuées selon une loi normale multivariée $\mathcal{N}(\mu; \mathbb{I}_m)$, observées sur $n = 60$ individus. On considère par ailleurs une variable x qualitative à deux modalités A et B, observées respectivement sur des échantillons de taille $n_A = n_B = n/2$. Pour un sous-ensemble de variables $\mathcal{M}_0$ de taille $m_0 = 400$, on a $\mu_k^A - \mu_k^B = 0$, où $\mu_k^A = \mathbb{E}(Y_k | x ='' A'')$ et $\mu_k^B = \mathbb{E}(Y_k | x ='' B'')$. Pour les $m_1 = 100$ autres variables, du sous-ensemble complémentaire $\mathcal{M}_1$, on a $\mu_k^A - \mu_k^B = \delta \neq 0$. La valeur de δ est définie afin d'assurer au test individuel de Student une puissance donnée, pour un risque de type-I α fixé, ici $\alpha = 0,05$. Le paramètre de non-centralité est défini par $\tau = \delta/\sqrt{2/n}$.

On note ϕ_0 et ϕ_τ les densités de la loi de Student à $n-2$ degrés de liberté et de paramètres de non-centralité respectifs 0 et τ, et $q_a = \phi_0^{-1}(a)$. Il est facile de voir que $g_1(p) = \frac{\phi_\tau(q_{1-p/2}) + \phi_\tau(q_{p/2})}{2 \times \phi_0(q_{1-p/2})}$ (voir FIGURE 1.1). Si la puissance de test individuelle est élevée, alors $g_1(1)$ est proche de 0. Dans le cas contraire, les distributions sous l'hypothèse nulle et sous l'hypothèse alternative ne sont pas bien séparées. Cela peut induire des problèmes d'identifiabilité des composantes du modèle (1.4) et pour l'estimation du paramètre π_0 (voir SECTION 1.2).

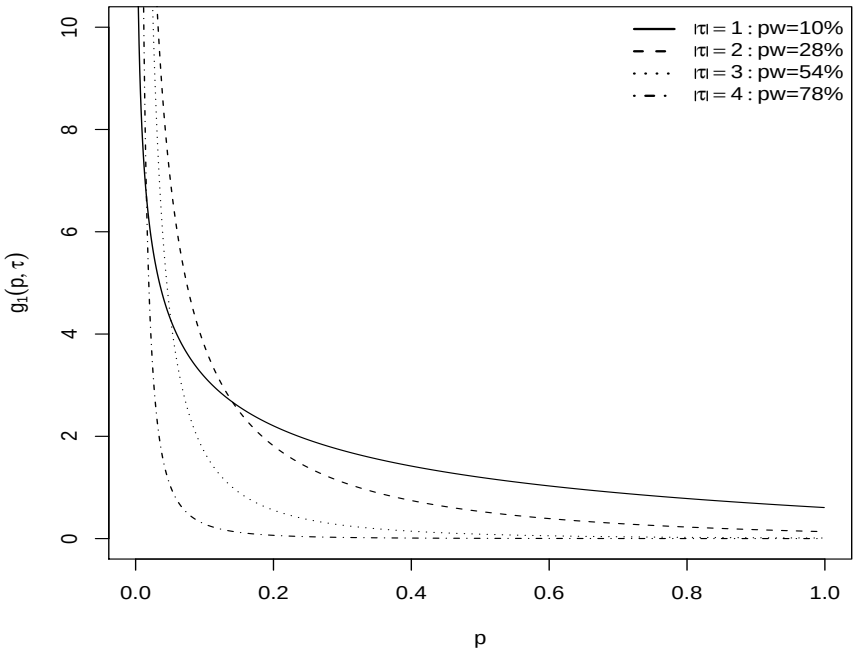


FIG. 1.1 – Représentation graphique de g_1 pour différentes valeurs de τ - pw : puissance du test individuel

1.1.3 Approche semi-paramétrique

EXEMPLE 2. DONNÉES GOLUB Cet exemple s'appuie sur une étude menée dans le cadre de la recherche sur la leucémie à travers l'analyse de données d'expressions géniques. L'étude réalisée par Golub [Golub et al., 1999] vise à mettre en évidence les gènes différemment exprimés entre deux types de tumeurs (AML et ALL). Les expressions géniques de 3051 gènes pour 38 tumeurs (27 de type ALL et 11 de type AML) ont été analysées par puces à ADN. Les données Y se présentent donc sous la forme d'une matrice $n \times m$ où $m = 3051$ (mesures d'expressions géniques) et $n = 38$ (prélèvements), le type de tumeur (AML ou ALL) étant la variable explicative x de l'étude. Une étape de normalisation a été effectuée au préalable [Dudoit et al., 2002]. Les données originales sont déposées sur un site public (`http://www.broadinstitute.org/cgi-bin/cancer/datasets.cgi`) et accessibles via le package R `multtest` [Pollard et al.] pour les données normalisées. Des tests de Student pour la comparaison de moyenne sont mis en œuvre pour chacune des variables.

La FIGURE 1.2 illustre la modélisation de la distribution des probabilités critiques. Le modèle de mélange (1.4) est ici estimé par une approche semi-paramétrique (voir Robin et al. [2007]) implémentée dans le package `kerfdr` de `R` [Guedj et al., 2007]. Dans le modèle (1.4), il s'agit donc d'estimer π_0 et g_1. Différentes méthodes d'estimation de la proportion d'hypothèses nulle π_0 sont détaillées dans la SECTION 1.2 de ce chapitre. Ce paramètre est ici estimé à $0,499$ par la méthode proposée dans

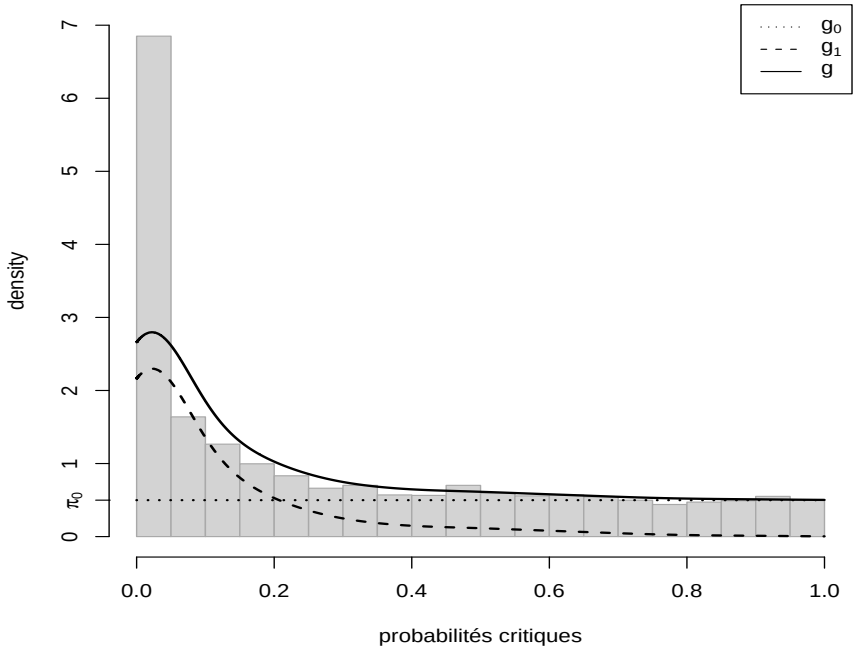


FIG. 1.2 – Distribution des probabilités critiques pour les données Golub : densité estimée et composantes du modèle de mélange (1.4)

le package. La densité sous l'hypothèse alternative g_1 est estimée en utilisant un estimateur à noyau [Silverman, 1986, Robin et al., 2007].

L'estimateur de la densité g_1 avec le noyau K et le paramètre de lissage ω est :

$$\hat{g}_1(p) = \frac{1}{m_1 \omega} \sum_{k \in \mathcal{M}_1} K\left(\frac{p - p_k}{\omega}\right) \tag{1.5}$$

L'expression (1.5) suppose la connaissance de $\mathcal{M}_1$. Une solution consiste à pondérer chaque observation par sa probabilité à postériori : $\eta(p_k) = \frac{(1-\pi_0)g_1(p_k)}{g(p_k)}$ [Efron et al., 2001]. On définit alors l'estimation de g_1 par un estimateur à noyau pondéré, où chaque probabilité critique p_k est pondérée par la probabilité que l'hypothèse nulle associée soit fausse :

$$\hat{g}_1(p) = \frac{1}{\omega \sum_{k \in \mathcal{M}} \eta(p_k)} \sum_{k \in \mathcal{M}} \eta(p_k) K\left(\frac{p - p_k}{\omega}\right) \tag{1.6}$$

On alterne alors des étapes d'estimation de g_1 (et g) puis de mise à jour de η [Robin et al., 2007]. Pour le choix du paramètre ω, on peut se référer à Silverman [1986].

1.2. Estimation de la proportion d'hypothèses nulles

Au delà de son intérêt en terme d'interprétation dans le contexte d'étude, la proportion d'hypothèses nulles est un paramètre crucial dans la mise en œuvre des procédures de tests multiples. Il est en

effet impliqué dans la définition du modèle de mélange présenté dans la SECTION 1.1 mais également dans le niveau de contrôle des taux d'erreurs (voir SECTION 1.3). De nombreuses méthodes ont été développées dans la littérature pour estimer ce paramètre et font l'objet de cette section.

Une première solution intuitive consiste à procéder de manière adaptative, en deux étapes. On utilise dans un premier temps $\hat{m}_0 = m$ dans une procédure de tests multiples en contrôlant le taux d'erreur à un niveau α, ce qui aboutit au non-rejet d'un certain nombre d'hypothèses qui va servir d'estimation de m_0 lors d'une seconde application de la procédure [Benjamini et al., 2006].

D'autres méthodes plus élaborées ont été développées. Elles reposent sur l'hypothèse d'un modèle de mélange à deux composantes (1.4), où la composante nulle de g est prépondérante pour les probabilités critiques proches de 1. On peut regrouper ces méthodes selon deux approches : des méthodes se basant sur la définition d'un estimateur empirique du paramètre d'une part, et des méthodes se basant sur l'estimation de la densité des probabilités critiques, en se servant de l'hypothèse selon laquelle $g(1) = \pi_0$, d'autre part.

1.2.1 Estimateur empirique

Définition L'estimateur empirique est défini à partir de W_λ, le nombre de probabilités critiques supérieures à λ :

$$W_\lambda = \#\{p_k > \lambda\} \tag{1.7}$$

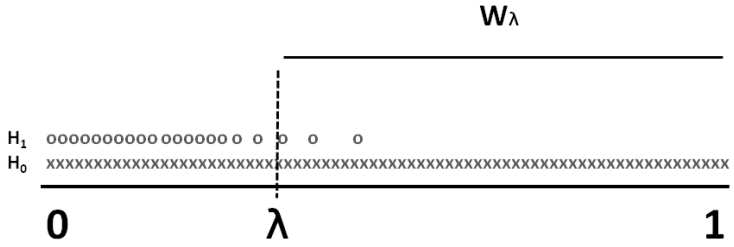


FIG. 1.3 – Définition de W_λ : seuil λ et répartition des probabilités critiques sous H_0 et sous H_1

Pour un choix judicieux de λ et sous l'hypothèse du modèle (1.4), les probabilités critiques sous H_1 sont en majorité distribuées sur un intervalle $[0; \lambda]$. Un grand nombre de probabilités critiques comprises dans $[\lambda; 1]$ sont alors associées à de vraies hypothèses nulles, et uniformément distribuées sur cet intervalle. On en déduit une approximation de l'espérance de $W(\lambda)$:

$$\mathbb{E}\left(W(\lambda)\right) \approx m\pi_0(1-\lambda) \tag{1.8}$$

Un estimateur naturel de π_0 est donc, pour un λ donné :

DÉFINITION 1.2.1 (Estimateur empirique de π_0 [Schweder and Spjotvoll, 1982])**.**

$$\hat{\pi}_0(\lambda) = \frac{\#(p_k > \lambda)}{(1-\lambda)m} = \frac{W_\lambda}{(1-\lambda)m}$$

Sous l'hypothèse d'indépendance entre probabilités critiques, la proposition suivante donne le biais et la variance de cet estimateur, à λ fixé.

PROPOSITION 1.2.1 (Propriétés de $\hat{\pi}_0(\lambda)$)**.**

$$\mathbb{E}\left(\hat{\pi}_0(\lambda)\right) = \pi_0 + \frac{1 - G_1(\lambda)}{1 - \lambda}(1 - \pi_0) \tag{1.9}$$

$$\mathbb{V}\left(\hat{\pi}_0(\lambda)\right) = \frac{\lambda \pi_0}{m(1-\lambda)} + \frac{G_1(\lambda)(1 - G_1(\lambda))(1 - \pi_0)}{m(1-\lambda)^2} \tag{1.10}$$

Démonstration. On a $W_\lambda = U_\lambda + T_\lambda$, où $U_\lambda = \sum_{k \in \mathcal{M}_0} \mathbb{1}_{p_k \geq \lambda}$ et $T_\lambda = \sum_{k \in \mathcal{M}_1} \mathbb{1}_{p_k \geq \lambda}$ (voir TABLEAU 1.1). Sous l'hypothèse d'indépendance des probabilités critiques, on a :

$$\begin{aligned} U_\lambda &\sim \mathcal{B}in(m_0, 1 - \lambda) \\ T_\lambda &\sim \mathcal{B}in(m - m_0, 1 - G_1(\lambda)) \end{aligned}$$

On a alors pour W_λ :

$$\begin{aligned} \mathbb{E}(W_\lambda) &= \mathbb{E}\left(U_\lambda\right) + \mathbb{E}\left(T_\lambda\right) = m_0(1-\lambda) + (m - m_0)(1 - G_1(\lambda)) \\ \mathbb{V}(W_\lambda) &= \mathbb{V}\left(U_\lambda\right) + \mathbb{V}\left(T_\lambda\right) = m_0\lambda(1-\lambda) + (m - m_0)(1 - G_1(\lambda))G_1(\lambda) \end{aligned}$$

On en déduit l'espérance et la variance de $\hat{\pi}_0$:

$$\begin{aligned} \mathbb{E}(\hat{\pi}_0(\lambda)) &= \frac{m_0(1-\lambda) + (m - m_0)(1 - G_1(\lambda))}{m(1-\lambda)} = \pi_0 + (1 - \pi_0)\frac{1 - G_1(\lambda)}{1 - \lambda} \\ \mathbb{V}(\hat{\pi}_0(\lambda)) &= \frac{\pi_0 \lambda}{m(1-\lambda)} + \frac{(1 - \pi_0)(1 - G_1(\lambda))G_1(\lambda)}{m(1-\lambda)^2} \end{aligned}$$

□

Le biais de l'estimateur empirique dépend du choix de λ. D'après (1.9), celui-ci est faible si chaque probabilité critique issue du test d'une variable sous H_1 a une probabilité faible d'être supérieure à λ ($G_1(\lambda) \approx 1$ pour λ petit). Cela équivaut au cas où les distributions des deux populations (sous H_0 et sous H_1) sont bien séparées Black [2004].

Plus particulièrement, le biais de l'estimateur dépend du paramètre $|\tau|$ de la distribution des probabilités critiques sous l'hypothèse alternative, qui caractérise la séparation entre les deux populations. La FIGURE 1.4 montre l'évolution du biais de l'estimation de π_0 en fonction de λ, pour différentes valeurs du paramètre de non-centralité, dans le cas où les probabilités critiques sont issues d'un test de Student (EXEMPLE 1).

La valeur minimale du biais est atteinte en $\lambda = 1$, quelle que soit la valeur de $|\tau|$. Elle dépend du rapport des densités sous H_1 et sous H_0 en ce point :

$$\mathcal{B}(\hat{\pi}_0(\lambda)) = (1 - \pi_0)\frac{1 - G_1(\lambda)}{1 - \lambda} \xrightarrow{\lambda=1} (1 - \pi_0)\frac{\phi_{|\tau|}(0)}{\phi_0(0)}$$

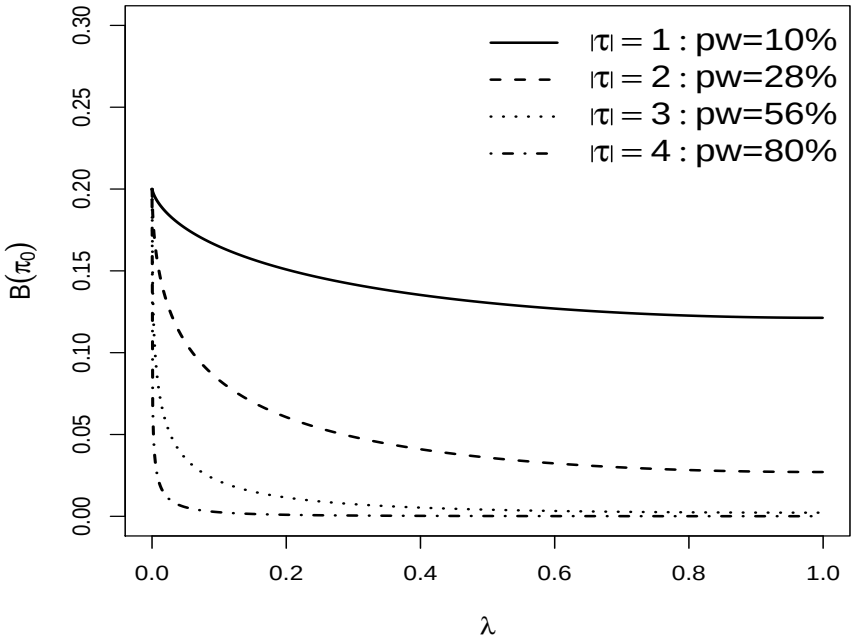


FIG. 1.4 – Évolution du biais de $\hat{\pi}_0$ en fonction du seuil λ, pour différentes valeurs du paramètre de non-centralité τ. Dans chaque cas, la puissance du test individuel (pw) est calculée - $\pi_0 = 0,80$

Cette quantité est strictement positive et tend vers 0 lorsque le paramètre de non-centralité $|\tau|$ augmente.

La variance de $\hat{\pi}_0$, au contraire, augmente quand λ tend vers 1.

Choix de λ Une première solution est de fixer une valeur arbitraire, par exemple $\lambda = 0,5$ [Tusher et al., 2001, Storey et al., 2004]. Mais cela ne tient pas compte des remarques précédentes sur le biais et la variance de l'estimateur. On peut, de manière moins arbitraire, choisir λ de façon à atteindre un compromis entre une estimation la moins biaisée possible (λ proche de 1) et de variance minimale (λ proche de 0), par exemple en minimisant l'Erreur Quadratique Moyenne de $\hat{\pi}_0$ définie à partir de (1.9) et (1.10) :

$$EQM\left(\hat{\pi}_0(\lambda)\right) \;=\; \mathbb{E}\left[(\hat{\pi}_0(\lambda) - \pi_0(\lambda))^2\right] = \mathbb{V}\left(\hat{\pi}_0(\lambda)\right) + \left(\mathcal{B}[\hat{\pi}_0(\lambda)]\right)^2 \tag{1.11}$$

Pour estimer (1.11), on peut utiliser une méthode de type *bootstrap* (méthode de rééchantillonnage) dont le principe est le suivant [Storey et al., 2004] :

1. On calcule un estimateur "plug-in" pour π_0 à partir de la DÉFINITION 1.2.1 [Storey, 2002] :

$$\hat{\pi}_0^{plug} \;=\; min_{\lambda' \in [0;1]} \left\{\hat{\pi}_0(\lambda')\right\}$$

2. On calcul L versions de $\hat{\pi}_0(\lambda)$ notées $\hat{\pi}_0^{*\ell}(\lambda)$, $\ell \in [1; L]$. Pour cela, on réalise L tirages avec remise de m probabilités critiques parmi les m probabilités critiques observées et pour chaque échantillon ℓ, on calcule l'estimateur $\hat{\pi}_0^{*\ell}(\lambda)$ de π_0.
3. On calcule ensuite l'estimateur *bootstrap* de l'EQM pour chaque valeur de λ :

$$\widehat{EQM}(\lambda) = \frac{1}{L}\sum_{\ell=1}^{L}\left(\hat{\pi}_0^{*\ell}(\lambda) - \hat{\pi}_0^{plug}\right)^2 \tag{1.12}$$

On choisit alors la valeur λ de $[0; 1]$ qui minimise $\widehat{EQM}(\lambda)$: $\hat{\lambda} = \text{argmin}_{\lambda\in\mathcal{R}}\{\widehat{EQM}(\lambda)\}$

Méthode par lissage Une autre approche a été développée, à partir de l'estimateur empirique Storey and Tibshirani [2003]. On calcule $\hat{\pi}_0(\lambda)$ pour $\lambda \in [0; 1]$ puis on effectue un lissage de la courbe $\lambda \longrightarrow \hat{\pi}_0(\lambda)$ par des splines cubiques pondérées. L'estimation de π_0 se fait en $\hat{\pi}_0(1)$[1].

1.2.2 Estimations basées sur un estimateur de la densité

Nous considérons maintenant l'estimation de π_0 à partir d'un estimateur non paramétrique de la densité g des probabilités critiques. En effet, sous l'HYPOTHÈSE 1, $\pi_0 = g(1)$. On peut alors estimer π_0 par $\hat{g}(1)$.

Plusieurs approches sont envisagées pour estimer la densité g. En pratique, comme g_0 est connue, il s'agit finalement d'estimer g_1.

Plusieurs auteurs [Langaas et al., 2005, Robin et al., 2007] proposent d'estimer g par une méthode à noyau (voir EXEMPLE 2). Dans la problématique d'estimation de π_0, il est fondamental de s'assurer d'une bonne estimation de la densité g en $p = 1$. Le choix de la fenêtre de lissage ω comme minimum de l'EQM en $p = 1$ étant instable en raison d'effets de bord, Langaas et al. [2005] proposent une solution consistant à étendre par symétrie la distribution des probabilités critiques sur $[1; 2]$ pour s'en affranchir.

Si g est une fonction supposée convexe et décroissante sur $[0; 1]$, alors nous pouvons estimer la densité des probabilités critiques à l'aide d'algorithmes itératifs de type "descente du gradient", dédiés à l'estimation de cette classe de fonctions Langaas et al. [2005].

Si G représente la fonction de répartition des probabilités critiques, $\hat{G}$ la fonction de répartition empirique et g la densité, on a :

$$\hat{G}(x) = \frac{1}{m}\sum \mathbb{1}_{[p<x]} = \frac{1}{m}\sum\left(1 - \mathbb{1}_{[p>x]}\right) = \frac{1 - W_x}{m}$$

d'où $1 - \hat{G}(x) = \frac{W_x}{m}$ or, $\hat{\pi}_0(x) = \frac{W_x}{m(1-x)}$ donc $\hat{\pi}_0(x) = \frac{1-\hat{G}(x)}{1-x}$.

[1] Dans le package R `qvalue`, cette méthode est implémentée avec $\hat{\pi_0} = \hat{\pi}_0(0.9)$ pour éviter les effets d'artéfacts à proximité de 1 [Dabney et al., 2009]

Le développement limité de Taylor de $G(p)$ en $p = x$ est :

$$G(p) = G(x) + (p - x)g(p) + o(p)$$

En $p = 1$, on a :

$$G(1) = 1 = G(x) + (1 - x)g(x) + o(1)$$

et donc

$$\frac{1 - G(x)}{1 - x} = g(x) + o(1)$$

Estimer π_0 par une estimation de la densité g en $p = 1$ est donc équivalent asymptotiquement à l'estimation par lissage de l'estimateur empirique [Storey and Tibshirani, 2003].

Les approches par estimation semi-paramétrique de la densité sont plus performantes en terme d'erreur quadratique moyenne que les méthodes basées sur l'estimateur empirique (voir étude détaillée de Langaas et al. [2005]). Néanmoins, la méthode d'estimation de π_0 basée sur l'estimateur défini en 1.2.1 avec choix de λ par *bootstrap* notamment est une des méthodes les plus utilisées en pratique, car plus simple à implémenter.

1.3. Taux d'erreurs

Les procédures de tests multiples ont été initialement développées dans l'objectif de contrôler le risque de faire une seule erreur parmi l'ensemble des rejets de H_0 (voir INTRODUCTION), mais d'autres définitions du taux d'erreur de type-I sont également utilisées, plus adaptées à la grande dimension. De même, dans les travaux concernant les tests multiples, on rencontre plusieurs définitions de l'erreur de type-II [Dudoit and VanDerLaan, 2008].

1.3.1 Taux d'erreurs de type-I

Les taux d'erreurs de type-I présentés ici sont les plus courants. On peut les regrouper en deux familles, selon qu'ils sont définis à partir de l'espérance du nombre de faux-positifs où de la probabilité que ce nombre dépasse une valeur donnée.

Taux d'erreurs basés sur l'espérance du nombre de faux-positifs V_t On définit le Per Comparison Error Rate (PCER) et le Per Family Error Rate (PFER) par :

$$PCER_t = \frac{\mathbb{E}(V_t)}{m} \qquad PFER_t = \mathbb{E}(V_t)$$

Le False Discovery Rate (FDR) [Benjamini and Hochberg, 1995] est l'espérance de la proportion d'erreurs de type-I parmi les hypothèses rejetées : $FDR_t = \mathbb{E}(FDP_t)$, où $FDP_t = \frac{V_t}{R_t}$ si $R_t > 0$ et 0 si $R_t = 0$.

On a donc :

$$FDR_t = \mathbb{E}(FDP_t | R_t > 0).\mathbb{P}(R_t > 0)$$

Par ailleurs, Storey et al. [2004] définissent le *positive FDR* (pFDR) par :

$$pFDR = \mathbb{E}(\frac{V_t}{R_t}|R_t > 0) = \frac{FDR_t}{\mathbb{P}(R_t > 0)}$$

Lorsque le nombre d'hypothèses testées est grand, $pFDR \approx FDR$ car $\mathbb{P}(R_t > 0) \overset{m\to\infty}{\to} 1$

Taux d'erreurs basés sur la probabilité d'avoir au moins k faux-positifs Le Family-Wise Error Rate (FWER) est la probabilité de commettre au moins une erreur de type-I :

$$FWER_t \;=\; \mathbb{P}(V_t \geq 1) \;=\; 1 - \mathbb{P}(V_t = 0)$$

D'une façon générale, on définit le *generalized*-Family-Wise Error Rate (g-FWER), où pour $k \in \mathcal{M}$

$$\text{g-}FWER_t(k) = \mathbb{P}(V_t > k) \;=\; 1 - \mathbb{P}(V_t \leq k)$$

Choix d'un taux d'erreurs de type I Pour une valeur fixée du seuil α sur le taux d'erreurs de type-I, les procédures qui contrôlent le PFER sont en général plus conservatrices (nombre plus faible d'hypothèses rejetées) que celles qui contrôlent le FWER ou le PCER, et les procédures qui contrôlent le FWER sont moins conservatrices que celles qui contrôlent le PCER [Dudoit and VanDerLaan, 2008]. Traditionnellement, le FWER a été le critère d'intérêt en matière de taux d'erreurs. Les premières procédures tenant compte de la multiplicité se sont intéressées au contrôle de ce taux d'erreurs [Bonferroni, 1936, Sidak, 1967, Holm, 1979], en particulier dans le cadre de l'analyse de variance, portant sur les tests de plusieurs contrastes d'un modèle linéaire. On choisit de contrôler le FWER lorsqu'on vise à une liste d'hypothèses rejetées comportant très peu de faux-positifs et qu'on accepte en contre-partie de ne pas détecter un certain nombre de variables sous H_1. Le contrôle du FWER est donc préconisé dans des analyses à but décisionnel par exemple.

Par ailleurs, choisir de contrôler le FDR (ou le p-FDR), c'est accepter de sélectionner à tort une proportion donnée de fausses hypothèses nulles. Cette approche est donc particulièrement intéressante en phase exploratoire. Ainsi, le FDR apparaît comme un bon compromis, entre le PCER, qui peut aboutir à un grand nombre de faux-positifs, et le FWER, qui diminue le nombre d'erreurs commises mais au détriment de la puissance. Ce taux d'erreur a d'ailleurs reçu beaucoup d'attention dans la littérature récente concernant les tests multiples en grande dimension [Benjamini and Hochberg, 1995, Benjamini and Yekutieli, 2001, Genovese and Wasserman, 2002, Storey, 2002, Sarkar, 2002, Storey, 2003]. Néanmoins, ces deux approches ne sont pas forcément à opposer et peuvent même apparaitre comme complémentaires. Ainsi, en phase exploratoire comme par exemple lors des premières étapes de l'élaboration d'un médicament, on peut autoriser quelques faux-positifs, et donc déterminer les tests positifs par une procédure qui contrôle le FDR. Par la suite, l'augmentation d'exigence en matière de précision conduit à une préférence pour le FWER, lors de phases finales de tests où aucun faux-positif n'est toléré.

1.3.2 Taux d'erreurs de type-II

La puissance d'une procédure de tests multiples se définit sous la forme d'une déclinaison similaire à celle du taux d'erreurs de type-I.

Taux d'erreurs basés sur l'espérance du nombre de faux-négatifs T_t On peut s'intéresser à la proportion de faux-négatifs $FNP_t = \frac{T_t}{m_1}$ et à son espérance $NDR_t = \mathbb{E}(FNP_t)$

Taux d'erreurs basés sur la probabilité d'avoir au moins un vrai-positif $\mathbb{P}(S_t \geq 1) = \mathbb{P}(T_t \leq m_1 - 1)$

1.4. Procédures de tests multiples

L'objectif d'une procédure de tests multiples est d'identifier les hypothèses nulles parmi les m hypothèses testées. D'une manière générale, il s'agit donc de déterminer un seuil t pour les probabilités critiques, tel que $p_k \leq t$ amène à rejeter l'hypothèse $H_0^{(k)}$. La valeur du seuil t peut être déterminée de différentes manières : fixée *à priori* ou dépendante des données. On dit qu'une procédure de tests multiples contrôle un risque d'erreurs de première espèce donné au risque α quand ce risque d'erreurs est inférieur ou égal à α. Si ce contrôle est assuré, quelle que soit la valeur de π_0, alors on parle de contrôle fort. Si il est assuré uniquement lorsque $\pi_0 = 1$, on parle alors de contrôle faible.

De nombreuses procédures peuvent être appliquées pour contrôler un des risques d'erreurs présentés précédemment (SECTION 1.3.1), voir par exemple le livre de Dudoit and VanDerLaan [2008] pour une présentation détaillée. Nous présentons dans cette section les principales méthodes de choix de t, généralement implémentées dans les logiciels. Le principe des différentes étapes des procédures de tests multiples est résumé sous forme de schéma à la FIGURE 1.5.

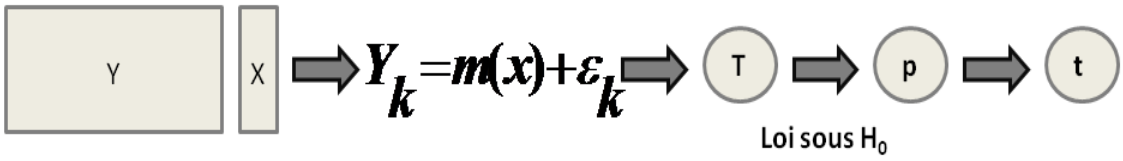


FIG. 1.5 – Étapes des procédures de tests multiples

1.4.1 Définitions et principe

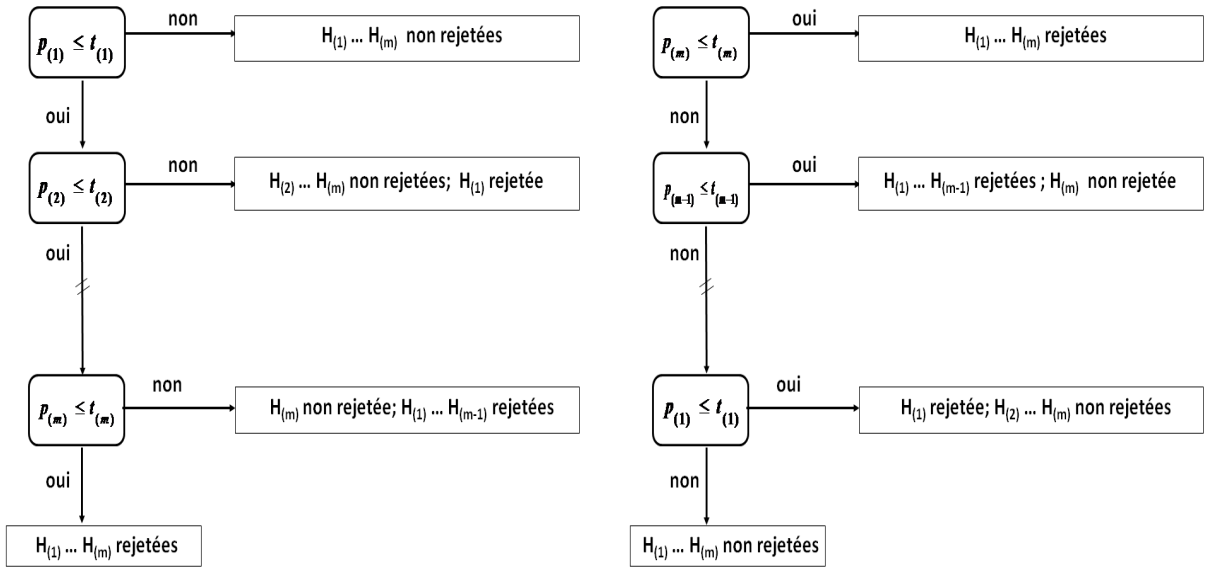
On considère habituellement deux types de procédures :

Procédures en une étape Un même seuil est considéré pour toutes les probabilités critiques, sans prendre en compte l'ordre de ces probabilités critiques, ni leur nombre. Chaque hypothèse est donc évaluée indépendamment des résultats obtenus pour les tests des autres hypothèses.

Procédures séquentielles Les tests sont considérés successivement dans l'ordre croissant (pour les procédures descendantes) ou décroissant (pour les procédures ascendantes) de leurs probabilités critiques, chaque test prenant en compte les résultats obtenus pour les tests des hypothèses précédentes. Dès qu'on ne rejette pas une hypothèse nulle, la procédure séquentielle descendante s'arrête et aucune autre hypothèse n'est rejetée - voir FIGURE 1.6(a). Dès qu'on décide de rejeter

une hypothèse nulle, la procédure séquentielle ascendante s'arrête et toutes les autres hypothèses sont rejetées - voir FIGURE 1.6(b)

Dans les procédures séquentielles, le seuil de rejet des hypothèses est donc redéfini à chaque étape.



(a) Procédures séquentielles descendantes (b) Procédures séquentielles ascendantes

FIG. 1.6 – Principe des procédures de tests multiples séquentielles. Les probabilités critiques sont ordonnées par ordre croissant, et $p_{(k)}$ représente la k^{eme} probabilité critique ordonnée

Cette section décrit les principales procédures de tests multiples. Elles ne permettent pas toutes de s'assurer que le même taux d'erreurs soit contrôlé à un niveau prédéfini α. Nous présentons dans un premier temps deux procédures permettant de définir le seuil t pour les probabilités critiques qui assure que $FWER_t \leq \alpha$, puis celles définissant t tel que $FDR_t \leq \alpha$.

1.4.2 Contrôle du FWER

Généralement, deux procédures sont utilisées. La première est celle proposée par Bonferroni [1936]. Elle est basée sur le principe d'un seuil $t = \alpha/m$, permettant d'assurer que le FWER est inférieur au seuil α, sans hypothèse particulière sur la distribution des probabilités critiques :

$$\begin{aligned}\mathbb{P}(V_t > 0) &= \mathbb{P}\left(\sum_{k\in\mathcal{M}_0} \mathbb{1}_{p_k<\alpha/m} > 0\right) = \mathbb{P}\left(\bigcup_{k\in\mathcal{M}_0} [p_k < \alpha/m]\right) \\ &\leq \sum_{k\in\mathcal{M}_0} \mathbb{P}\left(p_k < \alpha/m\right) = m_0\frac{\alpha}{m} \leq \alpha\end{aligned}$$

d'après l'inégalité de Boole.

La procédure de Sidak [1967] s'appuie sur l'hypothèse d'indépendance des probabilités critiques sous

H_0. Si $t = 1 - (1-\alpha)^{1/m}$ alors :

$$\begin{aligned}\mathbb{P}(V_t > 0) &= 1 - \mathbb{P}(V_t = 0) = 1 - \mathbb{P}\left(\bigcap_{k\in\mathcal{M}_0}\left[p_k > 1-(1-\alpha)^{1/m}\right]\right)\\ &= 1 - \prod_{k\in\mathcal{M}_0}\mathbb{P}\left(p_k > 1-(1-\alpha)^{1/m}\right) = 1-(1-\alpha)^{m_0/m} \leq \alpha\end{aligned}$$

On remarque que pour α petit (en particulier pour $\alpha = 0,05$, seuil usuel de niveau de contrôle du taux d'erreurs de type-I en partique) et m grand : $1-(1-\alpha)^{1/m} \approx \alpha/m$. Les procédures de Sidak et de Bonferroni ont des niveaux de contrôle du FWER qui s'apparentent.

1.4.3 Contrôle du (p)FDR

On note $\mathcal{R}_t = \{k|p_k \leq t\}$, $\mathcal{R}_t \subset \mathcal{M}$, de cardinal R_t. Intuitivement, on cherche à définir le seuil t qui permet de construire le plus grand ensemble $\mathcal{R}-t$ tel que $FDR_t \leq \alpha$. Pour cela, on approche FDR_t par le rapport d'estimateurs des espérances de V_t et de R_t : $FDR_t \approx \frac{\mathbb{E}(V_t)}{R_t} = \frac{\sum_{k\in\mathcal{M}_0}\mathbb{P}(p_k\leq t)}{R_t} \leq \frac{m_0 t}{R_t} \leq \frac{mt}{R_t}$. On a alors : $t \leq \alpha R_t/m \Rightarrow FDR_t \leq \alpha$.

D'une façon générale, pour garantir le contrôle du FDR au niveau α, l'algorithme de construction de $\mathcal{R}_t$ doit vérifier la condition $\mathcal{R}_t \subset \{k|p_k \leq \alpha\beta(R_t)/m\}$, dite d'autoconsistance [Blanchard and Roquain, 2008], où $\beta(x)$ est une fonction croissante. Cette condition est vérifiée en particulier par les procédures séquentielles ascendantes.

Dans le cas de l'indépendance, $\beta(x) = x$ correspond à la procédure de Benjamini and Hochberg [1995] (ci-après appelée procédure BH). En notant $p_{(1)}, \ldots, p_{(m)}$ les probabilités critiques ordonnées par ordre croissant, la procédure BH définit ainsi le seuil : $t = p_{(k^*)}$, avec $k^* = \text{argmax}\left(k \mid p_{(k)} < \frac{k}{m}\alpha\right)$. Le contrôle du FDR par cette procédure a initialement été démontré à l'aide de l'inégalité de Simes [Simes, 1986].

Étant donné un niveau α fixé pour le contrôle du FDR, la procédure BH consiste à trouver le seuil t_α, dépendant des données. Une autre approche consiste à fixer le seuil de rejet t et de calculer le taux d'erreurs correspondant :

$$t_\alpha = \underset{t\in[0;1]}{\text{argmax}}\{\widehat{FDR}_t(\lambda) \leq \alpha\}$$

où $\widehat{FDR}_t(\lambda) = \frac{\hat{m}_0(\lambda)t}{R_t \vee 1}$, avec $\hat{m}_0(\lambda) = m\hat{\pi}_0(\lambda)$ et λ judicieusement choisi (voir SECTION 1.2). C'est le point de vue adopté en particulier par Storey [2002]. Les deux approches sont équivalentes, si m est remplacé par $\hat{m}_0(\lambda)$ dans la procédure BH [Storey et al., 2004].

Storey [2003] établit un parallèle entre les procédures de tests simples et multiple en introduisant la *q-value*, définie par : $q_k = \text{argmin}_{t>p_k}\{pFDR_t\}$. D'un point de vue pratique, on estime la *q-value* par : $\hat{q}_k = \frac{\hat{m}_0(\lambda)p_{(k)}}{k}$ avec $\hat{m}_0(\lambda) = \frac{\#\{p_k>\lambda\}}{(1-\lambda)}$ (voir SECTION 1.2). Les hypothèses rejetées sont celles correspondant à une *q-value* inférieures à α, pour un contrôle du pFDR à ce niveau.

Conclusion : Amélioration des procédures

Nous avons présenté un cadre général pour les procédures de tests multiples en grande dimension, et introduit les principales notions qui y sont liées. Ce cadre suppose l'hypothèse d'un modèle de mélange à deux composantes pour la densité des probabilités critiques et leur indépendance. Une attention particulière est portée à l'estimation de la proportion d'hypothèse nulle et à la définition du taux d'erreurs de type-I qu'on souhaite contrôler.

De nombreux auteurs se sont ensuite intéressés à l'amélioration en terme de puissance de ces procédures : il s'agit, pour un même niveau de contrôle du taux d'erreurs de type-I, d'augmenter le nombre de vrais-positifs. Les propositions se sont concentrées sur les algorithmes de contrôle des taux d'erreurs en modifiant la procédure de choix du seuil t.

Les procédures en une étape sont caractérisées par une région de rejet identique pour l'ensemble des hypothèses testées. Un gain en puissance peut être atteint par les procédures séquentielles, ascendantes ou descendantes, pour un même niveau de contrôle du taux d'erreurs. Ainsi, en particulier, il existe des versions séquentielles de la procédure de Bonferroni ou de celle de Sidak [Holm, 1979], qui sont moins conservatrices que la version en une étape.

L'amélioration des procédures de tests multiples par l'estimation de π_0 est l'une des orientations les plus largement étudiées aujourd'hui [Storey, 2002, Kim and Van de Wiel, 2008]. En effet, les procédures présentées précédemment utilisent dans leur version initiale une approximation de m_0 par m. Cela rend les procédures conservatrices, les méthodes contrôlant alors les taux d'erreurs au niveau $\pi_0\alpha$ et non α.

On voit en particulier que considérer $\hat{m}_0 = m \Leftrightarrow \hat{\pi}_0 = 1$ conduit à choisir un seuil t plus faible, ce qui donne alors un nombre de rejets moins important. Si la vraie valeur de π_0 est éloignée de 1, il est donc intéressant d'inclure son estimation dans les procédures pour gagner en puissance [Black, 2004].

Par ailleurs, dans de nombreuses situations, la forte dépendance entre les variables mesurées induit une autre dépendance entre les tests, pouvant conduire à une remise en cause des performances des procédures de tests multiples. La prise en compte de la dépendance dans les procédures de tests multiples constitue alors une voie d'amélioration des méthodes.

Dans ce but, de nombreux auteurs se sont intéressés aux algorithmes de contrôle des taux d'erreurs en étudiant leurs propriétés en présence de certaines formes de dépendance ou en en modifiant la définition du seuil t. On peut montrer ainsi que dans certains cas de dépendance, le contrôle du FDR par la procédure BH est assuré : par exemple lorsque la matrice de corrélation présente une structure par blocs [Storey et al., 2004] ou dans le cas de dépendance positive [Benjamini and Yekutieli, 2001]. Cette dernière corrige le seuil de la procédure BH pour des données non-indépendantes. Les deux procédures (BH et BY) diffèrent simplement dans le coefficient appliqué au seuil de rejet pour les probabilités critiques : $\frac{m}{k}$ pour BH ou $\frac{m}{k}\sum_{j=1}^{m}\frac{1}{j}$ pour BY. Toutefois, cette méthode est en général très conservatrice.

On peut également adapter la forme de la fonction β pour les procédures séquentielles ascendantes permettant de contrôler le FDR à des situations de dépendances assez générales [Blanchard and Roquain, 2008] : $\int_0^x u d\nu(u)$, avec ν une distribution a priori sur le nombre de rejets de la procédure. En particulier, en prenant $\nu(k) \propto 1/k$, on retrouve la procédure BY [Benjamini and Yekutieli, 2001].

D'autres approches ont été étudiées, comme la modification de la statistique de test [Lönnstedt and Speed, 2002, Smyth, 2004, Storey et al., 2007] ou de la loi sous l'hypothèse nulle pour le calcul des statistiques de test [Efron, 2004]. De même, les observations étant indépendantes, on peut prendre en compte la structure (inconnue, mais potentiellement importante) de corrélation des statistiques de test (sous l'hypothèse nulle) à l'aide de techniques de permutations ou de rééchantillonnage [Westfall and Young, 1993, Tusher et al., 2001]. En pratique, le problème des méthodes de rééchantillonnage en grande dimension vient principalement du temps de calcul nécessaire, et de la fiabilité en terme de contrôle de taux d'erreurs [Yifan et al., 2006].

Dans la suite, nous nous intéressons tout particulièrement à l'impact de la dépendance sur les procédures de tests multiples en général. Récemment, l'idée de prendre en compte la dépendance à travers la modélisation de l'information partagée par l'ensemble des variables a été évoquée dans la littérature [Kendziorski et al., 2003, Leek and Storey, 2008].

Cette étude est présentée dans le CHAPITRE 2.

CHAPITRE 2

DÉPENDANCE ET TESTS MULTIPLES

Résumé La prise en compte de la dépendance est un des sujets cruciaux de la mise en œuvre des procédures de tests multiples en grande dimension. La présence d'une structure de corrélation induit en effet une instabilité dans le contrôle des taux d'erreurs et dans l'estimation de paramètres comme la proportion d'hypothèses nulles π_0. Dans de nombreux domaines, la dépendance observée dans les données peut être expliquée par l'existence d'une structure sous-jacente non observée. A travers la modélisation de cette structure de corrélation par un modèle d'Analyse en Facteurs, nous proposons donc de prendre en compte l'information partagée par l'ensemble des variables observées grâce à un petit nombre de variables latentes, les facteurs communs. Nous obtenons alors une expression exacte de la variance du nombre de faux-positifs en présence de dépendance, ainsi que pour l'estimateur empirique de π_0.

Sommaire

Introduction

Les procédures de tests multiples définies au chapitre précédent sont basées sur l'hypothèse que les m probabilités critiques issues des tests sont distribuées indépendamment selon un modèle de mélange à deux composantes (Hypothèse 1). Le modèle défini en (1.1) suppose en effet l'indépendance entre les résidus ϵ_k, ce qui n'est pas une hypothèse réaliste dans de nombreux contextes d'applications. Les données expérimentales sont affectées par un certain nombre de facteurs environnementaux, biologiques ou encore techniques, non-observés (ou non observables) qui induisent de la variabilité dans les mesures.

Dans le modèle (1.1), les sources de variabilité x sont explicitement modélisées à travers $m_k(x)$. Les résidus quant à eux concentrent la variabilité non expliquée par $m_k(x)$, imputables à des sources de variabilités autres que x non observées dans l'étude (facteurs environnementaux, démographiques, techniques, lié au plan d'expériences, etc.) ainsi que la variabilité spécifique à chaque variable d'intérêt.

Ces facteurs non observés induisent une forme de dépendance générale entre les variables d'intérêt. Dans de nombreuses situations, cette dépendance induit une autre dépendance entre les tests, pouvant conduire à une remise en cause des performance des procédures de tests multiples. La prise en compte de la dépendance dans ces procédures constitue donc un enjeu majeur de leur fiabilité et offre des perspectives d'amélioration de ces procédures.

Dans la littérature dédiée aux tests multiples, le contrôle des taux d'erreurs et l'écart à la distribution nulle théorique en présence de dépendance sont les deux points majoritairement évoqués. La plupart des approches visant à prendre en compte la dépendance consistent alors à corriger les procédures de contrôle du risque de première espèce [Benjamini and Yekutieli, 2001] ou en une modification de la loi des statistiques de test sous l'hypothèse nulle [Efron, 2004].

Sous certaines formes de dépendance, les procédures basées sur l'hypothèse d'une loi de mélange pour les probabilités critiques sous l'hypothèse nulle restent valides, c'est-à-dire que le taux d'erreurs de type-I est toujours contrôlé. En particulier, on évoque ici deux formes de dépendance pour lesquelles la procédure BH assure le contrôle du FDR : PRDS (Positive Regression Dependance on Subset) [Benjamini and Yekutieli, 2001] et la dépendance faible [Storey, 2002]. Les cas de dépendance faible se rencontrent lorsque la matrice de corrélation présente une structure par blocs finis, ou encore lorsque les statistiques de tests sont vues comme un processus aléatoire stationnaire ergodique ou suivent certaines lois de mélange par exemple.

A l'instar de Leek and Storey [2008], l'approche proposée ci-après consiste à tirer profit d'une infor-

mation partagée par l'ensemble des variables, issue des facteurs de variabilités non observés.

On considère le modèle (1.1) pour chaque variable d'intérêt Y_k. Leek and Storey [2008] proposent un cadre général d'étude de la dépendance basé sur la proposition suivante, démontrée sans hypothèse particulière sur la distribution de ϵ :

PROPOSITION 2.0.1 (voir Leek et Storey (2008)). *Soit $\epsilon_k = Y_k - \mathbb{E}(Y_k|x)$ l'erreur résiduelle du modèle de régression de la kème variable sur les covariables x mesurant les conditions expérimentales. S'il n'existe aucune fonction g mesurable telle que $\epsilon_k = g(\epsilon_1, \ldots, \epsilon_{k-1}, \epsilon_{k+1}, \ldots, \epsilon_m)$ presque sûrement, alors il existe une matrice $Z = (Z_1, Z_2, \ldots, Z_Q)$, avec $0 \leq Q \leq m$ et, pour tout $k \in [1; m]$, il existe des vecteurs b_k tels que $\epsilon_k = b_k' Z + \varepsilon_k$, où les composantes de $E = [\varepsilon_1, \ldots, \varepsilon_m]$ sont indépendantes.*

La PROPOSITION 2.0.1 introduit un ensemble de vecteurs Z qui capture la variabilité partagée par l'ensemble des variables. Ainsi, si b_k est le $Q-$vecteur des coefficients de chaque variable Y_k dans les combinaisons linéaires les reliant aux Q variables latentes Z, le modèle (1.1) devient alors :

$$Y_k = m_k(x) + Z b_k' + \varepsilon_k \tag{2.1}$$

et les composantes de $E = [\varepsilon_1, \ldots, \varepsilon_m]$ sont indépendantes.

Dans le cadre du modèle linéaire décrit à la SECTION 1.1.2, le modèle (2.1) s'écrit sous forme matricielle :

$$Y_{n\times m} = X_{n\times(p+1)}\theta_{(p+1)\times m} + Z_{n\times Q} B'_{m\times Q} + E_{n\times m} \tag{2.2}$$

Nous proposons de représenter la dépendance conditionnellement aux conditions expérimentales x à travers le sous-espace de faible dimension engendré par $Z = [Z_1; \ldots; Z_Q]$. L'estimation des paramètres et du nombre de facteurs à inclure dans le modèle est traitée au CHAPITRE 4.

Néanmoins, nous pouvons d'ores et déjà remarquer que cette décomposition s'apparente à celle d'un modèle d'Analyse en Facteurs (Factor Analysis). C'est d'ailleurs par analogie avec cette méthode que nous proposons une méthode d'estimation des paramètres du modèle (2.1). Par la suite, nous utiliserons le vocabulaire traditionnellement lié à cette méthode : les variables latentes Z sont les facteurs communs ou scores, et les composantes de E sont les facteurs spécifiques.

Dans ce cadre général de dépendance, nous allons donc étudier les propriétés des procédures de tests multiples.

Dans le modèle (2.1), où les résidus $E = [\varepsilon_1, \ldots, \varepsilon_m]$ ont pour matrice de variance-covariance Ψ, diagonale, on suppose que $\mathbb{V}(Z) = \mathbb{I}$ comme en Analyse en Facteurs dite exploratoire [Mardia et al.,

1979]. On obtient alors la décomposition suivante pour la matrice de variance-covariance de Y, notée Σ :

$$\Sigma = BB' + \Psi \tag{2.3}$$

où B est une matrice de taille $m \times Q$, dont la k^{eme} ligne correspond au vecteur b_k. Dans la terminologie de l'Analyse en Facteurs, BB' représente l'information commune à l'ensemble des variables et les termes diagonaux de Ψ sont les variances spécifiques.

Dans un premier temps, nous nous appuyons sur des simulations pour illustrer l'impact de la dépendance selon le niveau de corrélation des données. Dans des études similaires réalisées dans le cadre de l'analyse de données génomiques, les schémas de corrélation des données simulées sont relativement simples : données équi-corrélées [Allison, 2002] ou équi-corrélées par blocs [Korn et al., 2004] par exemple. Ces structures de dépendance sont simples à mettre en œuvre, facilement interprétables, et le contrôle du niveau de dépendance au sein de chaque scénario est possible en faisant varier la valeur du(es) coefficient(s) de la matrice de corrélation. Néanmoins, cette modélisation de la dépendance est éloignée de la réalité, où les liaisons entre variables d'intérêts sont parfois plus complexes. Dans les études par simulations de ce chapitre, d'une façon similaire à Kim and Van de Wiel [2008], nous proposons de considérer un ensemble de matrices de corrélations, tout en imposant une contrainte d'indépendance conditionnelle. La matrice de variance-covariance théorique se décompose sous la forme (2.3). Le niveau de dépendance souhaité est atteint en pondérant la structure de dépendance commune (BB') par un coefficient.

Exemple 3. On simule des données à partir de 10 scénarios pour la matrice de variance-covariance Σ. Ces scénarios traduisent une augmentation graduelle de la structure de dépendance, de l'indépendance (scénario 0) à une forte structure de corrélation (scénario 9) (voir Tableau 2.1). On contrôle le niveau de dépendance de chaque scénario s par un indicateur de l'information partagée par l'ensemble des variables, la variance commune, défini par : $tr(BB')/tr(\Sigma)$.

scénario	0	1	2	3	4	5	6	7	8	9
variabilité commune (%)	0	4,65	15,56	28,56	40,31	50,76	57,90	65,45	70,09	75,19

Tab. 2.1 – Variabilité commune (%) pour 10 scénarios

1000 tableaux de données sont simulés pour chaque scénario s. Les m variables de chaque tableau de données sont générées selon une loi normale multivariée : $\mathcal{N}_m(\mu; \Sigma_s)$, où μ est un m-vecteur contenant les moyennes de chaque variable et s un niveau de dépendance donné. Chacun est composé de $m = 500$ variables pour $n = 60$ individus. On considère par ailleurs une variable x qualitative à deux modalités A et B, telle que $n_A = 30$ et $n_B = 30$. Les données sont tout d'abord simulées de façon à ce que la loi jointe des variables ne dépende pas de x.

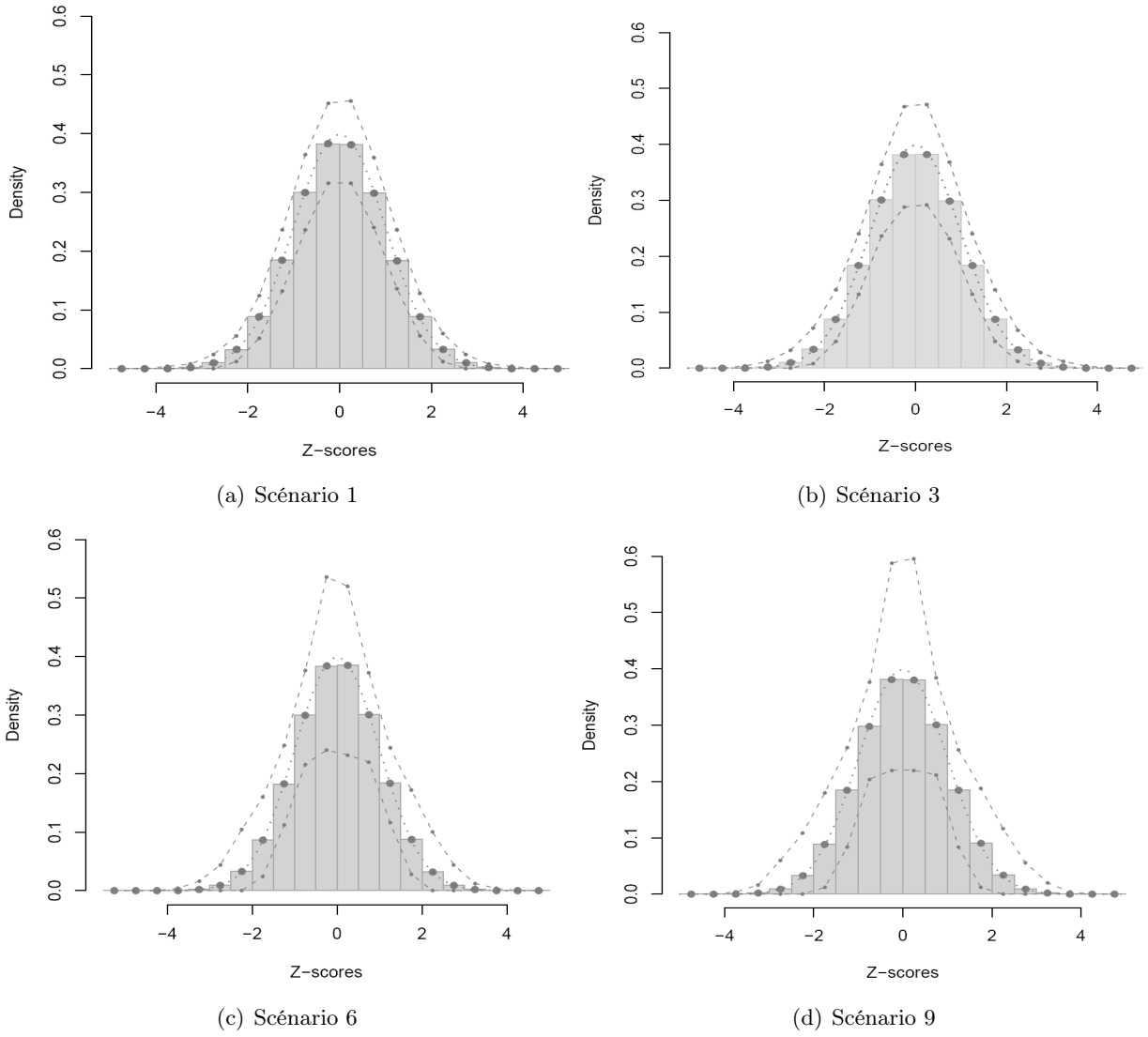
L'annexe A propose le code `R` permettant de simuler les tableaux de données de cet exemple.

Nous nous intéressons dans un premier temps à la première étape des procédures de tests multiples : le calcul des statistiques de tests et celui des probabilités critiques. Dans un deuxième temps, nous

étudions l'impact de la dépendance sur l'estimation de π_0 puis sur le nombre de faux-positifs et les taux d'erreurs.

2.1. Étude de l'impact de la dépendance sur la distribution des probabilités critiques

L'HYPOTHÈSE 1 suppose que les statistiques de tests suivent sous l'hypothèse nulle une même loi de distribution F_0. Comme suggéré par McLachlan et al. [2006] ou plus récemment par Efron [2007], les statistiques de test sont transformées en Z-scores pour avoir comme loi de référence sous H_0 la loi normale centrée-réduite : $Z_k = \Phi^{-1}(F_0(T_k)), k \in [1; m]$, où Φ est la fonction de répartition de la loi normale centrée-réduite. On a donc $Z_k \sim \mathcal{N}(0; 1)$, pour $k \in \mathcal{M}_0$.



(a) Scénario 1

(b) Scénario 3

(c) Scénario 6

(d) Scénario 9

FIG. 2.1 – Distribution des Z-scores sous H_0 - Histogramme moyen sur l'ensemble des 1000 simulations - lignes pointillées : quantiles à $2,5\%$ et $97,5\%$

En s'appuyant sur les données simulées de l'EXEMPLE 3, nous étudions l'impact de la dépendance sur la distribution des Z-scores sous H_0, puis sur celle des probabilités critiques des tests. Sur chaque tableau de données simulé, les $m = 500$ tests de l'effet de la variable x sont réalisés puis les Z-scores associés aux statistiques de tests sont calculés. Dans cet exemple, F_0 est la loi de Student à $n-2 = 58$ degrés de liberté.

Pour quatre des scénarios (1 : faible niveau de dépendance, 3 et 6 : niveaux intermédiaires, 9 : niveau élevé), on représente l'histogramme moyen ainsi que la variabilité autour de cette moyenne à travers les quantiles à $2,5\%$ et $97,5\%$ des distributions des Z-scores d'une part (FIGURE 2.1) et des probabilités critiques d'autre part (FIGURE 2.2).

Quel que soit le niveau de dépendance, la répartition moyenne des Z-scores est bien ajustée par la densité théorique (loi Normale centrée-réduite). Dans le cas de données indépendantes (2.1(a)), les histogrammes obtenus pour les Z-scores varient peu autour de l'histogramme moyen. Lorsque la corrélation entre les données augmente (FIGURE 2.1(b) puis FIGURES 2.1(c) et 2.1(d)), on observe en revanche une augmentation de la variabilité de la distribution des Z-scores autour de l'histogramme moyen. En présence de dépendance, les distributions des Z-scores se démarquent de la loi théorique par un aplatissement plus ou moins fort. On note ici qu'on s'intéresse à la distribution de l'histogramme des Z-scores, et non à un Z-score donné dont la distribution marginale reste une loi Normale centrée-réduite.

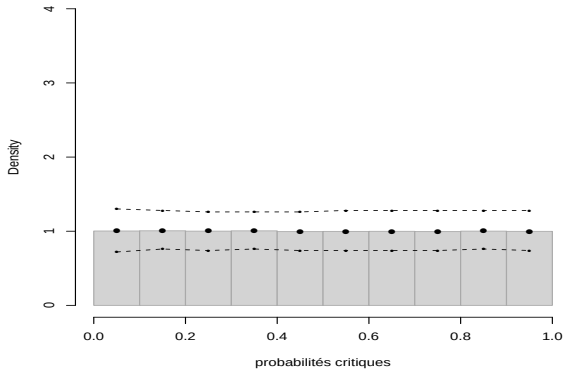
De même, la répartition moyenne des probabilités critiques est uniforme sur $[0;1]$. Comme pour la distribution des Z-scores, dans le cas de données indépendantes, les histogrammes obtenus pour les probabilités critiques varient peu autour de l'histogramme moyen. Lorsque la corrélation entre les données augmente (FIGURE 2.2(b) puis FIGURES 2.2(c) et 2.2(d)), la variabilité de la distribution des probabilités critique autour de l'histogramme moyen devient plus importante. Ainsi, la distribution empirique peut s'éloigner fortement de la distribution uniforme.

La FIGURE 2.3 représente les histogrammes de deux jeux de données simulées, parmi les 1000 jeux de données du scénario 9. En présence d'un niveau de dépendance élevé, les distributions des probabilités critiques se démarquent de la loi uniforme (loi théorique, en pointillé rouge) selon deux configurations :

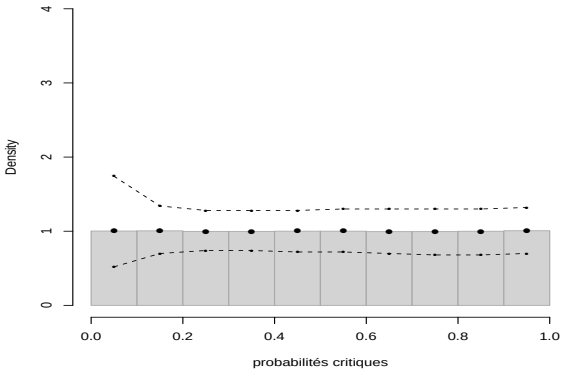
- soit en ayant une sur-représentation des probabilités critiques proches de 0 (à gauche), et donc une sous-représentation des probabilités critiques proches de 1,
- soit l'inverse (à droite), c'est-à-dire une sous-représentation des probabilités critiques proches de 0 et donc une sur-représentation des probabilités critiques proches de 1.

Efron [2007] remarque également ce phénomène sur deux jeux de données réelles issues d'expériences par biopuces. Il attribue directement l'écart entre les Z-scores et la loi théorique nulle (histogramme des Z-scores plus large ou moins large que la densité de la loi Normale centrée-réduite) à la présence de corrélation dans les données.

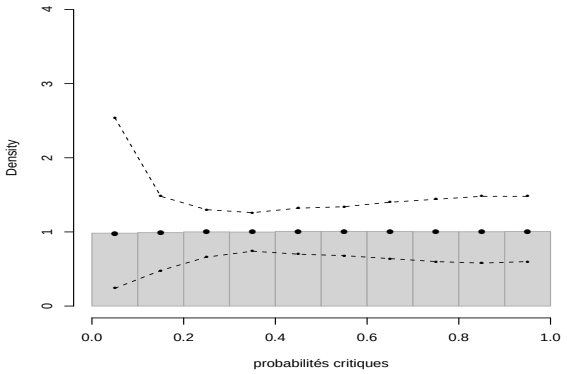
On peut tester l'uniformité des m probabilités critiques, pour les 1000 tableaux de donénes de chaque scénarios à l'aide d'un test de Kolmogorov-Smirnov. La FIGURE 2.4 représente les probabilités critiques de ces tests. Elles sont elles même réparties uniformément (test d'uniformité sur ces probabilités critiques non significatif au seuil $0,05$) pour les scénarios 0 et 1. Pour les autres scénarios, l'écart à



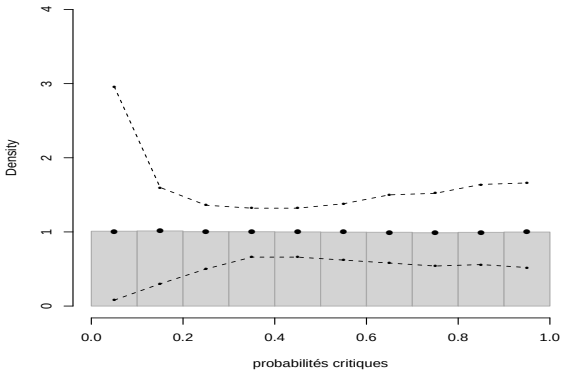
(a) Scénario 1



(b) Scénario 3

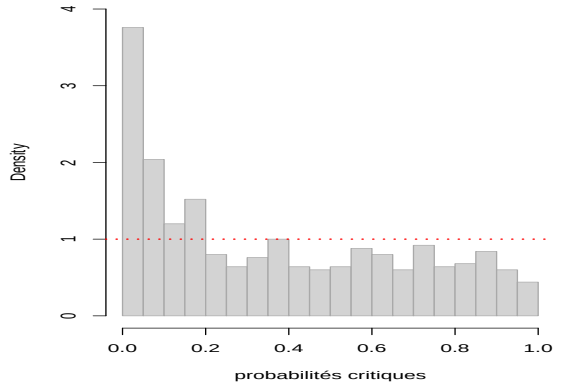


(c) Scénario 6



(d) Scénario 9

FIG. 2.2 – Distribution des probabilités critiques - Histogramme moyen sur l'ensemble des 1000 simulations - lignes pointillées : quantiles à $2,5\%$ et $97,5\%$



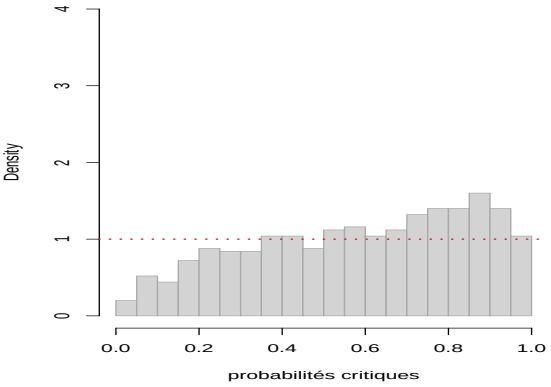


FIG. 2.3 – Distribution des probabilités critiques sous H_0 : exemples de deux jeux de données - scénario 9

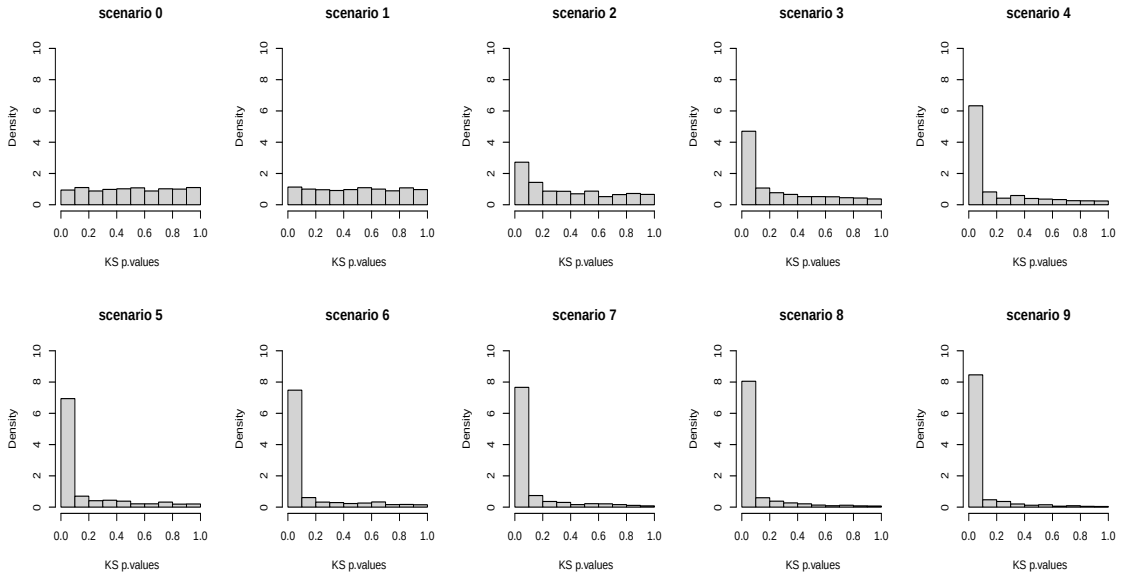


FIG. 2.4 – Distribution des probabilités critiques des tests de Kolmogorov-Smirnov obtenues pour les tests d'uniformité des m probabilités critiques des 1000 jeux de données simulés de chaque scénarios

Scenario	0	1	2	3	4	5	6	7	8	9
prop. de tests sig. (%)	4.4	7.1	20.2	38.0	56.1	63.8	68.5	71.6	75.5	80.1

TAB. 2.2 – Proportion de tests de l'uniformatité des probabilités critiques déclarés significatifs par scénario (seuil : $\alpha = 0.05$) - Tests de kolmogorov-Smirnov pour chaque ensemble de m probabilités critiques obtenus pour chaque tableau de données simulé

l'uniformité des probabiliéts critiques des tests de Kolmogorov Smirnov est mis en évidence de façon significative. Le TABLEAU 2.1 résume le nombre de tests déclarés significatifs, à un niveau $\alpha = 0,05$.

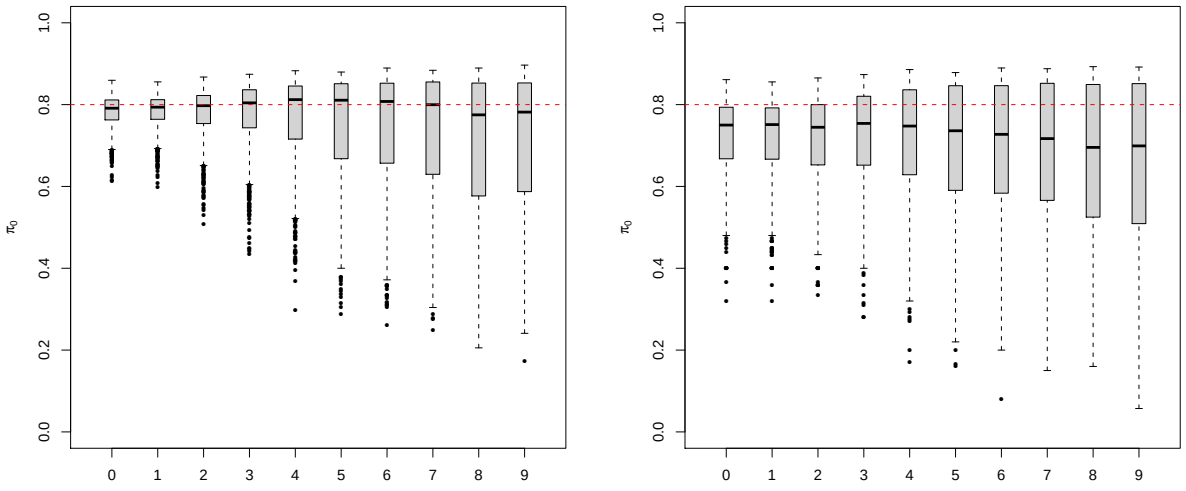
L'hypothèse d'uniformité de la distribution des probabilités critiques sous H_0 est le support de nombreuses procédures de tests multiples. Or, les deux exemples présentés sur la FIGURE 2.3 illustrent qu'en présence de dépendance, la distribution empirique des probabilités critiques peut s'éloigner fortement de cette distribution théorique, rendant alors plus complexe l'identification de la composante nulle du modèle de mélange.

2.2. Étude de l'impact de la dépendance sur l'estimation de la proportion d'hypothèses nulles (π_0)

Les simulations de l'EXEMPLE 3 montrent de grandes disparités sur la répartition des probabilités critiques sous l'hypothèse nulle en présence de dépendance, en particulier sur les bords de l'intervalle $[0; 1]$. Or, la séparation entre les deux composantes du modèle de mélange, c'est-à-dire la dominance

d'une composante au voisinage de 0 et de l'autre composante au voisinage de 1, est une condition nécessaire pour une bonne estimation de π_0 (voir SECTION 1.2). Les méthodes d'estimation de π_0 sont en effet toutes basées sur le comportement de la distribution des probabilités critiques aux alentours de 1.

EXEMPLE 4. Reprenons les données de l'EXEMPLE 3. Pour $m_1 = 100$ variables, une quantité δ est ajoutée pour les individus du groupe B. La valeur de δ est fixée de façon à assurer au test individuel de Student une puissance de 80%, pour un risque de type-I $\alpha = 5\%$, soit ici $\delta = 0.74$. La valeur théorique de π_0 est de $0,80$. Au sein de chaque scénario, pour chaque tableau de données simulé, on réalise les $m = 500$ tests de l'effet de la variable x.



(a) Estimation de la densité par une fonction convexe (b) Estimateur empirique avec choix de λ par *bootstrap*

FIG. 2.5 – Estimations de π_0 à partir de probabilités critiques issues de tests de Student sur des données simulées selon différents scénarios de dépendance - $\pi_0 = 0,80$

La FIGURE 2.5 montre les boites de dispersion de $\hat{\pi}_0$ pour des niveaux de dépendance croissants. Les résultats présenté sur la FIGURE 2.5 concernent deux méthodes : estimation de la densité par une fonction convexe décroissante sur $[0;1]$ [Langaas et al., 2005] et l'estimateur empirique avec choix de λ par *bootstrap* [Storey et al., 2004] pour chacun des scénarios. Les résultats pour les autres méthodes sont présenté en ANNEXE D, FIGURE D.1. Pour l'ensemble des méthodes, la variabilité d'estimation augmente nettement en présence de dépendance.

Nous introduisons un critère noté Δ définit en (2.4). Ce critère est calculé en faisant la différence, sur l'intervalle $[0;0,05]$, entre la densité empirique des probabilités critiques sous H_0 (histogramme empirique) et la densité nulle théorique sur ce même intervalle. Il caractérise la forme de l'histogramme des probabilités critiques : Δ est nul si l'histogramme est constant, et varie autour de 0

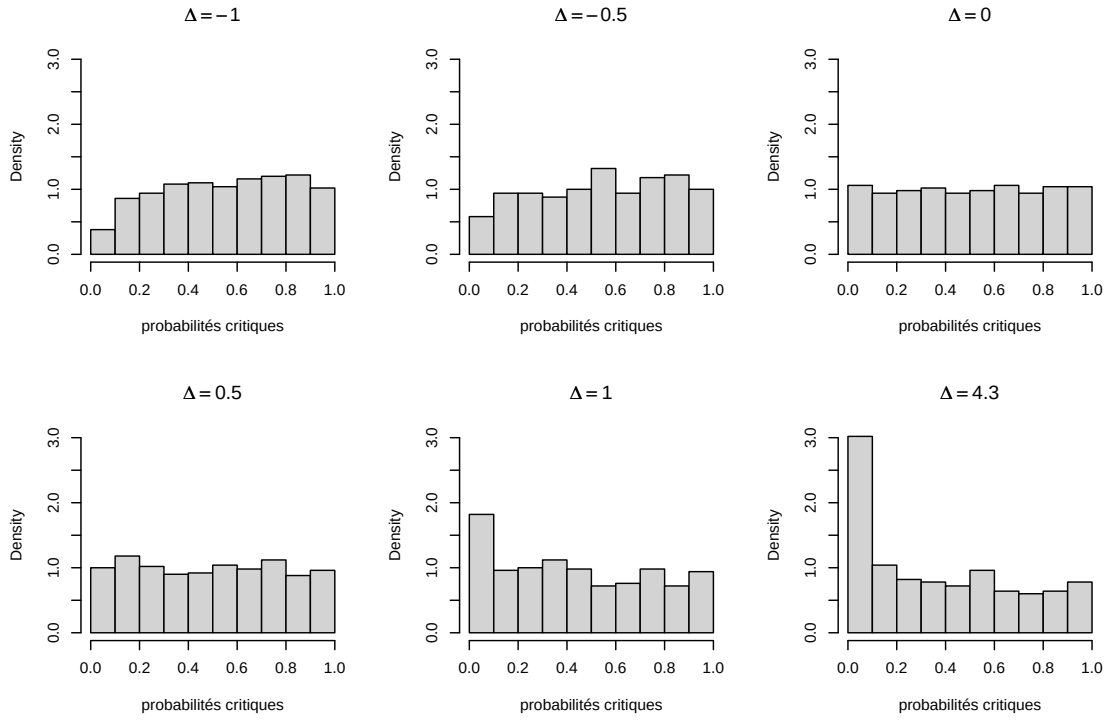


FIG. 2.6 – Exemple d'histogrammes de probabilités critiques sous H_0 pour 6 valeurs de Δ

en cas d'écarts. Il est positif lorsque les probabilités critiques proches de 0 sont sur-représentées et négatif lorsque les probabilités critiques proches de 1 sont sur-représentées (FIGURE 2.6). La FIGURE 2.7 représente la distribution de Δ pour les données simulées pour les 10 scénarios de dépendance.

$$\Delta \quad = \quad \{\#p_k \in [0;0,05]\} \times \frac{0,05}{card(\mathcal{M})} - 1 \tag{2.4}$$

En fait, si en présence de dépendance on observe une augmentation de la variabilité de l'estimation du paramètre (FIGURE 2.5), la dépendance induit surtout une sur- ou sous-estimation de π_0 en fonction de la configuration de l'histogramme (FIGURE 2.8 pour les deux mêmes méthodes que précédemment. Les résultats pour les autres méthodes sont présentés en ANNEXE D, FIGURE D.2).

L'estimation de π_0, quelle que soit la méthode utilisée, est donc affectée par la présence de dépendance dans les données.

Étude de l'estimateur empirique $\hat{\pi}_0(\lambda)$ en présence de dépendance Pour cette étude, nous introduisons la distribution conditionnelle des probabilités critiques : $G^k(t, Z) = \mathbb{P}(p_k \leq t|Z)$. Les propriétés de cette distributions sont données ci-après :

PROPOSITION 2.2.1 (Propriétés de $G^k(t, Z) = \mathbb{P}(p_k \leq t|Z)$)**.** *Soit* $G^k(t) = \mathbb{P}(p_k \leq t)$ *et* $G^{kk'}(t) =$

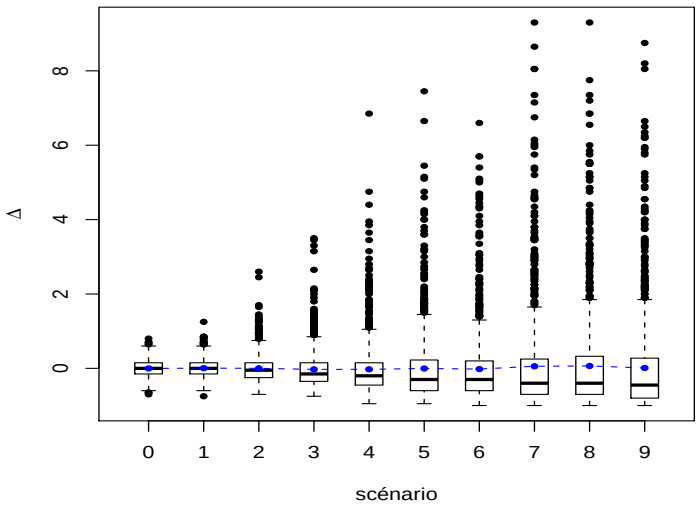


FIG. 2.7 – Valeurs de Δ pour les 10 scénarios de simulations

$\mathbb{P}(p_k \leq t; p_{k'} \leq t)$.

$$\begin{aligned} \mathbb{E}(G^k(t,Z)) &= G^k(t) \\ \mathrm{Cov}(G^k(t,Z); G^{k'}(t,Z)) &= G^{kk'}(t) - G^k(t)G^{k'}(t) \end{aligned}$$

En particulier, si $k \in \mathcal{M}_0$ *:* $\mathbb{E}(G^k(t,Z)) = G_0(t) = t$

Démonstration. On a : $G^k(t,Z) = \mathbb{P}(p_k \leq t|Z) = \mathbb{E}(\mathbb{1}_{p_k \leq t}|Z)$.

D'où $\mathbb{E}(G^k(t,Z)) = \mathbb{E}(\mathbb{1}_{p_k \leq t}) = G^k(t)$ □

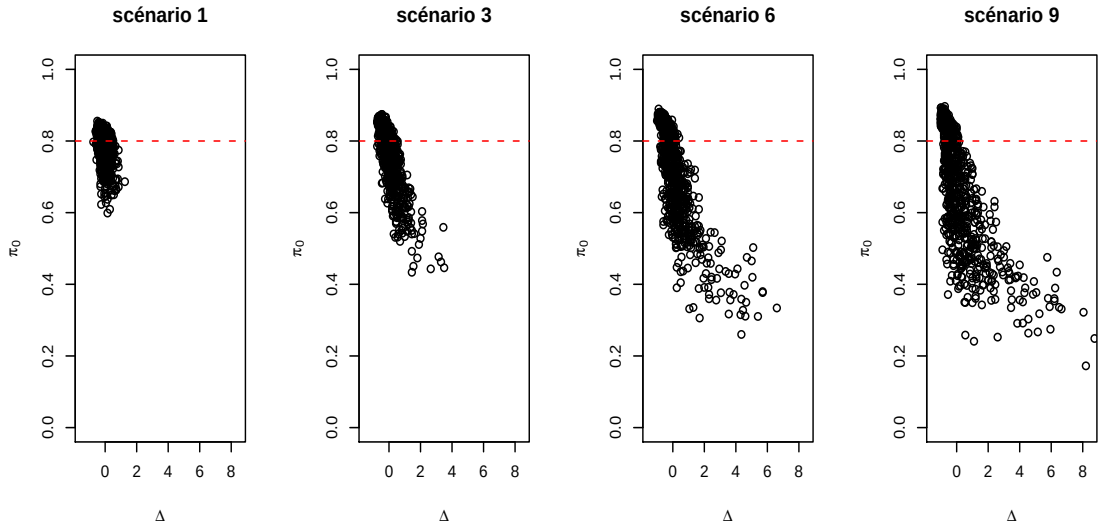
On définit également la fonction $D^{kk'}(t)$ par l'expression suivante :

DÉFINITION 2.2.1 (Fonction $D^{kk'}(t)$)**.**

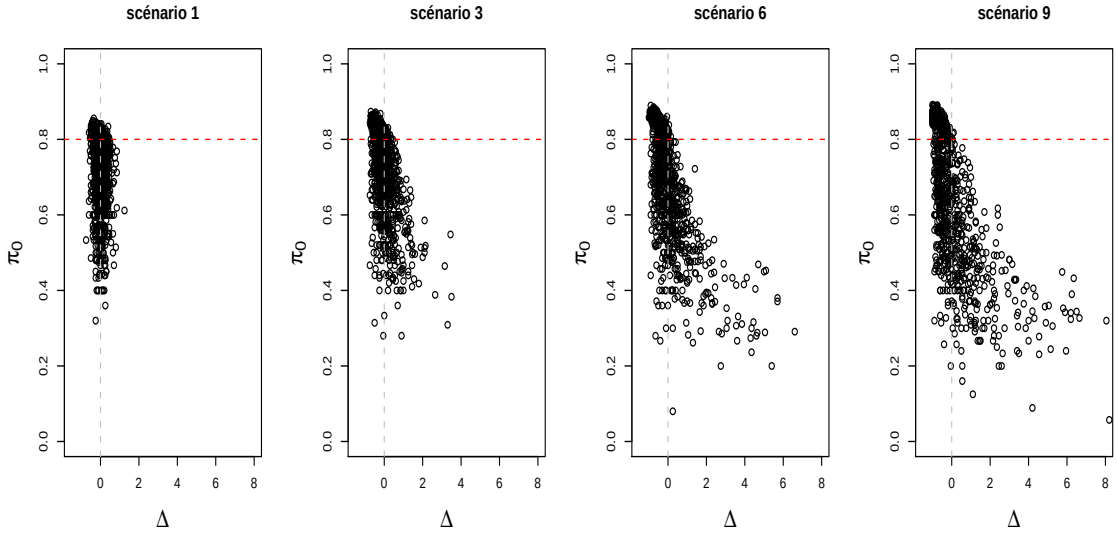
$$D^{kk'}(t) = \frac{G^{kk'}(t) - G^k(t)G^{k'}(t)}{t(1-t)}$$

Dans le TABLEAU 1.1, pour un seuil de rejet des probabilités critiques λ, U_λ désigne le nombre de vrai-négatifs : $U_\lambda = \sum_{k \in \mathcal{M}_0} \mathbb{1}_{(p_k > \lambda)}$, et T_λ est le nombre de faux-négatifs : $T_\lambda = \sum_{k \in \mathcal{M}_1} \mathbb{1}_{(p_k > \lambda)}$.On a $W_\lambda = U_\lambda + T_\lambda$.

Sous l'HYPOTHÈSE 1, $U_\lambda \sim \mathcal{B}in(m_0, 1-\lambda)$ et $T_\lambda \sim \mathcal{B}in(m_1, 1-G_1(\lambda))$, ce qui permet d'obtenir les propriétés de $\hat{\pi}_0$ données par la PROPOSITION 1.2.1.



(a) Estimation de la densité par une fonction convexe



(b) Estimateur empirique avec choix de λ par *bootstrap*

FIG. 2.8 – Estimations de π_0 à partir de probabilités critiques issues de tests de Student sur des données simulées selon différents scénarios de dépendance en fonction de Δ - $\pi_0 = 0,80$

Quand les données sont corrélées, sous l'hypothèse du modèle (2.1), on s'appuie sur l'indépendance conditionnelle aux facteurs Z pour étudier les propriétés de $\hat{\pi}_0$.

PROPOSITION 2.2.2 (Propriétés de $\hat{\pi}_0$). *$G_1^k(\lambda)$ désigne la fonction de répartition de la probabilité critique p_k pour $k \in \mathcal{M}_1$ et on note $\bar{G}_1(\lambda) = \frac{1}{m_1}\sum_{k\in\mathcal{M}_1}\left(G_1^k(\lambda)\right)$.*

$$\mathbb{E}(\hat{\pi}_0(\lambda)) = \pi_0 + (1-\pi_0)\frac{1-\bar{G}_1(\lambda)}{1-\lambda}$$

$$\mathbb{V}(\hat{\pi}_0(\lambda)) = \frac{\lambda\pi_0}{m(1-\lambda)} + \frac{\sum_{k\in\mathcal{M}_1}\left[G_1(\lambda,k)(1-G_1(\lambda,k))\right]}{m^2(1-\lambda)^2} + \frac{\lambda}{1-\lambda}\frac{1}{m^2}\sum_{k\neq k'\in\mathcal{M}} D^{kk'}(\lambda)$$

La démonstration de la PROPOSITION 2.2.2 passe par les lemmes suivants.

LEMME 2.2.1 (Propriétés de U_λ)**.**

$$\begin{aligned}\mathbb{E}(U_\lambda) &= m_0(1-\lambda)\\ \mathbb{V}(U_\lambda) &= \Big[m_0 + \sum_{k\neq k'\in\mathcal{M}_0} D^{kk'}(\lambda)\Big]\lambda(1-\lambda)\end{aligned}$$

Démonstration. On s'intéresse d'abord aux propriétés conditionnelles de U_λ.

$$\begin{aligned}\mathbb{E}(U_\lambda|Z) &= \sum_{k\in\mathcal{M}_0}\mathbb{P}\left(p_k\geq\lambda|Z\right) = \sum_{k\in\mathcal{M}_0}\Big(1-G^k(\lambda,Z)\Big)\\ \mathbb{V}(U_\lambda|Z) &= \sum_{k\in\mathcal{M}_0}\mathbb{P}\left(p_k\geq\lambda|Z\right)\left(1-\mathbb{P}\left(p_k\geq\lambda|Z\right)\right) = \sum_{k\in\mathcal{M}_0}\Big(1-G^k(\lambda,Z)\Big)G^k(\lambda,Z)\end{aligned}$$

Or, $card(\mathcal{M}_0) = m_0$ et $\mathbb{E}(G^k(\lambda,Z)) = G^k(\lambda) = \lambda$ pour $k\in\mathcal{M}_0$ d'après la PROPOSITION 2.2.1. On déduit donc que $\mathbb{E}(U_\lambda) = \mathbb{E}\left(\mathbb{E}(U_\lambda|Z)\right) = m_0(1-\lambda)$.

La variance de U_λ est déduite en utilisant : $\mathbb{V}(U_\lambda) = \mathbb{E}\left(\mathbb{V}(U_\lambda|Z)\right) + \mathbb{V}\left(\mathbb{E}(U_\lambda|Z)\right)$. On a :

$$\begin{aligned}\mathbb{V}[\mathbb{E}(U_\lambda|Z)] &= \mathbb{V}\left[\sum_{k\in\mathcal{M}_0}(1-G^k(\lambda,Z))\right] = \mathbb{V}\left[\sum_{k\in\mathcal{M}_0}(G^k(\lambda,Z))\right] = \sum_{k\in\mathcal{M}_0}\mathbb{V}\Big(G^k(\lambda,Z)\Big) + \sum_{k\neq k'\in\mathcal{M}_0}\mathrm{Cov}\Big(G^k(\lambda,Z);G^{k'}(\lambda,Z)\Big)\\ &= \sum_{k\in\mathcal{M}_0}\mathbb{E}\Big(G^k(\lambda,Z)^2\Big) - \sum_{k\in\mathcal{M}_0}\mathbb{E}\Big(G^k(\lambda,Z)\Big)^2 + \sum_{k\neq k'\in\mathcal{M}_0}\mathrm{Cov}\Big(G^k(\lambda,Z);G^{k'}(\lambda,Z)\Big)\\ &= \sum_{k\in\mathcal{M}_0}\mathbb{E}\Big(G^k(\lambda,Z)^2\Big) - m_0\lambda^2 + \sum_{k\neq k'\in\mathcal{M}_0}\mathrm{Cov}\Big(G^k(\lambda,Z);G^{k'}(\lambda,Z)\Big)\\ \mathbb{E}[\mathbb{V}(U_\lambda|Z)] &= \mathbb{E}\left[\sum_{k\in\mathcal{M}_0}\Big(1-G^k(\lambda,Z)\Big)G^k(\lambda,Z)\right]\\ &= \sum_{k\in\mathcal{M}_0}\mathbb{E}\Big(G^k(\lambda,Z)\Big) - \sum_{k\in\mathcal{M}_0}\mathbb{E}\Big(G^k(\lambda,Z)^2\Big)\\ &= m_0\lambda - \sum_{k\in\mathcal{M}_0}\mathbb{E}\Big(G^k(\lambda,Z)^2\Big)\end{aligned}$$

On en déduit la variance de U_λ, en introduisant la fonction $D^{kk'}(\lambda)$:

$$\begin{aligned}\Rightarrow\mathbb{V}(U_\lambda) &= m_0\lambda(1-\lambda) + \sum_{k\neq k'\in\mathcal{M}_0}\mathrm{Cov}\Big(G^k(\lambda,Z);G^{k'}(\lambda,Z)\Big)\\ &= \Big[m_0 + \sum_{k\neq k'\in\mathcal{M}_0}D^{kk'}(\lambda)\Big]\lambda(1-\lambda) \qquad (2.5)\end{aligned}$$

□

Lemme 2.2.2 (Propriétés de T_λ). *On note* $\bar{G}_1(\lambda) = \frac{1}{m_1}\sum_{k\in\mathcal{M}_1}\left(G_1^k(\lambda)\right)$

$$\begin{aligned}\mathbb{E}(T_\lambda) &= (m-m_0)(1-\bar{G}_1^k(\lambda))\\ \mathbb{V}(T_\lambda) &= \sum_{k\in\mathcal{M}_1}\Big[G_1^k(\lambda)(1-G_1^k(\lambda))\Big] + \Big[\sum_{k\neq k'\in\mathcal{M}_1} D^{kk'}(\lambda)\Big]\lambda(1-\lambda)\end{aligned}$$

Démonstration. On s'intéresse d'abord aux propriétés conditionnelles de T_λ.

$$\begin{aligned}\mathbb{E}(T_\lambda|Z) &= \sum_{k\in\mathcal{M}_1}\mathbb{P}\left(p_k\geq\lambda|Z\right) = \sum_{k\in\mathcal{M}_1}\Big(1-G^k(\lambda,Z)\Big)\\ \mathbb{V}(T_\lambda|Z) &= \sum_{k\in\mathcal{M}_1}\mathbb{P}\left(p_k\geq\lambda|Z\right)\left(1-\mathbb{P}\left(p_k\geq\lambda|Z\right)\right) = \sum_{k\in\mathcal{M}_1}(1-G^k(\lambda,Z))G^k(\lambda,Z)\end{aligned}$$

Or, $card(\mathcal{M}_1) = m-m_0$ et $\mathbb{E}(G^k(\lambda,Z)) = G_1^k(\lambda)$ pour $k\in\mathcal{M}_1$ d'après la proposition 2.2.1. On note $\bar{G}_1(\lambda) = \frac{1}{m_1}\sum_{k\in\mathcal{M}_1}(G_1(\lambda,k))$. On déduit donc que $\mathbb{E}(T_\lambda) = \mathbb{E}\left(\mathbb{E}(T_\lambda|Z)\right) = (m-m_0)(1-\bar{G}_1(\lambda,k))$.
La variance de T_λ est déduite en utilisant : $\mathbb{V}(T_\lambda) = \mathbb{E}\left(\mathbb{V}(T_\lambda|Z)\right) + \mathbb{V}\left(\mathbb{E}(T_\lambda|Z)\right)$. On a :

$$\begin{aligned}\mathbb{V}[\mathbb{E}(T_\lambda|Z)] &= \mathbb{V}\left[\sum_{k\in\mathcal{M}_1}(1-G^k(\lambda,Z))\right]\mathbb{V}\left[\sum_{k\in\mathcal{M}_1}(G^k(\lambda,Z))\right] = \sum_{k\in\mathcal{M}_1}\mathbb{V}\Big(G^k(\lambda,Z)\Big) + \sum_{k\neq k'\in\mathcal{M}_1}\mathrm{Cov}\Big(G^k(\lambda,Z));G^{k'}(\lambda,Z)\Big)\\ &= \sum_{k\in\mathcal{M}_1}\mathbb{E}\Big(G^k(\lambda,Z)^2\Big) - \sum_{k\in\mathcal{M}_1}\mathbb{E}\Big(G^k(\lambda,Z)\Big)^2 + \sum_{k\neq k'\in\mathcal{M}_1}\mathrm{Cov}\Big(G^k(\lambda,Z);G^{k'}(\lambda,Z)\Big)\\ \mathbb{E}[\mathbb{V}(T_\lambda|Z)] &= \mathbb{E}\left[\sum_{k\in\mathcal{M}_1}\Big(1-G^k(\lambda,Z)\Big)G^k(\lambda,Z)\right]\\ &= \sum_{k\in\mathcal{M}_1}\mathbb{E}\Big(G^k(\lambda,Z)\Big) - \sum_{k\in\mathcal{M}_1}\mathbb{E}\Big(G^k(\lambda,Z)^2\Big)\end{aligned}$$

On en déduit la variance de T_λ, en introduisant la fonction $D^{kk'}(\lambda)$:

$$\begin{aligned}\Rightarrow\mathbb{V}(T_\lambda) &= \sum_{k\in\mathcal{M}_1}\left[\mathbb{E}\Big(G^k(\lambda,Z)\Big) - \mathbb{E}\Big(G^k(\lambda,Z)\Big)^2\right] + \sum_{k\neq k'\in\mathcal{M}_1}\mathrm{Cov}\Big(G^k(\lambda,Z);G^k(\lambda,Z)\Big)\\ &= \sum_{k\in\mathcal{M}_1}\Big[G_1^k(\lambda)(1-G_1^k(\lambda))\Big] + \sum_{k\neq k'\in\mathcal{M}_1}\Big[G^{kk'}(\lambda) - G^k(\lambda)G^{k'}(\lambda)\Big]\\ &= \sum_{k\in\mathcal{M}_1}\Big[G_1^k(\lambda)(1-G_1^k(\lambda))\Big] + \Big[\sum_{k\neq k'\in\mathcal{M}_0}D^{kk'}(\lambda)\Big]\lambda(1-\lambda) \qquad (2.6)\end{aligned}$$

□

On peut alors démontrer la proposition 2.2.2, en se servant des lemmes 2.2.1 et 2.2.2.

Démonstration. On s'intéresse d'abord à l'espérance de $\hat{\pi}_0$.
En notant $\bar{G}_1(\lambda) = \frac{1}{m_1}\sum_{k\in\mathcal{M}_1}(G_1(\lambda,_k))$, on en déduit l'espérance de $\hat{\pi}_0$.

$$\begin{aligned}\mathbb{E}(\hat{\pi}_0(\lambda)) &= \frac{\mathbb{E}(W_\lambda)}{m(1-\lambda)} = \frac{\mathbb{E}(U_\lambda)+\mathbb{E}(T_\lambda)}{m(1-\lambda)}\\ &= \frac{m_0(1-\lambda)+m_1(1-\bar{G}_1(\lambda))}{m(1-\lambda)} = \pi_0 + (1-\pi_0)\frac{(1-\bar{G}_1(\lambda))}{(1-\lambda)}\end{aligned}$$

On s'intéresse maintenant à la variance de $\hat{\pi}_0$. On utilise : $\mathbb{V}(W_\lambda) = \mathbb{V}(U_\lambda) + \mathbb{V}(T_\lambda) + 2\,\mathrm{Cov}(U_\lambda; T_\lambda)$, où $\mathbb{V}(U_\lambda)$ et $\mathbb{V}(T_\lambda)$ sont donnés dans les LEMMES 2.2.1 et 2.2.2. Il reste donc à calculer $\mathrm{Cov}(U_\lambda; T_\lambda)$. On se sert de $\mathrm{Cov}(U_\lambda; T_\lambda) = \mathbb{E}\left(\mathrm{Cov}(U_\lambda; T_\lambda|Z)\right) + \mathrm{Cov}(\mathbb{E}(U_\lambda|Z); \mathbb{E}(T_\lambda|Z))$. D'après la propriété d'indépendance conditionnelle, $\mathrm{Cov}(U_\lambda; T_\lambda|Z) = 0$.

$$\begin{aligned}
\mathrm{Cov}(\mathbb{E}(U_\lambda|Z); \mathbb{E}(T_\lambda|Z)) &= \mathrm{Cov}\left(\sum_{k\in\mathcal{M}_0}(1-G^k(\lambda,Z)); \sum_{k\in\mathcal{M}_1}(1-G^k(\lambda,Z))\right) \\
&= \sum_{k\in\mathcal{M}_0}\sum_{k'\in\mathcal{M}_1}\mathrm{Cov}(G^k(\lambda,Z); G^{k'}(\lambda,Z)) \\
&= \sum_{k\in\mathcal{M}_0}\sum_{k'\in\mathcal{M}_1}\left[G^{kk'}(\lambda) - G^k(\lambda,Z)G^{k'}(\lambda,Z)\right] \\
&= \sum_{k\in\mathcal{M}_0}\sum_{k'\in\mathcal{M}_1}\left[D^{kk'}(\lambda)\right]\lambda(1-\lambda)
\end{aligned}$$

On a alors la variance de $\hat{\pi}_0(\lambda)$:

$$\mathbb{V}(\hat{\pi}_0(\lambda)) = \frac{\lambda\pi_0}{m(1-\lambda)} + \frac{\sum_{k\in\mathcal{M}_1}\left[G_1(\lambda,k)(1-G_1(\lambda,k))\right]}{m^2(1-\lambda)^2} + \frac{\lambda}{1-\lambda}\frac{1}{m^2}\sum_{k\neq k'\in\mathcal{M}} D^{kk'}(\lambda)$$

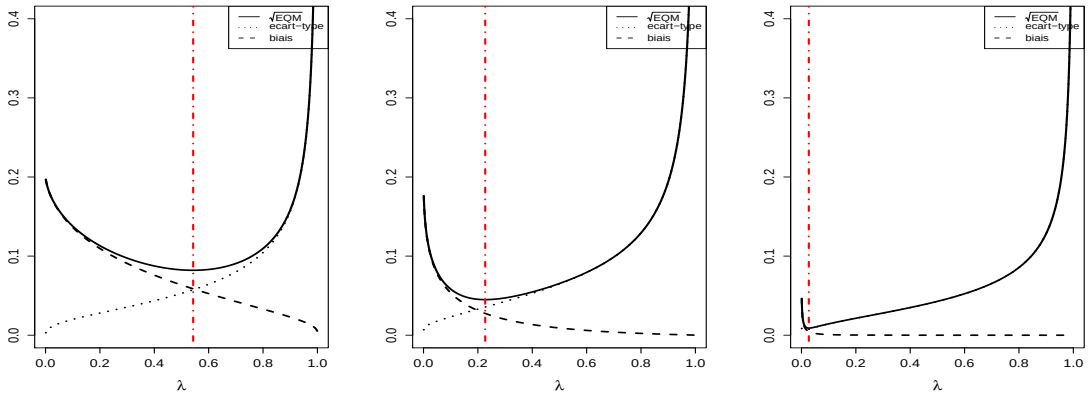
□

Cette expression de la variance de l'estimateur empirique se réduit à la formule de la PROPOSITION 1.2.1 dans le cadre de l'HYPOTHÈSE 1. Sous cette hypothèse, les probabilités critiques sont identiquement distribuées sous H_1 donc $\sum_{k\in\mathcal{M}_1}\left[G_1(\lambda,k)(1-G_1(\lambda,k))\right] = (m-m_0)G_1(\lambda,k)(1-G_1(\lambda,k)$ et le dernier terme est nul dans le cas de l'indépendance. La fonction $D^{kk'}(\lambda)$ prend en effet une valeur nulle si la corrélation entre les variables Y_k et $Y_{k'}$ est nulle.

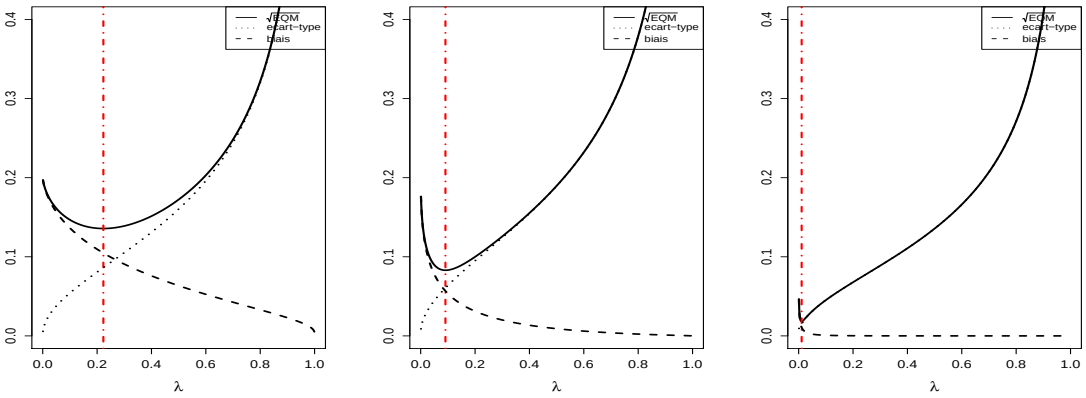
La PROPOSITION 2.2.2 permet d'étudier l'impact de la dépendance sur l'estimation de π_0, et en particulier, dans quelle mesure la variance de l'estimateur de ce paramètre dépend de la corrélation entre les variables.

Choix de λ par minimisation de l'EQM Pour illustrer l'impact de la corrélation sur le choix du seuil λ optimal, nous calculons les valeurs du biais, de la variance et de l'EQM de $\hat{\pi}_0(\lambda)$ pour différents scénarios de corrélation à partir des matrices de variances-covariances Σ_s utilisées dans le cadre théorique de l'EXEMPLE 4.

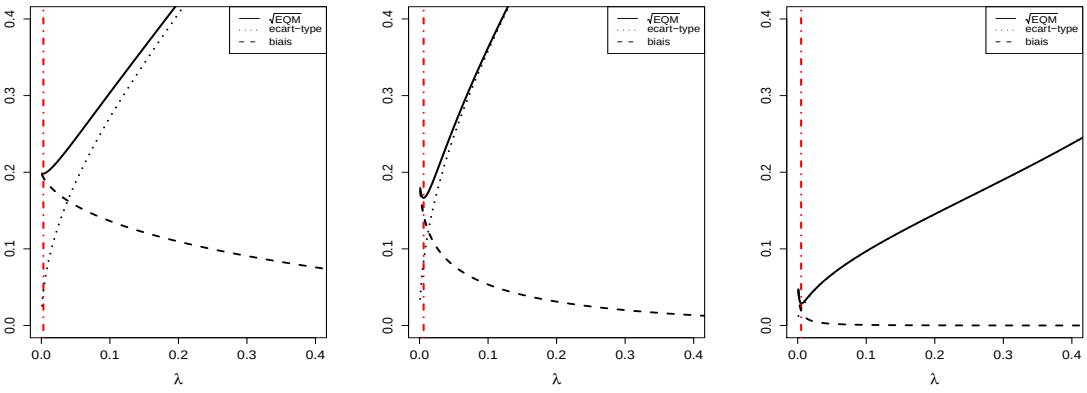
Dans un cadre non-paramétrique, la probabilité bivariée notée $G_{kk'}$ apparaissant au numérateur de la fonction $D(t)$, dans la formule de la variance donnée à la PROPOSITION 2.2.2, peut être calculée à partir de permutations. Dans le cas des tests de Student de l'EXEMPLE 4, nous pouvons utiliser l'expression exacte de cette fonction, définie ci-après.



(a) Scénario 1



(b) Scénario 4



(c) Scénario 8

FIG. 2.9 – Biais, variance et EQM de $\hat{\pi}_0(\lambda)$: courbes théoriques obtenues à partir des matrices de variances covariances utilisées pour la simlation des données de chacun des scénarios - Niveau de dépendance : 1 : niveau faible, 4 : niveau intermédiaire, 8 : niveau élevé - à gauche : $\tau = 1$ (puissance : 17%), au milieu : $\tau = 2,8$ (puissance : 80%), à droite : $\tau = 4$ (puissance : 97%) - $\pi_0 = 0,80$

PROPOSITION 2.2.3. *On note $\Phi^{(2)}(J;\rho,\tau_1,\tau_2,ddl)$ la probabilité qu'un vecteur, suivant une loi de Student bivariée de paramètres de non-centralité $(\tau_k;\tau_{k'})$, de degrés de liberté ddl et de corrélation ρ n'appartienne pas à $J \subseteq \mathbb{R}^2$. On a alors $G_{kk'} = \Phi^{(2)}([-u_t;u_t];\rho_{kk'},\tau_k,\tau_{k'},n-2)$, où $\rho_{kk'}$ la corrélation entre T_k et $T_{k'}$ et $u_t = \phi_0^{-1}\left(1-\frac{t}{2}\right)$, avec ϕ_0 la densité de la loi de Student centrée $(\tau = 0)$ à $ddl = n-2$ degrés de liberté. On définit par ailleurs $\Phi^{(1)}(I,\tau,ddl)$ qui représente la probabilité qu'une variable aléatoire distribuée selon une loi de Student de paramètre de non centralité τ et à ddl degrés de liberté n'appartienne pas à l'intervalle $I \subseteq \mathbb{R}$.*

$$\begin{aligned} D^{kk'}(t) &= \frac{1}{t(1-t)}\Big[\Phi^{(2)}\Big([-u_t;u_t]^2;\rho_{kk'};\tau_k;\tau_{k'};n-2\Big) \\ &\quad -\Phi^{(1)}\left([-u_t;u_t];\tau_k,n-2\right)\Phi^{(1)}\left([-u_t;u_t];\tau_{k'};n-2\right)\Big] \end{aligned}$$

La FIGURE 2.9 donne les résultats pour trois des scénarios (1 : niveau faible, 4 : niveau intermédiaire, 8 : niveau élevé), et pour différentes valeurs de τ, caractérisant différents niveau de puissance des tests individuels. La ligne pointillée rouge donne le minimum de l'EQM pour chaque cas. Lorsque τ est élevé (figures de droite), le choix de lambda est similaire quel que soit le niveau de dépendance, et l'EQM associée est faible. Par contre, quand ce paramètre est plus faible (figures de gauche et au milieu), ignorer la dépendance induit une surestimation de λ. L'EQM associée à une telle valeur de λ pour les scénarios 4 et 8 est beaucoup plus élevée, et la variance de $\hat{\pi}_0$ est largement sous-estimée.

2.3. Étude de l'impact de la dépendance sur les taux d'erreurs

La FIGURE 2.10 reprend les deux exemples de la FIGURE 2.3, la ligne pointillée indiquant un seuil de $0,05$ pour les probabilités critiques. En déclarant positifs les tests associés aux probabilités critiques inférieures à ce seuil, on sur-estime (à gauche) ou sous-estime (à droite) sensiblement le nombre attendu de faux-positifs.

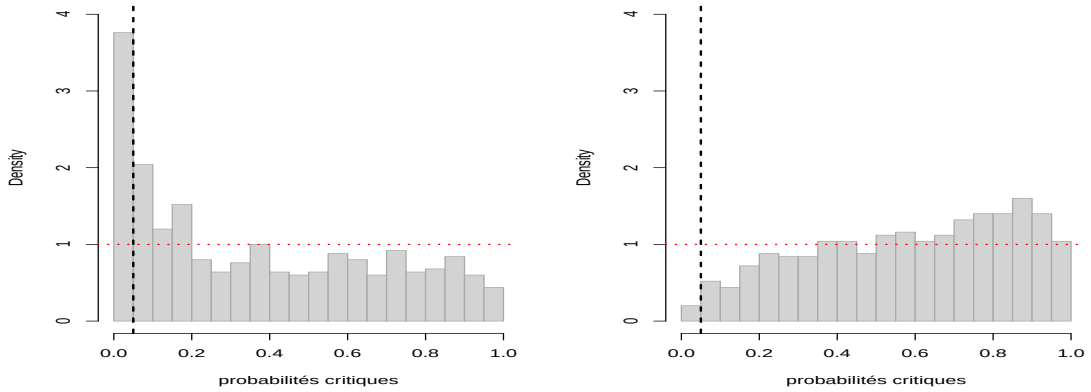


FIG. 2.10 – Exemples de deux jeux de données : distribution des probabilités critiques - scénario 9 - EXEMPLE 3. En pointillé : seuil de $0,05$ pour les probabilités critiques

2.3.1 Impact de la dépendance sur le nombre de faux-positifs (V_t)

Pour chaque tableau de données simulées (EXEMPLE 4), la quantité V_t, c'est à dire le nombre de rejets à tort d'hypothèses nulles, est calculée en considérant un seuil de rejet $t = 0,05$ pour les probabilités critiques.

Le TABLEAU 2.3 donne les statistiques descriptives des résultats.

scénario	0	1	2	3	4	5	6	7	8	9
var. commune (%)	0	4,65	15,56	28,56	40,31	50,76	57,90	65,45	70,09	75,19
min	6.00	5.00	6.00	5.00	1.00	1.00	0.00	0.00	0.00	0.00
$q_{0,25}$	17.00	17.00	15.00	13.00	11.00	8.00	8.00	6.00	6.00	4.00
médiane	20.00	20.00	19.00	17.00	16.00	14.00	14.00	12.00	12.00	11.00
moyenne	**19.96**	**20.12**	**20.03**	**19.44**	**19.48**	**19.94**	**19.63**	**21.12**	**21.29**	**20.22**
$q_{0,75}$	23.00	23.00	23.00	23.00	23.00	24.25	24.00	25.00	26.25	25.25
max	36.00	45.00	72.00	90.00	157.00	169.00	152.00	206.00	206.00	195.00
écart-type	**4.43**	**4.96**	**7.33**	**10.43**	**14.26**	**18.73**	**19.83**	**26.19**	**26.14**	**26.35**

TAB. 2.3 – Statistiques descriptives de V_t pour les 10 scénarios de données simulées

Si on suppose que les probabilités critiques sont distribuée selon le modèle (1.4), on a sous l'HYPOTHÈSE 1 : $V_t \sim \mathcal{B}in(m_0, t)$. L'espérance d'une telle variable aléatoire est $m_0 t = 400 \times 0,05 = 20$ et sa variance $m_0 t(1-t) = 400 \times 0,05 \times 0,95 = 19$. Pour l'ensemble des scénarios, d'après le TABLEAU 2.3, la moyenne du nombre de faux-positifs est proche de 20 hypothèses rejetées à tort. L'espérance de la variable aléatoire V_t ne semble donc pas affectée par la présence de dépendance.

Pour les scénarios 0 et 1, l'écart-type de V_t a bien une valeur semblable à la valeur théorique ($\sqrt{19} = 4,35$). Mais à mesure que le niveau de dépendance augmente, la variance du nombre de faux-positifs augmente également. La dépendance entraine donc une instabilité du nombre de faux-positifs, en augmentant sa variabilité.

La FIGURE 2.11 représente sous forme d'histogrammes la distribution de V_t pour quatre des 10 scénarios, caractérisés par des niveaux croissants de dépendance, ce qui illustre les commentaires faits précédemment à partir du TABLEAU 2.3.

Nous étudions maintenant les propriétés du nombre de faux-positifs en présence de corrélation.

PROPOSITION 2.3.1 (Propriétés de V_t)**.**

$$\begin{aligned} \mathbb{E}(V_t) &= m_0 t \\ \mathbb{V}(V_t) &= \Big[m_0 + \sum_{k \neq k' \in \mathcal{M}_0} D^{kk'}(t)\Big] t(1-t) \end{aligned}$$

Démonstration. On se sert de la démonstration du LEMME 2.2.1 car $V_t = m_0 - U_t$. D'où $\mathbb{E}(V_t) = m_0 - m_0(1-t) = m_0 t$ et $\mathbb{V}(V_t) = \mathbb{V}(m_0 - U_t) = \mathbb{V}(U_t) = \Big[m_0 + \sum_{k \neq k' \in \mathcal{M}_0} D^{kk'}(t)\Big] t(1-t)$. □

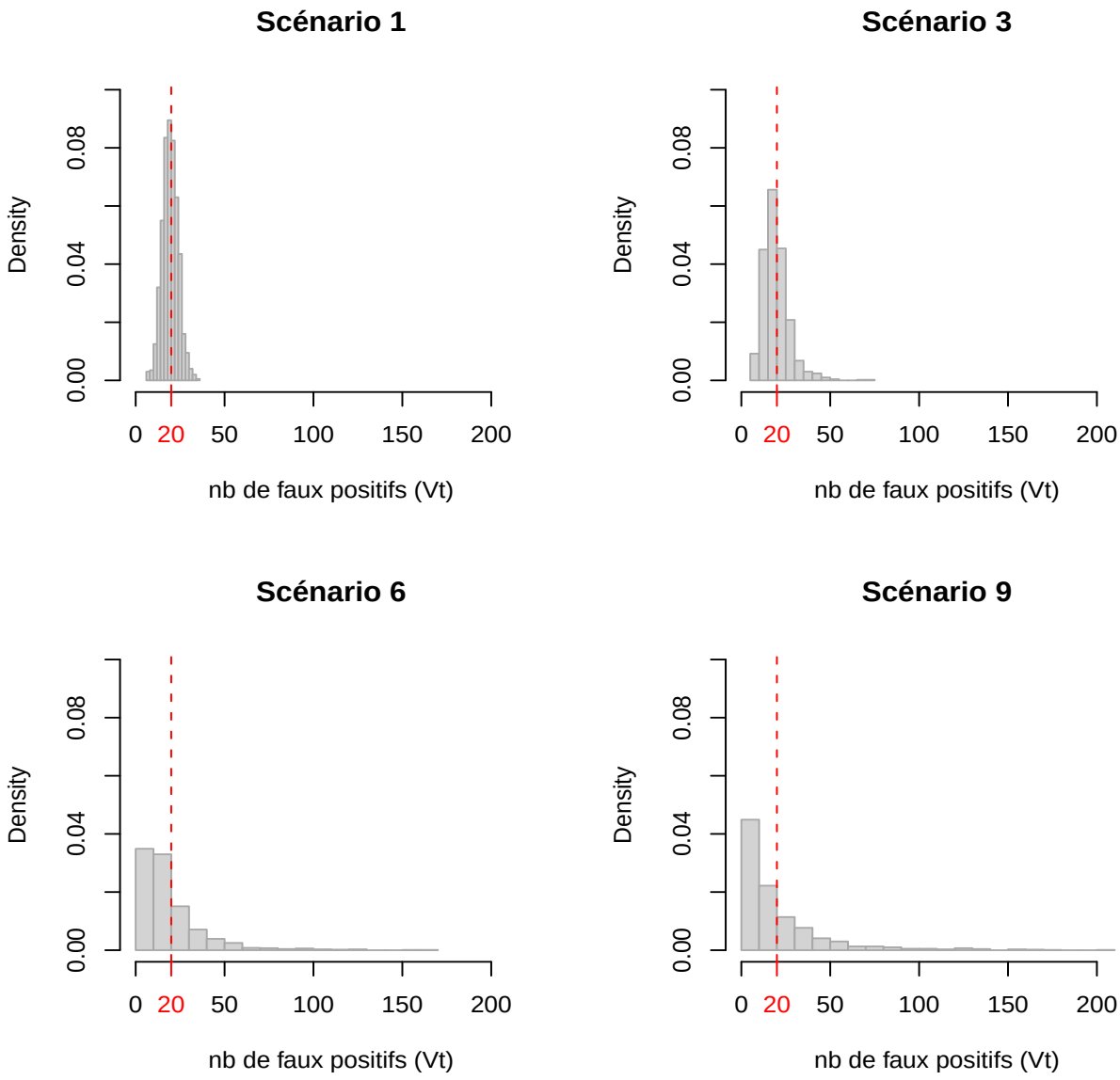


FIG. 2.11 – Distribution du nombre de faux-positifs pour les données simulées (scénarios 1, 3, 6 et 9)

D'après la PROPOSITION 2.3.1, la présence de dépendance n'affecte pas le nombre moyen de faux-positifs. Par contre, la variance de V_t introduit un terme dépendant directement de la corrélation, à travers la fonction $D^{kk'}(t)$: l'impact de la dépendance sur la variance du nombre de faux-positifs est mesuré par la somme des quantités $D^{kk'}(t)$, calculées pour toutes les paires de variables $\{k, k'\}$ pour lesquelles l'hypothèse nulle est vraie.

Nous définissons la fonction $D^{kk'}(t)$ dans le cas particulier où $\{k, k'\} \in \mathcal{M}_0$ et dans le cadre du test de Student.

Définition 2.3.1 (Fonction $D^{kk'}(t)$ - $k, k' \in \mathcal{M}_0$). *On note $\Phi^{(2)}(J;\rho)$ la probabilité qu'un vecteur, suivant une loi de Student bivariée, de paramètres de non-centralité nuls et de matrice de variance covariance $\begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$, n'appartienne pas à $J \subseteq \mathbb{R}^2$.*

$$D_0^{kk'}(t) = \frac{\Phi^{(2)}\Big([-u_t;u_t]^2;\rho_{kk'}\Big) - t^2}{t(1-t)}$$

En particulier, $0 \leq D_0^{kk'}(t) \leq 1$. La borne inférieure 0 est atteinte en $\rho = 0$ et la borne supérieure 1 est atteinte en $\rho = \pm 1$. La FIGURE 2.12 montre la courbe de la fonction $D_0^{kk'}(t)$ pour différentes valeurs de t.

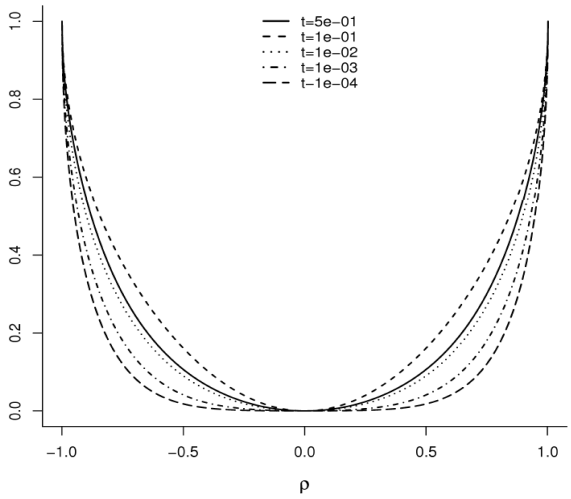


FIG. 2.12 – Courbe $D_0^{kk'}(t)$ pour différentes valeurs de t, $k, k' \in \mathcal{M}_0$

La FIGURE 2.13 représente la variance théorique de V_t selon différentes valeurs de t, pour chacun des 10 scénarios. La corrélation entre les variables est déterminée à partir des matrices de variances-covariances qui ont permis de générer les données simulées de l'EXEMPLE 4. Cela confirme que la présence de dépendance conduit à une instabilité de la distribution du nombre de faux-positifs. Même si il semble que pour de faibles valeurs de t, l'écart de chacun des scénarios de dépendance avec le cas indépendant diminue, il faut rappeler que l'espérance du nombre de faux-positifs diminue également avec le seuil choisi. Le TABLEAU 2.4 donne la variance théorique de V_t pour $t = 0,05$. On retrouve bien des valeurs similaires à celles observées pour les écarts-types du TABLEAU 2.3.

Certains auteurs [Owen, 2005, Efron, 2007] ont également obtenu des résultats similaires pour la modélisation de la variance du nombre de faux-positifs. En particulier, Owen [2005] propose une expression à base d'une intégrale simple de la probabilité bivariée $\Phi^{(2)}$, lorsque les covariables x sont supposées être des variables indépendantes gaussiennes. Il propose également une approximation ne nécessitant pas le calcul de la somme pour toutes les paires de corrélations (de l'ordre de m_0^2). Cette approximation est utile pour le calcul de la variance de V_t en pratique. Nous proposons une approche

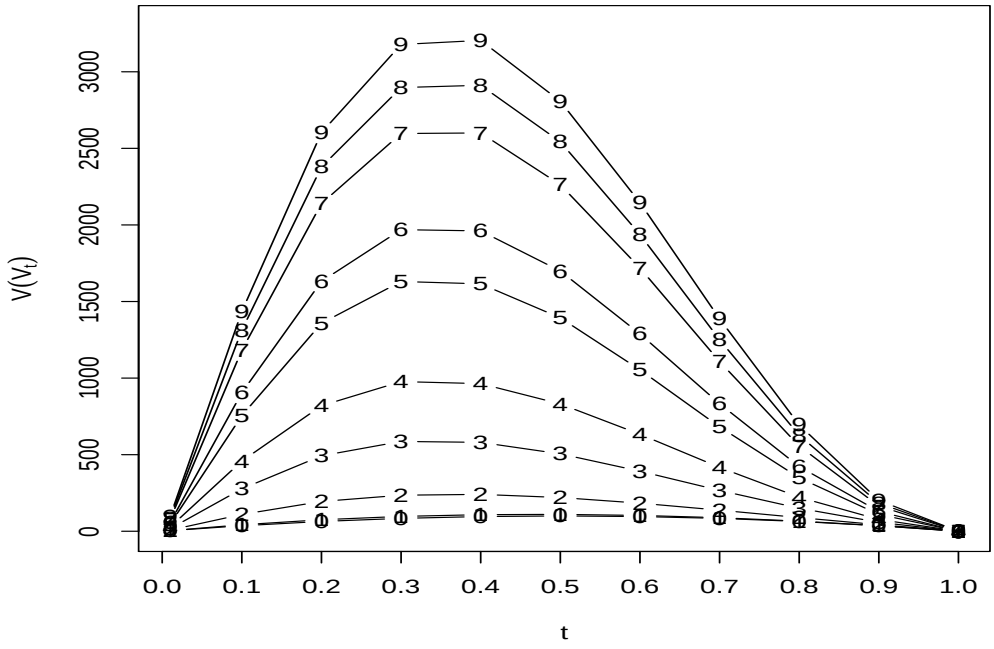


FIG. 2.13 – Variance de V_t selon différentes valeurs de t, pour chacun des 10 scénarios

scénario	0	1	2	3	4	5	6	7	8	9
var. commune (%)	0	4,65	15,56	28,56	40,31	50,76	57,90	65,45	70,09	75,19
variance V_t	19	21,89	52,71	129,52	213,77	352,88	424,28	556,68	619,70	678,83
écart-type V_t	4,36	4,68	7,26	11,38	14,62	18,79	20,60	23,59	24,89	26,05

TAB. 2.4 – Variances et écarts-types théoriques de $V_{t=0,05}$ calculés à partir des matrices de variances-covariances qui ont permis de générer les données simulées de l'EXEMPLE 4

similaire, présentée plus en détail dans la SECTION 4.3.

2.3.2 Impact de la dépendance sur le FWER

Nous considérons ici la procédure de Sidak [Sidak, 1967], qui consiste à définir le seuil suivant pour le rejet des probabilités critiques : $t_{sidak} = 1 - (1 - \alpha)^{1/m}$, où α est le niveau de contrôle global souhaité, ici $\alpha = 0,05$. Les résultats obtenus avec la procédures de Bonferroni [Bonferroni, 1936] sont similaires.

Les résultats de l'application de cette procédure sur les données simulées de l'EXEMPLE 4 sont fournis au le TABLEAU 2.5 et sur la FIGURE 2.14. Quel que soit le niveau de dépendance, le FWER est inférieur au seuil α. Pour les scénarios avec un faible niveau de dépendance, le niveau de contrôle du FWER est en réalité de $\pi_0\alpha$. En présence de dépendance, le risque de commettre une erreur est encore plus faible. Le contrôle du FWER plus strict qu'attendu en présence de dépendance conduit à conclure à une perte de puissance des procédures de tests multiples dans ce cadre. C'est une affirmation qu'on

retrouve fréquemment dans des articles traitant des propriétés des procédures de tests multiples en présence de dépendance [Efron, 2007, Storey, 2007]. Néanmoins, il arrive que pour certains tableaux de données simulés le nombre d'erreurs commises soit en réalité très important.

Pour information, on donne entre parenthèse dans le TABLEAU 2.5 les résultats obtenu par la procédure de Sidak en considérant m_0 connu, soit avec le seuil $t_{sidak} = 1-(1-\alpha)^{1/m_0}$. Les commentaires sont similaires, sauf que le niveau de contrôle du FWER est plus proche du seuil fixé α pour des scénarios de faible dépendance. En présence de dépendance, on observe également un nombre d'erreurs qui peut être très important.

scénario	0	1	2	3	4	5	6	7	8	9
var. comm.(%)	0	4,65	15,56	28,56	40,31	50,76	57,90	65,45	70,09	75,19
fréquences										
0	957(949)	958(952)	970(958)	973(969)	966(955)	971(965)	969(957)	964(960)	970(963)	968(968)
1	42(49)	41(46)	29(40)	24(28)	33(42)	26(31)	27(35)	24(26)	22(26)	25(28)
2	(1)	1(2)	1(2)	3(1)	(2)	2(2)	3(7)	4(5)	3(5)	3(4)
3	1(1)			(1)	1	(1)		5(3)	3(3)	2(3)
4				(1)		1(1)	1	1(3)	(1)	(1)
5							(1)	2(3)		
6									1	
7					(1)				(1)	
8										2
9										(1)
10										(1)
11										
12										
13										
14										
15									1	
16										
17									(1)	
FWER										
$\#\{V_t > 0\}/m$	4,3(5,1)	4,2(4,8)	3,0(4,2)	2,7(3,1)	3,4(4,5)	2,9(3,5)	3,1(4,3)	3,6(4,0)	2,9(3,7)	3,2(3,2)

TAB. 2.5 – FWER estimé pour les 10 scénarios de données simulées (résultats en %) et tableau des fréquences observées pour V_t - Procédure de Sidak [Sidak, 1967], avec un niveau α fixé à $0,05$ pour le risque de type-I (seuil $t = 1,0258.10^{-4}$). Entre parenthèse : Même procédure en considérant m_0 connu (seuil $t = 1,2822.10^{-4}$)

2.3.3 Impact de la dépendance sur le FDR

Pour estimer le FDR, nous disposons de deux estimateurs : l'estimateur empirique $\widehat{FDR}_t = \frac{m_0 t}{R_t}$ et l'estimateur donné par la procédure de Storey [Storey et al., 2004] où le FDR est estimé en considérant le seuil t correspondant à la valeur de la probabilité critique $p_{(k^*)}$ telle que $k^* = \text{argmax}_k \left(p_{(k)} < \frac{\alpha}{m_0} k \right)$. L'estimateur est donc finalement le même que l'estimateur empirique avec $t = p_{(k*)}$: pour le premier estimateur, t est fixé quel que soit le jeu de données, et pour le second estimateur, le seuil dépend des données considérées. Dans les deux cas, si m_0 est inconnu, il est alors estimé par $\hat{m}_0(\lambda)$ (voir SECTION 1.2).

Storey et al. [2004] démontrent la propriété suivante :

PROPOSITION 2.3.2 (Propriété de l'estimateur du FDR [Storey et al., 2004])**.** *Sous les conditions de l'*HYPOTHÈSE *1, pour* $\lambda \in [0;1[$ *fixé :*

$$\mathbb{E}(\widehat{FDR}_t(\lambda)) \geq FDR_t$$

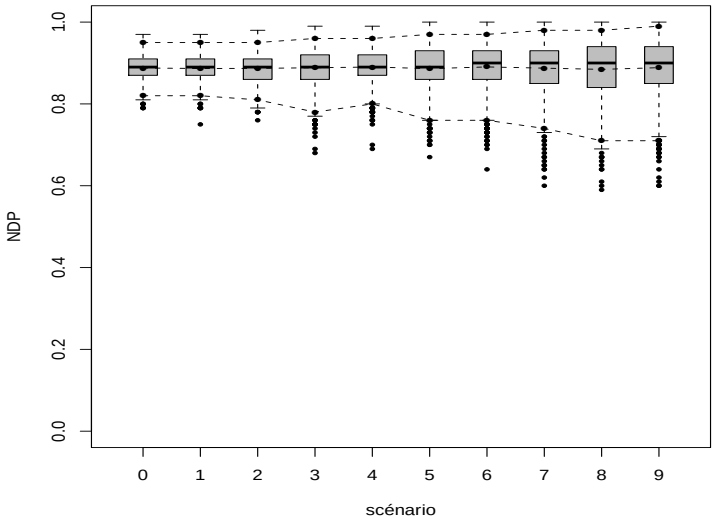


FIG. 2.14 – Nombre de non-découvertes (Non Discovery Proportion) en fonction du niveau de dépendance pour les 10 scénarios - Procédure de Sidak [Sidak, 1967] sur les probabilités critiques usuelles

L'estimation empirique du FDR $\widehat{FDR_t} = \frac{m_0 t}{R_t}$ est basée sur l'approximation $\mathbb{E}(\frac{V_t}{R_t}) \approx \frac{E(V_t)}{\mathbb{E}(R_t)}$. Cette approximation est maintenant étudiée en présence de dépendance en considérant tout d'abord le développement limité de Taylor d'ordre 2 d'une fonction f de deux variables $z = (x, y)$ en $a = (x_0; y_0)$:

$$f(z) = f(a) + (z-a)\nabla f(a) + \frac{1}{2}(z-a)'\mathbb{H}(z-a) + \circ(\|z-a\|^2)$$

Posons $f(x,y) = \frac{x}{x+y}$. En prenant $x = V_t$, $y = S_t$, $x_0 = \mathbb{E}(V_t)$ et $y_0 = \mathbb{E}(S_t)$, on obtient le développement limité d'ordre 2 de la proportion de faux-positifs V_t/R_t :

$$\begin{aligned} FDP_t = \frac{V_t}{R_t} &= \frac{\mathbb{E}(V_t)}{\mathbb{E}(R_t)} + \frac{V_t\mathbb{E}(S_t) - S_t\mathbb{E}(V_t)}{\mathbb{E}(R_t)^2} \\ &+ \frac{\mathbb{E}(V_t)(S_t - \mathbb{E}(S_t))^2 - \mathbb{E}(S_t)(V_t - \mathbb{E}(V_t))^2 + (\mathbb{E}(V_t) - \mathbb{E}(S_t))(V_t - \mathbb{E}(V_t))(S_t - \mathbb{E}(S_t))}{\mathbb{E}(R_t)^3} \\ &+ \circ(\|(V_t - \mathbb{E}(V_t); S_t - \mathbb{E}(S_t))\|^2) \end{aligned}$$

On calcule ensuite l'espérance de l'expression précédente, ce qui permet de déduire :

$$\begin{aligned} FDR_t = \mathbb{E}\left(\frac{V_t}{R_t}\right) &= \frac{\mathbb{E}(V_t)}{\mathbb{E}(R_t)} + \frac{\mathbb{E}(V_t)\mathbb{V}(S_t) - \mathbb{E}(S_t)\mathbb{V}(V_t) + (\mathbb{E}(V_t) - \mathbb{E}(S_t))\,\mathrm{Cov}(V_t, S_t)}{\mathbb{E}(R_t)^3} \\ &+ \circ(\|(V_t - \mathbb{E}(V_t); S_t - \mathbb{E}(S_t))\|^2) \end{aligned} \quad (2.7)$$

La FIGURE 2.15 montre l'évolution du deuxième terme de (2.7) en fonction de t pour chacun des 10 scénarios de l'EXEMPLE 4. La corrélation entre les variables est déterminée à partir des matrices de variances-covariances qui ont permis de générer les données simulées. Le deuxième terme de (2.7) est négatif (voir FIGURE 2.15) et tend vers 0, quel que soit t, en cas d'indépendance. On illustre ici

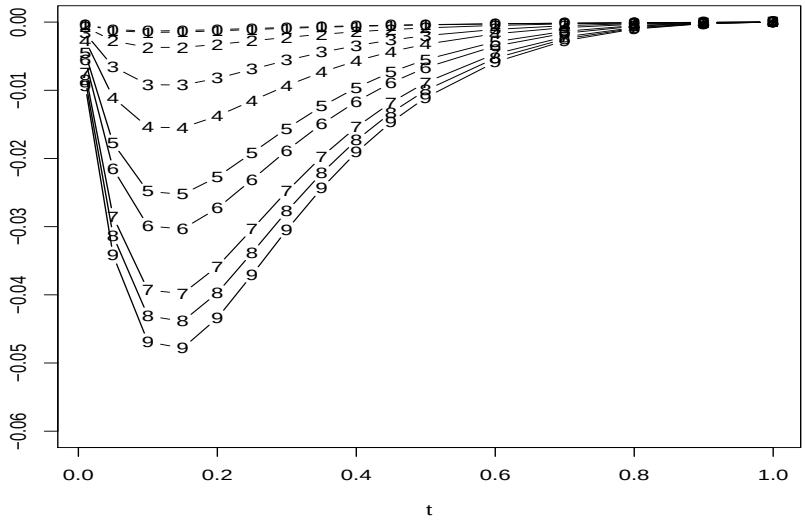


FIG. 2.15 – Evolution du deuxième terme de (2.7) en fonction du seuil t, pour les 10 scénarios

la PROPOSITION 2.3.2. Néanmoins, même pour des scénarios de forte dépendance, il semble que la valeur de ce terme reste négligeable par rapport à la valeur de $F\hat{D}R_t$. De plus, si l'estimation du FDR est biaisé, il s'agit d'un biais positif ce qui se traduit par un contrôle plus stricte du FDR réel par les procédures. Les conséquences pratiques de ce biais dans les décisions sont donc uniquement une diminution de la puissance par rapport à ce qui est attendu [Sarkar, 2008].

Pour chaque tableau de données simulé (EXEMPLE 4), et pour les deux estimateurs, on calcule le FDR estimé et on détermine la vraie proportion d'hypothèses rejetées à tort : $FDP_t = V_t/R_t$, ainsi que la proportion de non découvertes (faux-négatifs) : $NDP_t = T_t/m_1$.

On choisit un seuil $t = 0,05$ pour les probabilités critiques. L'espérance du nombre de faux-positifs est de 20, et l'espérance du nombre de rejets est de 100 car on a fixé la puissance individuelle des tests à 80%. Afin d'éviter toute confusion, et n'étudier que l'impact de la dépendance sur l'estimation des taux d'erreurs, on suppose que π_0 est connu, soit $\pi_0 = 0,80$.

Pour l'estimation du FDR par la *q-value*, on considère le seuil $t_\alpha(\widehat{FDR})$ avec $\alpha = 0,20$.

Le TABLEAU 2.6 donne les statistiques descriptives des estimations du FDR par les deux méthodes, pour les différents scénarios. Quel que soit le niveau de dépendance, la moyenne du FDR estimé est stable, pour les deux estimateurs. La variabilité est nettement plus faible dans le cas de l'estimation par la *q-value*, pour tous les niveaux de dépendance : proposer un seuil dépendant des données semble réduire la variabilité de l'estimation du FDR.

Le TABLEAU 2.7 donne les statistiques descriptives de la vraie proportion de faux-positifs (FDP) pour les différents scénarios. Quel que soit le niveau de dépendance, la moyenne de la proportion de faux-positifs (vrai FDR) est inférieure à la moyenne des FDR estimés : on retrouve ici la PROPOSITION 2.3.2. Par contre, on observe une forte augmentation de la variabilité autour de cette valeur cible à mesure que le niveau de dépendance augmente (voir figure 2.16(a)).

scénario	0	1	2	3	4	5	6	7	8	9
var. commune (%)	0	4,65	15,56	28,56	40,31	50,76	57,90	65,45	70,09	75,19
$\widehat{FDR}_{t=0.05}$										
min	17,09	15,62	13,33	12,20	8,85	8,55	9,22	7,30	7,38	7,75
$q_{0,25}$	19,23	19,23	19,05	19,23	19,23	19,42	19,61	19,23	19,23	19,42
médiane	20,00	20,00	20,00	20,41	20,62	20,83	21,05	21,05	21,05	21,28
moyenne	**20,11**	**20,02**	**20,08**	**20,20**	**20,33**	**20,34**	**20,52**	**20,46**	**20,51**	**20,65**
$q_{0,75}$	20,83	20,83	21,05	21,28	21,74	21,98	21,98	22,47	22,47	22,47
max	25,32	24,10	24,69	25,00	25,64	25,00	27,40	26,67	28,17	30,30
écart-type	**1,23**	**1,25**	**1,47**	**1,78**	**2,11**	**2,39**	**2,57**	**3,08**	**3,12**	**3,19**
q-value										
min	14,69	14,88	15,99	15,94	15,83	15,18	13,37	15,07	13,63	12,99
$q_{0,25}$	19,16	19,17	19,16	19,10	19,12	19,15	19,09	19,10	19,11	19,02
médiane	19,59	19,59	19,59	19,55	19,62	19,59	19,59	19,56	19,60	19,56
moyenne	**19,41**	**19,40**	**19,42**	**19,37**	**19,38**	**19,38**	**19,33**	**19,34**	**19,34**	**19,27**
$q_{0,75}$	19,84	19,83	19,85	19,82	19,85	19,83	19,82	19,84	19,84	19,84
max	20,29	20,33	20,54	20,44	20,47	20,36	20,28	20,58	20,51	20,41
écart-type	**0,63**	**0,63**	**0,60**	**0,63**	**0,67**	**0,68**	**0,76**	**0,73**	**0,78**	**0,87**

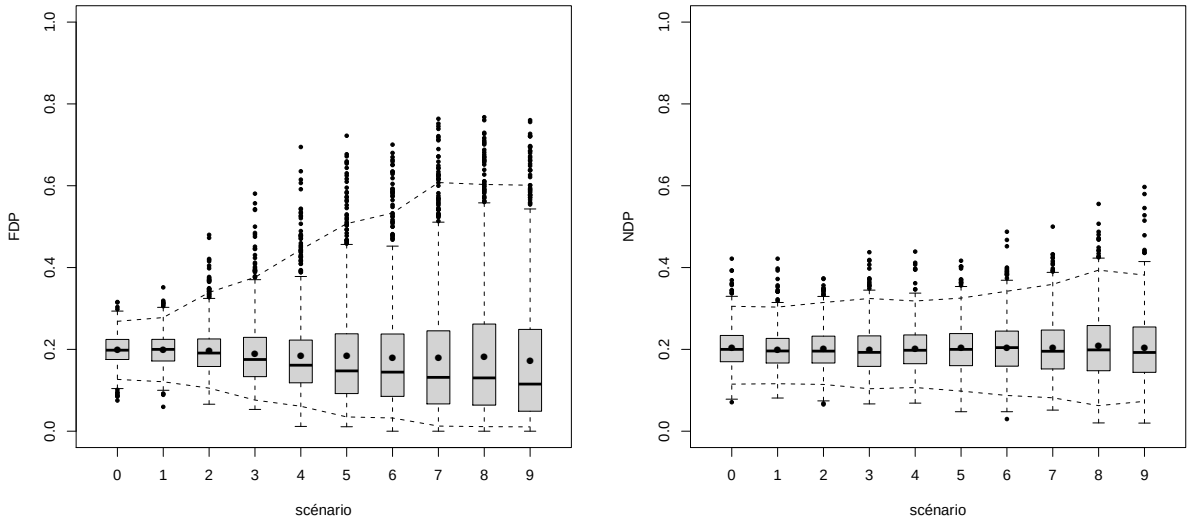
TAB. 2.6 – Statistiques descriptives de l'estimation du FDR pour les 10 scénarios de données simulées (résultats en %)

scénario	0	1	2	3	4	5	6	7	8	9
var. commune (%)	0	4,65	15,56	28,56	40,31	50,76	57,90	65,45	70,09	75,19
FDP ($t = 0,05$)										
min	7,50	5,95	6,59	5,32	1,16	1,09	0,00	0,00	0,00	0,00
$q_{0,25}$	17,53	17,17	15,83	13,33	11,83	9,19	8,51	6,67	6,38	4,88
médiane	19,80	20,00	19,01	17,53	16,13	14,74	14,44	13,16	13,02	11,52
moyenne	**19,86**	**19,90**	**19,67**	**18,87**	**18,48**	**18,31**	**17,88**	**17,97**	**18,23**	**17,74**
$q_{0,75}$	22,43	22,45	22,55	22,93	22,25	23,79	23,75	24,53	26,18	24,82
max	31,58	35,16	48,00	58,07	69,47	72,22	70,05	76,37	76,75	76,03
écart-type	**3,64**	**4,04**	**5,70**	**7,82**	**9,69**	**12,36**	**12,79**	**15,13**	**15,80**	**16,08**
FDP (seuil *q-value*)										
min	4,17	4,94	4,76	3,49	1,16	0,00	0,00	0,00	0,00	0,00
$q_{0,25}$	17,17	16,83	15,05	12,37	10,78	8,05	7,30	5,60	5,32	3,90
médiane	19,82	20,21	19,00	17,20	15,53	14,00	13,40	12,07	11,43	10,01
moyenne	**19,96**	**20,02**	**19,82**	**19,03**	**18,78**	**18,60**	**18,07**	**18,34**	**18,68**	**17,62**
$q_{0,75}$	22,86	23,17	23,15	23,48	23,00	24,16	24,30	25,05	26,6	25,60
max	32,80	36,50	54,26	63,59	74,83	76,27	76,38	78,07	79,18	79,75
écart-type	**4,38**	**4,84**	**6,91**	**9,33**	**11,65**	**14,55**	**14,98**	**17,64**	**18,37**	**18,59**

TAB. 2.7 – Statistiques descriptives de la vraie proportion d'erreurs pour les 10 scénarios de données simulées (résultats en %)

La proportion de faux-négatifs (Non Discovery Proportion) reste par contre semblable pour tous les scénarios de dépendance (FIGURE 2.16(b)).

On se propose alors de comparer l'estimation du FDR et sa valeur attendue, pour chacun des jeux de données, en fonction du niveau de dépendance.



(a) FDP (b) NDP

FIG. 2.16 – Proportion de faux-positifs (FDP) et de faux-négatifs (NDP), en fonction du niveau de dépendance pour les 10 scénarios de données simulées

La FIGURE 2.17 représente sous forme graphique la valeur de l'estimation du FDR en fonction de la vraie proportion de faux-positifs (FDP) pour quatre des 10 scénarios, ce qui illustre les commentaires faits précédemment à partir des TABLEAUX 2.6 et 2.7. L'estimation empirique du FDR conduit même à une interprétation inversée de la proportion de faux-positifs, la pente de régression entre l'estimation et la vraie proportion étant négative (voir TABLEAU 2.8). L'estimation par la *q-value* corrige en partie ce problème. Cela s'explique par la construction même du seuil de rejet par la méthode, différent pour chaque tableau de données.

Ainsi, les deux estimateurs du FDR rendent bien compte de la valeur moyenne de la proportion de faux-positifs, même en présence de dépendance. Néanmoins, la détermination de t à partir des données permet d'obtenir une estimation du FDR plus précise. Cependant, en présence de corrélation, la vraie proportion de faux-positifs varie énormément. Ces deux indicateurs ne sont donc pas adaptés pour rendre compte de la vraie proportion d'erreurs de type-I commises : dans une situation de grande dépendance, le contrôle du FDR peut donc s'avérer inefficace.

scénario	0	1	2	3	4	5	6	7	8	9
var. commune (%)	0	4,65	15,56	28,56	40,31	50,76	57,90	65,45	70,09	75,19
$\widehat{FDR}_{t=0.05}$	-0,197	-0.196	-0,189	-0,176	-0.187	-0,171	-0,175	-0,182	-0,169	-0,171
q-value	0,025	0,026	0,017	0,016	0,015	0,013	0,017	0,013	0,014	0,016

TAB. 2.8 – Coefficients de pente dans la régression entre le FDR estimé et la vraie proportion de faux-positifs (FDP)

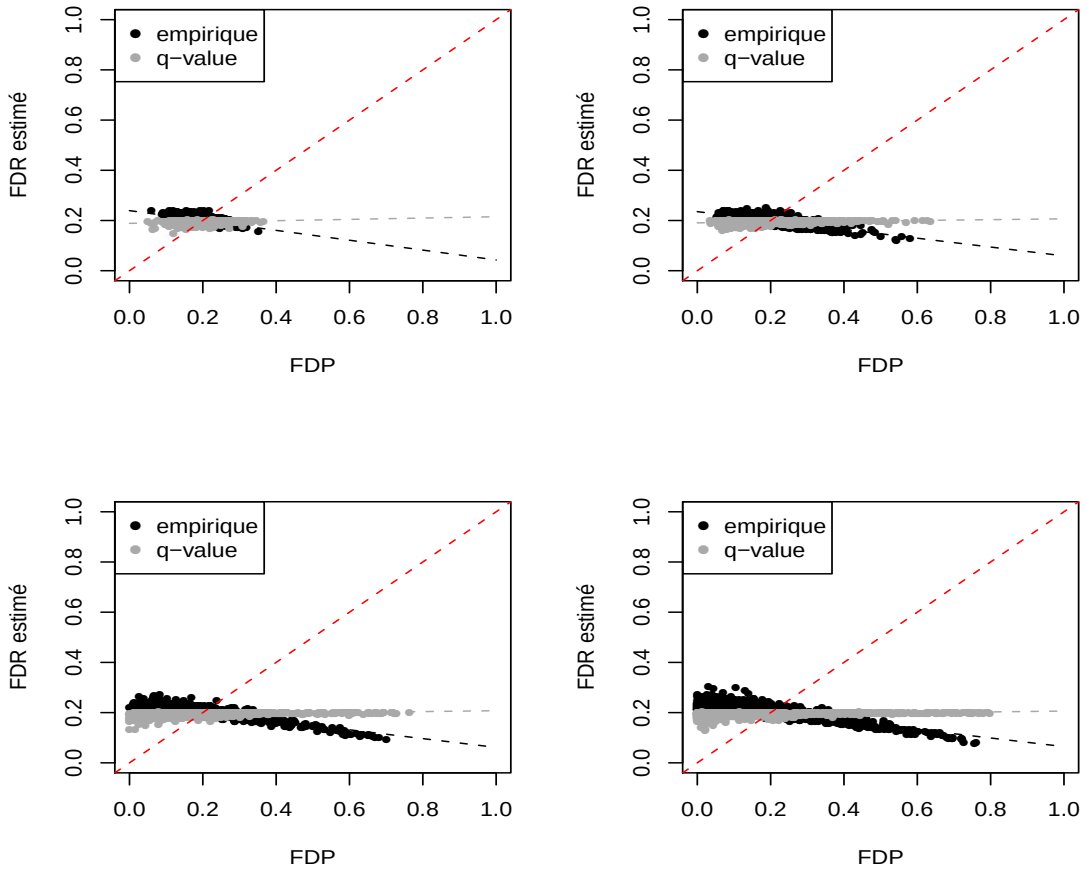


FIG. 2.17 – Comparaison entre le FDR estimé et la vraie proportion de faux-positifs (FDP) - scénarios 1, 3, 6 et 9

Conclusion

La dépendance induit une variabilité qui perturbe notamment la distribution empirique des probabilités critiques sous l'hypothèse nulle, celle-ci pouvant s'éloigner fortement de la distribution théorique. La conséquence directe est une augmentation de la variabilité du nombre de faux-positifs (V_t) en présence de dépendance, entraînant alors une instabilité des procédures de tests multiples. Les taux d'erreurs contrôlés par ces procédures sont en effet définis à partir du nombre de faux-positifs. L'estimation de π_0, la proportion d'hypothèses nulles, est quant à elle biaisée en présence de dépendance. Ce biais est positif ou négatif selon que les probabilités critiques proches de 0 sont sous- ou sur-représentées.

Nous proposons dans ce chapitre de modéliser la structure de dépendance par un modèle à variables latentes (2.1), qui capturent l'information commune à l'ensemble des variables. Ce cadre nous permet de donner l'expression exacte de la variance du nombre de faux-positifs et de π_0, dans un cadre général de dépendance. Nous mettons en évidence dans ces deux expressions un terme directement fonction

de la corrélation entre les variables.

Le chapitre suivant décrit une méthode qui permet de tirer profit des propriétés obtenues par une approche conditionnelle aux variables latentes.

CHAPITRE 3

APPROCHE CONDITIONNELLE DES TESTS MULTIPLES EN GRANDE DIMENSION EN PRÉSENCE DE DÉPENDANCE

Résumé La modélisation de la dépendance à travers un ensemble de variables latentes permet de définir un cadre d'analyse de l'effet de la corrélation. Dans ce cadre, nous définissons ici des statistiques de test conditionnellement indépendantes. L'avantage apporté par la prise en compte de la dépendance dans le modèle est le contrôle de la variance des taux d'erreurs et de celle de π_0. Les propriétés des procédures de tests sont grandement améliorées, notamment au regard du taux de non découvertes (puissance des tests) par rapport aux méthodes classiques. Nous définissons également un estimateur conditionnel pour le taux de faux-positifs (FDR) et un estimateur conditionnel pour la proportion d'hypothèses nulles (π_0).

Sommaire

Introduction

Il y a principalement deux enjeux dans les procédures de tests multiples : le contrôle des taux d'erreurs et indirectement l'estimation de la proportion d'hypothèses nulles. En général, les méthodes développées se basent sur les probabilités critiques issues des tests simultanés pour chacune des variables. Pour la plupart des méthodes (voir CHAPITRE 1) les probabilités critiques sont supposées être indépendamment distribuées.

La prise en compte de la dépendance est un des sujets cruciaux de l'étude des procédures de test multiples en grande dimension. On a vu au CHAPITRE 2 que la présence de corrélation entre variables affectait toutes les étapes des procédures. En particulier, si certaines méthodes contrôlent l'espérance du taux de faux-positifs même en présence de corrélation, les taux d'erreurs estimés présentent une grande variabilité.

On rappelle le modèle (2.1) défini au CHAPITRE 2 qui suppose l'existance d'un ensemble de vecteurs Z modélisant l'information commune à l'ensemble des m variables étudiées. Les variables Z peuvent être vues comme des facteurs de variations communes à l'ensemble des données. Ainsi, si b_k est le $Q-$vecteur des coefficients de chaque variable Y_k, $k \in [1; m]$, dans les combinaisons linéaires les reliant aux Q variables latentes Z, on a (2.1) :

$$Y_k \quad = \quad m_k(x) + Zb_k' + \varepsilon_k$$

où $E = [\varepsilon_1, \ldots, \varepsilon_m]$ est composé de vecteurs indépendants.

Dans un premier temps, on définit une procédure d'ajustement des variables d'intérêt par rapport aux variables latentes Z. En particulier, on présente l'effet de cet ajustement sur l'estimation de π_0 et des taux d'erreurs. Dans un second temps, des estimateurs conditionnels sont proposés dans l'optique de réduire l'impact de la dépendance sur l'estimation mis en évidence au CHAPITRE 2.

3.1. Données ajustées

3.1.1 Construction de statistiques de test indépendantes

Avant de présenter une méthodologie permettant de prendre en compte la dépendance à travers la structure en facteurs, nous considérons un petit exemple dont la problématique est similaire mais plus simple, à savoir le test d'une seule hypothèse en présence de covariables.

3.1.1.1 Construction d'un test ajusté en présence de covariables

On se place ici dans le cadre du modèle linéaire décrit dans la SECTION 1.1.2 : considérons m variables d'intérêt mesurées sur un échantillon de taille n. On s'intéresse au lien entre chaque mesure Y_k et une covariable d'intérêt x à travers le test de nullité du contraste $c'\theta_k$. L'estimateur classique de θ_k (estimateur des moindres carrés) noté $\hat{\theta}_k$ permet de définir la statistique du test et sa loi sous l'hypothèse nulle (DÉFINITION 1.1.1).

Par ailleurs, soit $Z = [Z_1, ..., Z_Q]$, Q variables aléatoires indépendantes de X, avec $Q < n$, d'espérance nulle et de matrice de variance identité. En pratique, par exemple dans le cadre de l'analyse de données génomiques, ces variables peuvent être les mesures de l'expression de gènes constitutifs, c'est-à-dire de gènes s'exprimant de la même manière dans toutes les cellules d'un organisme, sans mécanisme de régulation (*housekeeping genes*) et qui servent de mesures de contrôle pour identifier les erreurs de mesures techniques notamment. En épidémiologie, on dispose également d'un certain nombre de mesures sur les patients de l'étude, qui peuvent jouer ce rôle de covariables indépendantes de x dans l'analyse. Cette information à disposition n'est pas prise en compte en général par l'estimateur $\hat{\theta}_k$. Pourtant, intégrer de l'information externe améliore la précision de l'estimation. Cette méthodologie a été très étudiée entre autre dans le cadre des essais cliniques par une approche semi-paramétrique [Tsiatis et al., 2000, Leon et al., 2003, Davidian et al., 2005]. C'est également le principe du double-échantillonnage [Kloareg and Causeur, 2007], lorsque l'information externe des covariables Z est mesurée sur un ensemble d'individus plus large que l'échantillon pour lequel la variable d'intérêt à été mesurée. La puissance du test pour la variable Y_k est améliorée en utilisant l'information apportée par Z, plus particulièrement lorsque la corrélation entre la variable d'intérêt et Z est importante.

Pour améliorer l'estimateur usuel de θ_k en exploitant l'information contenue dans Z, et à partir des relations entre les variables observées Y_k et Z, on considère le modèle suivant :

$$\begin{cases} \mathbb{E}(Y_k|x, Z) = X\theta_k + Zb'_k \\ \mathbb{E}(Z_q|x) = \mu_q \ , \ q \in [1, Q] \end{cases}$$

Ou bien de façon matricielle :

$$\begin{cases} \mathbb{E}(Y|x, Z) = X\theta + ZB' \\ \mathbb{E}(Z|x) = \mu \end{cases}$$

Nous considérons le cas ici où $p = 1$ (une seule variable explicative). X est de dimension $n \times 2$, et $\theta = [\theta_1, \ldots, \theta_m]$ de dimension $2 \times m$. Z est de dimension $n \times Q$ et b_k représente une ligne de B, matrice de dimension $m \times Q$. $\mu = [\mu_1, \ldots, \mu_Q]$ est un $Q-$vecteur.

L'estimateur régulier, asymptotiquement linéaire le plus efficace, qui est de la forme [Robins et al., 1994] :

$$\tilde{\theta}_k = \hat{\theta}_k - b'_k\hat{\theta}_Z \tag{3.1}$$

où $\hat{\theta}_Z$, de dimension $(p+1) \times Q$, représente l'estimateur des moindres carrés des coefficients dans le modèle de régression reliant X et Z. (3.1) montre que les estimateurs de θ_k peuvent être vus comme

des estimateurs construits à partir de l'estimateur non biaisé $\hat{\theta}_k$, qui ne prend pas en compte les covariables, qu'on ajuste ensuite grâce à l'information apportée par les covariables de façon explicite à travers le second terme $b'_k\hat{\theta}_Z$.

Remarque : On choisit d'introduire θ_Z pour donner du sens à $\hat{\theta_Z}$. Effectivement, ici $\theta_Z = 0$ et l'espérance de $\hat{\theta_Z}$ est nulle également. $\tilde{\theta}_k$ est finalement une version ajustée de $\hat{\theta_k}$: on retire de l'estimateur usuel une quantité d'espérance nulle (on n'introduit pas de biais) et qui permet de réduire la variance de l'estimateur obtenu par rapport à l'estimateur usuel, et ce d'autant plus que la covariance entre Y_k et Z est importante. C'est cette propriété qu'on va vouloir conserver par la suite.

L'estimateur $\tilde{\theta}_k$ est non biaisé : $\mathbb{E}(\tilde{\theta}_k) = \mathbb{E}(\hat{\theta}_k - b'_k\hat{\theta}_Z) = \theta_k$ car $\mathbb{E}(\hat{\theta}_Z) = 0$. De plus, sa variance est plus faible que celle de $\hat{\theta}_k$: $\mathbb{V}(\tilde{\theta}_k) = \mathbb{V}(\hat{\theta}_k) + \mathbb{V}(b'_k\hat{\theta}_Z) - 2\,\mathrm{Cov}(\hat{\theta}_k; b'_k\hat{\theta}_Z) = (\sigma^2_y - b_k b'_k)(X'X)^{-1} = \sigma^2_{Y_k|Z}(X'X)^{-1}$, où $\sigma^2_{Y_k|Z}$ est la variance résiduelle dans le modèle reliant Y_k à X et Z.

L'estimation de l'effet de x sur la variable d'intérêt est donc améliorée, en terme de variance, par rapport au test classique en considérant la statistique de test :

$$\tilde{T}_k \quad = \quad \frac{c'\tilde{\theta}_k}{\mathbb{V}(c'\tilde{\theta}_k)} = \frac{c'\hat{\theta}_k - c'b'_k\hat{\theta}_Z}{\sqrt{\sigma^2_{Y_k|Z}c'S_x^{-1}c}} \tag{3.2}$$

Remarque : Si on note β le $Q+p+1-$vecteur : $\beta = [\theta_Z; \theta]$, on retrouve le test classique dans le cadre du modèle linéaire généralisé d'une combinaison particulière des composantes de β, qui suppose la nullité des Q premiers éléments. La statistique de test (3.2) correspond alors à la statistique du test du rapport des vraisemblances.

Le petit exemple suivant permet d'illustrer l'apport de la statistique de test (3.2) en terme d'erreurs de types-I et II.

Exemple 5. On simule 10 000 tableaux de données, composés de 2 variables notées Y et Z, et mesurées sur des échantillons de taille n répartis en 2 groupes de taille $n_A = n_B = 12$. La variable X code pour l'appartenance à un de ces groupes. On suppose qu'il n'y a pas d'effet du groupe pour la variable Z, qui est simulée selon une loi $\mathcal{N}(0, 1)$.

La variable Y est simulée selon une loi $\mathcal{N}(0, 1)$. Une quantité δ donnée est ajoutée aux individus du groupe B. On contrôle la corrélation ρ_{YZ} entre les variables Y et Z. On pourra ainsi étudier l'impact de la corrélation sur les erreurs de type-I et II en faisant varier ce paramètre entre les deux variables entre 0 et 0.9.

Pour chaque tableau de données, on teste l'effet de X sur la variable Y, en tenant compte de l'information apportée par Z.

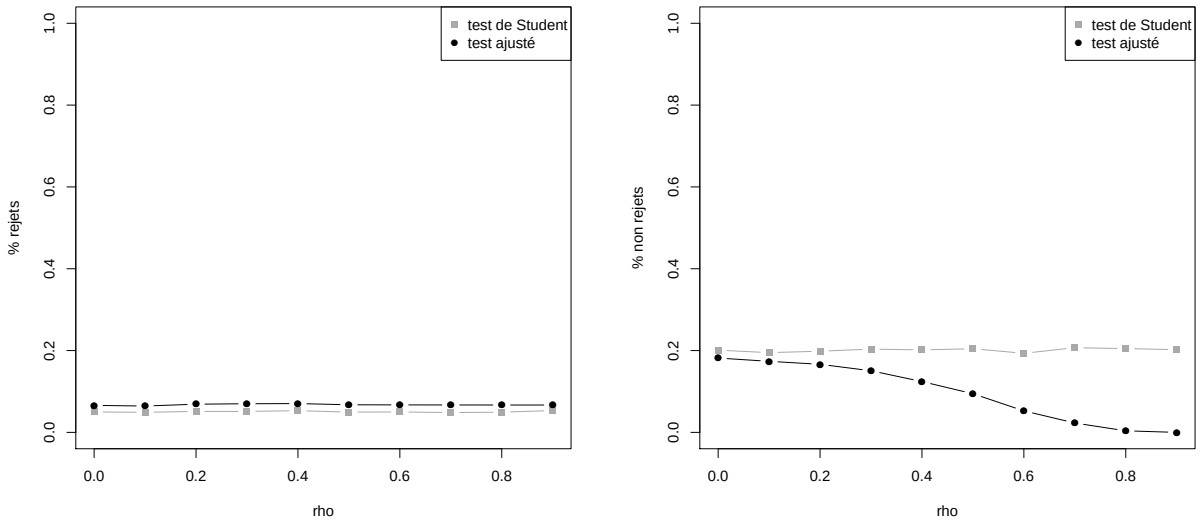
On va considérer deux scénarios : $\delta = 0$ et $\delta \neq 0$. Dans ce dernier cas, la valeur de δ est fixée de façon a assurer au test individuel de Student une puissance de 80%, soit ici $\delta = 1.19$.

L'annexe A propose le code - permettant de simuler les tabkeaux de donénes de cet exemple.

Nous allons utiliser les données de l'EXEMPLE 5 pour étudier l'intérêt de ce test sur les erreurs de type-I et II. L'erreur de type-I est estimée en comptant, sur l'ensemble des simulations, la proportion d'hypothèses nulles rejetées alors que les données ont été simulées sous H_0, c'est-à-dire sans différence entre les moyennes des deux groupes pour Y (scénario $\delta = 0$) et pour Z. Le taux moyen d'erreur de type-II sera estimé en comptant, sur l'ensemble des simulations, la proportion d'hypothèses nulle non rejetées alors que les données ont été simulées sous H_1, c'est-à-dire sans différence entre les moyennes des deux groupes pour Z et avec une différence $\delta \neq 0$ pour Y.

Ici, la statistique de test (3.2) s'exprime de la façon suivante :

$$\tilde{T} = \frac{(\bar{Y}^{(B)} - \bar{Y}^{(A)}) - \frac{\rho_{YZ}}{\sigma_Z^2}(\bar{Z}^{(B)} - \bar{Z}^{(A)})}{\sqrt{(\frac{1}{n_A} + \frac{1}{n_B})\sigma^2_{Y|Z}}}$$



(a) Erreur de type I (% rejets sous H_0)

(b) Erreur de type II (% non rejets sous H_1)

FIG. 3.1 – Comparaison du test de Student et du test ajusté : évolution des erreurs de type-I et II en fonction de ρ (10 000 jeux de données de deux variables (Y et Z) simulés pour chaque valeur de ρ) - Seuil de rejet $\alpha = 5\%$

La FIGURE 3.1(a) illustre l'apport de la variable Z sur l'erreur de type-I, en fonction de différentes valeurs de ρ_{YZ} (de $0,1$ à $0,9$). On montre ainsi que le test proposé est semblable au test de Student en ce qui concerne les erreurs de type-I : quelle que soit la corrélation entre les variables, le test de Student et le test ajusté permettent de maintenir un nombre de rejets à tort des hypothèses nulles à un seuil de 5%. Par contre, le test ajusté permet d'améliorer la puissance par rapport au test de Student, et ce d'autant plus lorsque la corrélation entre les variables est importante, comme le montre le graphique 3.1(b) : en effet, le nombre d'hypothèses alternatives non rejetées diminue lorsque la corrélation augmente. Lorsque les variables sont très peu corrélées, les deux tests sont équivalents.

Ainsi, le test présenté permet de tirer profit, dans le test de l'effet de x sur Y, de l'information

apportée par Z. L'avantage est d'autant plus important que la corrélation entre Y et Z est élevée. En appliquant ce principe aux tests multiples, on peut espérer améliorer globalement la puissance des procédures.

Les facteurs Z sont, jusqu'ici, des variables aléatoires observables. On se sert de ce cas favorable, étudié en particulier dans le cadre des essais cliniques [Davidian et al., 2005] ou dans le cas du double-échantillonnage citepkloacaus07 par exemple, pour introduire par la suite le cas où il s'agit de facteurs latents, et donc non-observables. En effet, nous n'avons pas toujours à disposition de covariables Z sous forme d'information externe qui vérifient les propriétés nécessaires pour que la correction du test soit optimale, c'est-à-dire indépendantes de la condition expérimentale x mais surtout les plus corrélées possibles aux variables d'intérêts, conditionnellement à x. Dans le cas où m est de l'ordre de plusieurs milliers, les données comportent un grand nombre de variables pour lesquelles l'hypothèse nulle est vraie. Nous proposons alors de tirer profit de ce grand ensemble de variables pour définir des statistiques de test ajustées, en s'inspirant de (3.2).

Dans la suite, l'idée est d'utiliser les facteurs extraits de la structure de dépendance des variables d'intérêt conditionnellement à X comme variables d'ajustement Z dans l'estimation des effets fixes du modèle. Au sein des variables de $\mathcal{M}_0$, il s'agit de définir celles qui seront utilisées comme "covariables" Z. Le test ajusté sera d'autant plus puissant que ces variables seront fortement corrélées aux variables à tester.

3.1.1.2 Utilisation des facteurs comme covariables

On se place dans le cadre défini par la PROPOSITION 2.0.1. Chaque variable Y_k s'écrit donc d'après le modèle (2.1) :

$$Y_k = m_k(x) + Zb_k' + \varepsilon_k$$

Le noyau de dépendance ZB' est une composante indépendante des covariables X. On se propose de centrer les données Y_k par rapport à ce noyau de dépendance :

$$\tilde{Y_k} = Y_k - Zb_k' = m_k(x) + \varepsilon_k \tag{3.3}$$

Le modèle défini en (3.3) permet de se ramener à un modèle de régression multivarié dont le vecteur des résidus $E = [\varepsilon_1, \ldots, \varepsilon_m]$ est constitué de composantes indépendantes. On s'intéresse toujours au lien entre la variable Y_k et les covariables X à travers les m tests simultanés de

$$\begin{cases} H_0^k : & m_k(x) = m_k^{(0)}(x) \\ H_1^k : & m_k(x) \neq m_k^{(0)}(x) \end{cases}$$

On base alors la décision du test sur des statistiques de tests conditionnelles, ajustées de l'effet de la dépendance par rapport aux statistiques de tests usuelles $T_k = s(Y_k)$.

DÉFINITION 3.1.1 (Statistique de test et probabilité critique ajustées)**.**

$$\begin{aligned} \tilde{T_k} &= s(\tilde{Y_k}) = s(Y_k - Zb_k') \\ \tilde{p_k} &= 1 - F_0(\tilde{T_k}) \end{aligned}$$

PROPOSITION 3.1.1 (Probabilités critiques ajustées). *Soit $\tilde{G}^k(t) = \mathbb{P}(\tilde{p}_k \leq t)$ la fonction de répartition des probabilités critiques ajustées définies précédemment. On suppose que la densité des erreurs du modèle* (3.3) *est une certaine loi ϕ_k, qui ne diffèrent que d'un facteur d'échelle ψ_k, et qui ont pour fonction standardisée φ, comme pour le modèle* (1.1) *: $\phi_k(\varepsilon_k) = \varphi(\varepsilon_k/\psi_k)/\psi_k$.*

On a alors :

$$\begin{aligned} \tilde{p}_k &\sim \tilde{G}_0(t) \equiv G_0(t) = t, \forall k \in \mathcal{M}_0 \\ \tilde{p}_k &\sim \tilde{G}_1^k(t) \equiv G_1(t; \tilde{\tau}_k), \forall k \in \mathcal{M}_1 \end{aligned}$$

où τ_k et $\tilde{\tau}_k$ sont les paramètres de non-centralité sous l'hypothèse alternative respectivement pour la distribution usuelle et pour la distribution conditionnelle : $\tilde{\tau}_k = \tau_k.\sigma_k/\psi_k$

Démonstration. On peut ré-écrire $\tilde{G}^k(t)$ tel que :

$$\tilde{G}^k(t) = \int \mathbb{1}_{[1-F_0(s(m_k(x)-\varepsilon_k))\leq t]}\phi_k(\varepsilon_k)d\varepsilon_k$$

On a : $\phi_k(\varepsilon_k) = \varphi(\varepsilon_k/\psi_k)/\psi_k$. On pose $u_k = \varepsilon_k/\psi_k$, et donc $d\varepsilon_k = \psi_k du_k$.

$$\tilde{G}^k(t) = \int \mathbb{1}_{[1-F_0(s(m_k(x)/\Psi_k-u_k))\leq t]}\varphi(u_k)du_k = G^k(t; \tilde{\tau}_k)$$

De plus, $\tilde{\tau}_k = m_k(x)/\psi_k = \tau_k\sigma_k/\psi_k$. □

Étude de la puissance σ_k^2 et ψ_k^2 correspondent respectivement à la variance de Y_k et à celle de ε_k dans le modèle (2.1), qui représente la variabilité spécifique à Y_k selon la décomposition de la variabilité donnée par le modèle d'Analyse en Facteurs (2.3). On a alors $\sigma_k^2 \geq \psi_k^2$. Le paramètre de non-centralité de la distribution des probabilités critiques ajustées $(\tilde{\tau}_k)$ est donc plus élevé que celui des probabilités critiques non-ajustées (τ_k). Le test réalisé à partir des probabilités critiques ajustées est donc plus puissant que le même test dont la décision est basée sur les probabilités critiques non-ajutées.

De plus, notons que la puissance individuelle dépend de la contribution de chaque variable à la variabilité commune. Dans le cas de variables proches de l'indépendance, la structure en facteurs est faible et $\psi_k^2 \approx \sigma_k^2$. Dans ce cas, les résultats obtenus avec les probabilités ajustées de l'effet des facteurs seront semblables à ceux obtenus avec les probabilités critiques usuelles. Dans le cas d'une structure de dépendance marquée, avec une part de variabilité commune à l'ensemble des variables importante, $\exists k \in \mathcal{M}, \psi_k^2 << \sigma_k^2$. On profite donc de cette information commune pour augmenter la puissance des procédures de tests multiples.

3.1.1.3 Cas du modèle linéaire

La SECTION 1.1.2 présente le cas du modèle linéaire comme un cas usuel du modèle (1.1). Les statistiques de test utilisées dans ce cadre sont celles de tests de Student (DÉFINITION 1.1.1) et possèdent les propriétés suivantes :

Proposition 3.1.2. *Propriétés de* $T = [T_1, \ldots, T_m]$

$$T \sim \mathcal{T}_{n-p-1}(\tau, R)$$

où $\tau = [\tau_1; \ldots; \tau_m]$ *et* $R = cor(Y|X)$
On note que $\tau_k = 0$ *si* $k \in \mathcal{M}_0$, *et* $\tau_k = \frac{\sqrt{n}c'\theta_k}{\sigma_k\sqrt{c'S_x^{-1}c}}$ *sinon.*

Dans le cadre définit par la proposition 2.0.1, on s'intéresse maintenant aux propriétés conditionnelles de $T = [T_1, \ldots, T_m]$. On définit pour cela la quantité $\tau_Z = s(Z) = c'\hat{\beta}_Z/\sqrt{n^{-1}c'S_X^{-1}c}$, où $\hat{\beta}_Z$ est le $Q \times 1$-vecteur des estimateurs des Moindres Carrés des coefficients de pente de la régression multivariée de Z sur X. De plus, $\tau_Z \sim \mathcal{N}(0; \mathbb{I}_Q)$.

Proposition 3.1.3. *Propriétés conditionnelles de* T_k

$$\begin{aligned} \mathbb{E}(T_k|Z) &= \tau_k + \frac{b_k\tau_Z}{\sigma_k} \\ \mathbb{V}(T_k|Z) &= \psi_k^2/\sigma_k^2 \\ \mathrm{Cov}(T_k, T_{k'}|Z) &= 0 \end{aligned}$$

Démonstration. Les propriétés conditionnelles de T sont déduites des propriétés conditionnelles de $\hat{\theta} = (X'X)^{-1}X'Y$, avec $Y = X\theta + ZB' + E$.

$$\begin{aligned} \mathbb{E}(\hat{\theta}|Z) &= \mathbb{E}((X'X)^{-1}X'(X\theta + ZB')|Z) \\ &= \mathbb{E}((X'X)^{-1}X'X\theta|Z) + \mathbb{E}((X'X)^{-1}X'ZB'|Z) \\ &= \theta + \beta_Z B' \\ \mathbb{V}(\hat{\theta}|Z) &= \mathbb{V}((X'X)^{-1}X'Y|Z) \\ &= (X'X)^{-1}X'\mathbb{V}(Y|Z)X(X'X)^{-1} \\ &= \Psi(X'X)^{-1}X'X(X'X)^{-1} \\ &= \Psi n^{-1}S_X^{-1} \end{aligned}$$

On en déduit l'espérance et la variance de T_k :

$$\begin{aligned}
\mathbb{E}(T_k|Z) &= \mathbb{E}\left(\frac{c'\hat{\theta}_k}{\sqrt{\mathbb{V}(c'\hat{\theta}_k)}}|Z\right) = \mathbb{E}\left(\frac{c'\hat{\theta}_k}{\sqrt{\sigma_k^2 n^{-1} c' S_X^{-1} c}}|Z\right) \\
&= \frac{c\theta_k}{\sqrt{\hat{\sigma}_k^2 n^{-1} c' S_X^{-1} c}} + \frac{c\beta_Z b_k}{\sqrt{\sigma_k^2 n^{-1} c' S_X^{-1} c}} \\
&= \tau_k + \frac{b_k' \tau_Z}{\sigma_k} \\
\mathbb{V}(T_k|Z) &= \mathbb{V}\left(\frac{c'\hat{\theta}_k}{\sqrt{\mathbb{V}(c'\hat{\theta}_k)}}|Z\right) = \frac{\mathbb{V}\left(c'\hat{\theta}_k|Z\right)}{\sigma_k^2 n^{-1} c' S_X^{-1} c} \\
&= \frac{\psi_k^2 n^{-1} c' S_X^{-1} c}{\hat{\sigma}_k^2 n^{-1} c' S_X^{-1} c} = \frac{\psi_k^2}{\sigma_k^2}
\end{aligned}$$

L'indépendance conditionnelle des statistiques de test se déduit donc de l'indépendance conditionnelle des variables Y (Ψ est une matrice diagonale). □

On montre que la statistique de test conditionnelle $\tilde{T}_k$ se définit à partir de la statistique classique $T_k$par centrage et réduction conditionnellement à la structure en facteurs (PROPOSITION 3.1.3).

DÉFINITION 3.1.2.

$$\forall k \in \mathcal{M},\ \tilde{T}_k = \frac{\sigma_k}{\psi_k}\left[T_k - \frac{b_k' \tau_Z}{\sigma_k}\right]$$

Les propriétés de $\tilde{T} = [\tilde{T}_1, \ldots, \tilde{T}_m]$ sont décrites dans la proposition suivante :

PROPOSITION 3.1.4. *Sous l'hypothèse du modèle* (2.1)*, $\tilde{T}$ suit une loi normale multivariée, avec :*

$$\begin{aligned}
\mathbb{E}(\tilde{T}_k) &= \tau_k \frac{\sigma_k}{\psi_k}, \forall k \in \mathcal{M} \\
\mathbb{V}(\tilde{T}) &= \mathbb{I}_m
\end{aligned}$$

Démonstration. Les statistiques de test (DÉFINITION 3.1.2) sont des combinaisons linéaires des résidus du modèle (3.3) : $\varepsilon_k = (Y_k - Zb_k') - m_k(x)$, qui sont distribués selon une loi normale. Ainsi, $\tilde{T}$ est distribué selon une loi normale multivariée. D'après les propriétés conditionnelles des statistiques usuelles T_k énoncées dans la PROPOSITION 3.1.2,

$$\begin{aligned}
\mathbb{E}(\tilde{T}_k) &= \left(\mathbb{E}(T_k|Z) - \frac{b_k}{\sigma_k}\tau_Z\right)\sqrt{\frac{\sigma_k}{\psi_k}} = \tau_k\sqrt{\frac{\sigma_k}{\psi_k}} \\
\mathbb{V}(\tilde{T}_k) &= \left(\mathbb{V}(T_k|Z)\right)\frac{\sigma_k}{\psi_k} = \mathbb{I}_m
\end{aligned}$$

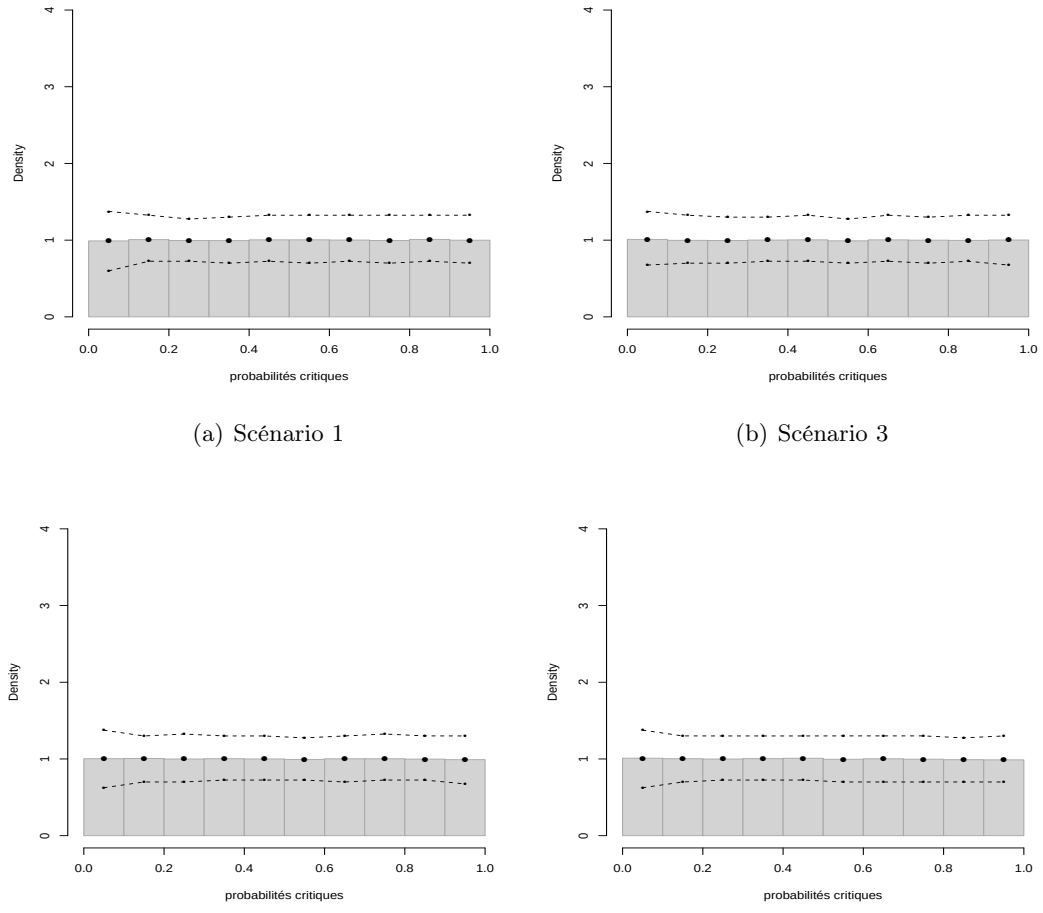
□

Au CHAPITRE 4, nous proposons de considérer les estimateurs du Maximum de Vraisemblance des paramètres du modèle (2.1). Par construction, il s'agit d'estimateurs consistants, l'estimation des paramètres de variance n'affecte donc pas la distribution asymptotique de la statistique de test $\tilde{T}$.

Dans le cas de petits échantillons, on va prendre en compte l'effet de l'estimation des paramètres de variance en considérant sous l'hypothèse nulle une distribution de Student, par analogie avec le test classique. Les degrés de liberté de cette loi sont ceux associés à la résiduelle du modèle (voir (4.7)).

Distribution des probabilités critiques ajustées sous H_0 Reprenons ici les données simulées dans le cadre de l'EXEMPLE 3 pour l'étude de l'impact de la dépendance sur la distribution des probabilités critiques.

Sur chaque jeu de données, les $m = 500$ tests de l'effet de la variable X sont réalisés, en considérant les statistiques de tests ajustées et les probabilités critiques associées (DÉFINITION 3.1.1).



(a) Scénario 1

(b) Scénario 3

(c) Scénario 6

(d) Scénario 9

FIG. 3.2 – Distribution des probabilités critiques ajustées - EXEMPLE 3

D'après la FIGURE 3.2, la distribution moyenne des probabilités critiques est uniforme, quel que soit le niveau de dépendance. De plus, contrairement aux résultats obtenus à l'EXEMPLE 3 dans le cas des

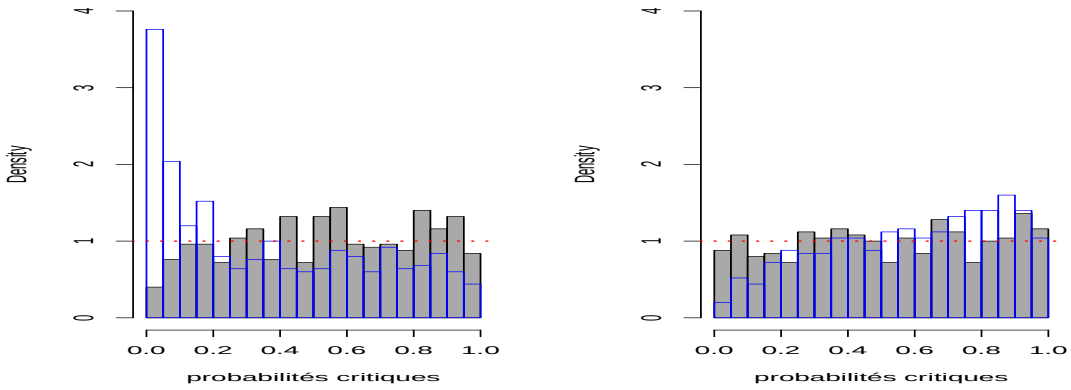


FIG. 3.3 – Exemples de deux jeux de données : distribution des probabilités critiques ajustées (en gris) - scénario 9. En bleu : histogramme des probabilités critiques usuelles

statistiques de test usuelles (FIGURE 2.2), les histogrammes obtenus pour les probabilités critiques ajustées varient peu autour de l'histogramme moyen, et ce pour chacun des scénarios.

Sur la FIGURE 3.3, on représente l'histogramme des deux tableaux de données simulés (scénario 9) de la FIGURE 2.3. En présence d'un niveau de dépendance élevé, les distributions des probabilités critiques se démarquent peu de la loi uniforme (loi théorique, en pointillé rouge) par rapport aux probabilités critiques usuelles. Le phénomène de sur-/sous-représentation des probabilités critiques proches de 0 ou proches de 1 observé en présence de dépendance est fortement minimisé ici.

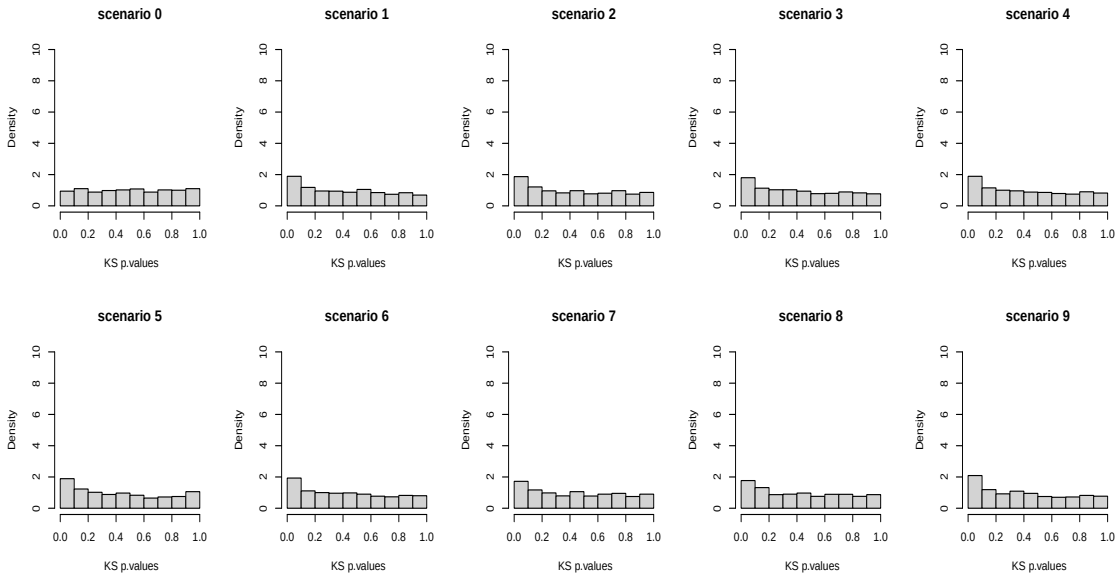


FIG. 3.4 – Distribution des probabilités critiques des tests de Kolmogorov-Smirnov obtenues pour les tests d'uniformité des m probabilités critiques ajustées des 1000 jeux de données simulés de chaque scénarios

On peut également tester l'uniformité des m probabilités critiques, pour les 1000 tableaux de donénes de chaque scénarios à l'aide d'un test de Kolmogorov-Smirnov. La FIGURE 3.4 représente les probabilités critiques de ces tests. Elles sont elles même réparties uniformément (test d'uniformité sur ces probabilités critiques non significatif au seuil $0,05$) pour les scénarios 0 et 1. Pour les autres scénarios, l'écart à l'uniformité des probabiliéts critiques des tests de Kolmogorov Smirnov est significatif, mais cet écart est nettement moindre que dans le cas des tests réalisés sur les probabilités critiques usuelles (FIGURE 2.4).

Étude de la puissance On a $\tau_k \frac{\sigma_k}{\psi_k} \geq \tau_k$ car $\sigma_k \geq \psi_k, \forall k \in \mathcal{M}$. La statistique de test ajustée $\tilde{T}_k$ permet donc un test individuel pour la variable k qui sera plus puissant que le test de Student.

De plus, le paramètre de non-centralité de la statistique de test $\tilde{T_k}$ est proportionnel à la communalité de la variable k : $h_k^2 = b_k' b_k / \sigma_k^2$:

$$\mathbb{E}(\tilde{T_k}) = \tau_k \sqrt{\frac{\sigma_k}{\psi_k}} = \frac{\tau_k}{\sqrt{1 - h_k^2}}$$

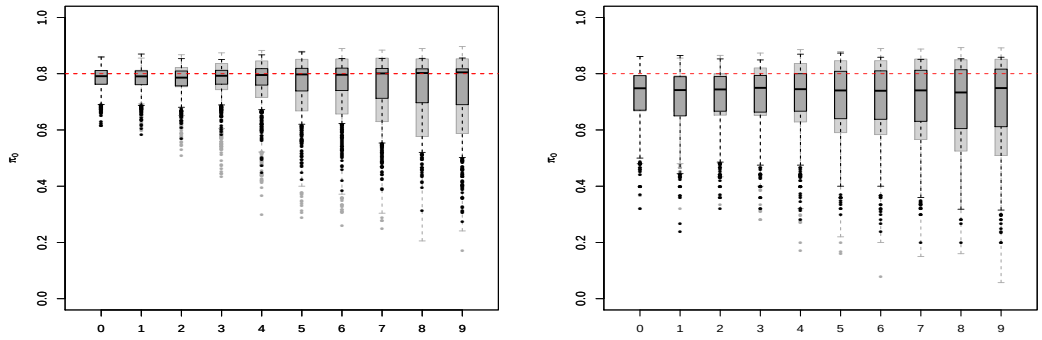
Ainsi, on améliore le test pour la variable k si sa contribution à la structure commune est importante et si cette variable a une contribution nulle, alors la statistique de test coïncide avec la statistique usuelle T_k.

Les statistiques de tests ajustées des facteurs sont donc indépendantes conditionnellement à la structure en facteurs, ce qui permet de se placer dans un cadre optimal pour l'utilisation des méthodes usuelles de tests multiples. Au CHAPITRE 2 un terme directement fonction de la corrélation dans les expressions des variances de π_0 et du nombre de faux-positifs a été identifié. Travailler à partir des données ajustées des facteurs permet d'atteindre un cadre d'indépendance conditionnelle qui corrige de l'effet de la dépendance dans l'estimation de ces quantités. Ce point est abordé dans les sections suivantes.

3.1.2 Estimation de la proportion d'hypothèses nulles

On se propose d'appliquer les méthodes d'estimation de π_0 décrites à la SECTION 1.2 sur les probabilités critiques ajustées, calculées sur les données simulées de l'EXEMPLE 4. Les résultats pour deux méthodes d'estimation de π_0 sont présentés sur les FIGURES 3.5 et 3.6, les résultats pour les autres méthodes sont donnés en ANNEXE D. Les mêmes méthodes d'estimations appliquées sur les probabilités critiques usuelles sont représentées en gris, pour rappel (voir FIGURE 2.8).

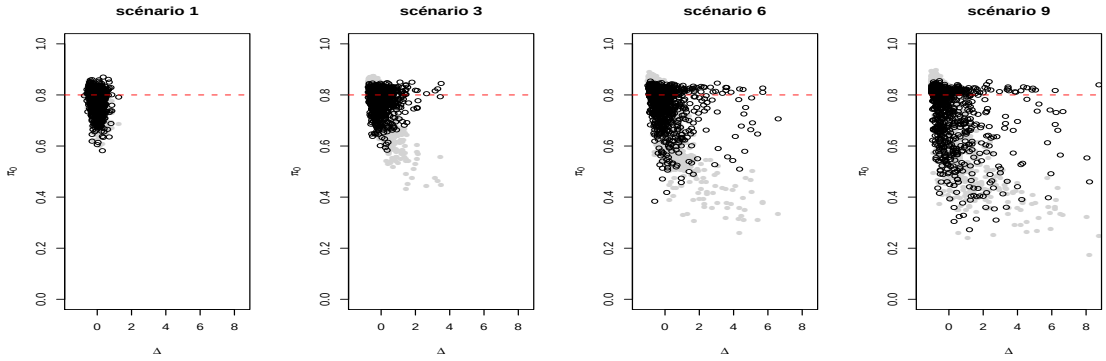
Pour les deux types de méthodes d'estimation de π_0, la variabilité d'estimation est fortement diminuée. Lorsque le paramètre est estimé à partir d'une estimation de la densité des probabilités critiques, la stabilisation de la distribution sous H_0 permet de diminuer l'erreur d'estimation aux alentours de $p = 1$ et donc de diminuer la variabilité d'estimation de π_0. D'autre part, dans le cadre d'indépendance conditionnelle, le terme dépendant de la corrélation dans l'expression de la variance de l'estimateur empirique de π_0 est nul et pour tous les scénarios de dépendance, la variance de l'estimation de $\hat{\pi}_0$ est stabilisé.



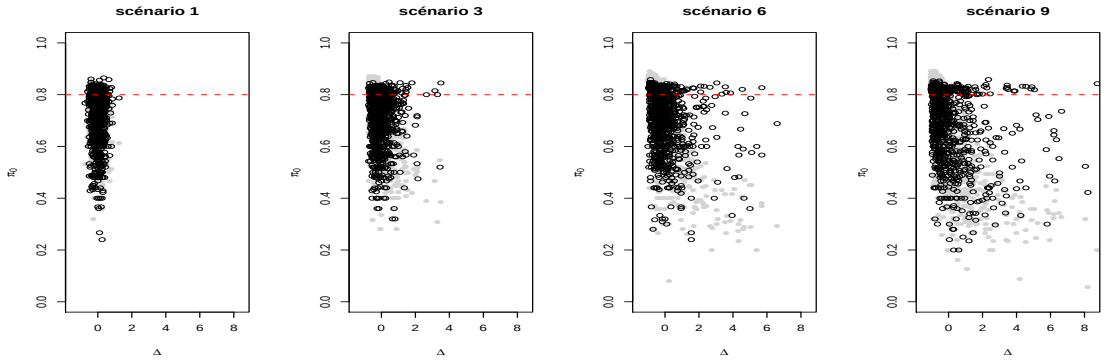
(a) Estimation de la densité par une fonction convexe

(b) Estimateur empirique avec choix de λ par *bootstrap*

FIG. 3.5 – Estimations de π_0 à partir de probabilités critiques ajustées sur des données simulées selon différents scénarios de dépendance - en gris : Mêmes méthodes d'estimation à partir des probabilités critiques usuelles (FIGURE 2.5) - $\pi_0 = 0,80$



(a) Estimation de la densité par une fonction convexe



(b) Estimateur empirique avec choix de λ par *bootstrap*

FIG. 3.6 – Estimations de π_0 à partir de probabilités critiques ajustées (en noir) sur des données indépendantes (scénario1), modérément (scénario 3 et scénario 6) ou très corrélées (scénario 9) avec deux méthodes. En gris : mêmes méthodes d'estimations à partir des probabilités critiques usuelles (FIGURE 2.8) - $\pi_0 = 0,80$

La variance de $\hat{\pi}_0$ comporte en effet un terme qui apparait être la somme des valeurs de la fonction $D^{kk'}(t)$ calculée pour les corrélations entre l'ensemble des variables. Les corrélations résiduelles (une fois le modèle d'analyse en facteurs ajusté) sont nulles. La fonction $D^{kk'}(t)$ est nulle pour $\rho_{kk'} = 0$, la somme des valeurs l'est donc aussi dans ce cas.

3.1.3 Contrôle du FWER et du FDR

FWER Une expression explicite du FWER, dans le cadre d'une approximation à un facteur de la structure de corrélation entre les statistiques de test est donnée par Hsu [Hsu, 1992, Hsu and Nelson, 1998]. Cette approximation semble appropriée en particulier en analyse de variance, pour les tests *post-hoc* (tests simultanés des contrastes linéaires suite au test global de l'analyse de variance).

Dans notre cas, seul un contraste est testé pour chacune des variables d'intérêt. La multiplicité des tests est dûe non pas au nombre de contrastes testés simultanément mais au nombre de variables réponses. La dépendance entre les statistiques de test n'est pas déduite du plan d'expérience mais provient de la corrélation entre les variables réponses elles mêmes. L'utilisation d'une approximation à un facteur de la structure de dépendance peut alors apparaitre trop simple étant donnée la dimension et la complexité des données étudiées.

Nous proposons ici d'estimer le FWER à partir des probabilités critiques ajustées (DÉFINITION 3.1.1) pour les données simulées de l'EXEMPLE 4. Le TABLEAU 3.1 donne les fréquences du nombre de faux-positifs pour les différents scénarios de dépendance. L'utilisation des probabilités critiques ajustées permet d'assurer un contrôle du FWER proche du seuil souhaité ($\alpha = 0,05$).

scénario	0	1	2	3	4	5	6	7	8	9
var. commune (%)	0	4,65	15,56	28,56	40,31	50,76	57,90	65,45	70,09	75,19
fréquences des valeurs de V_t										
0	957	954	965	951	960	966	967	963	960	955
1	42	45	34	49	39	32	30	35	36	34
2		1	1		1	2	3	1	3	9
3	1							1	1	1
4										1
FWER										
$\#\{V_t > 0\}/m$	4,3	4,6	3,5	4,9	4,0	3,4	3,3	3,7	4,0	4,5

TAB. 3.1 – FWER estimé pour les 10 scénarios de données simulées (résultats en %) et tableau des fréquences observées pour les valeurs de V_t - Procédure de Sidak [Sidak, 1967] sur les probabilités critiques ajustées, avec un niveau α fixé à $0,05$ pour le risque de type-I

Une autre propriété frappante de l'utilisation des probabilités critiques ajustées pour l'analyse est le gain en puissance obtenu en présence de dépendance (FIGURE 3.7).

Des résultats similaires sont obtenus pour la procédure de Bonferroni.

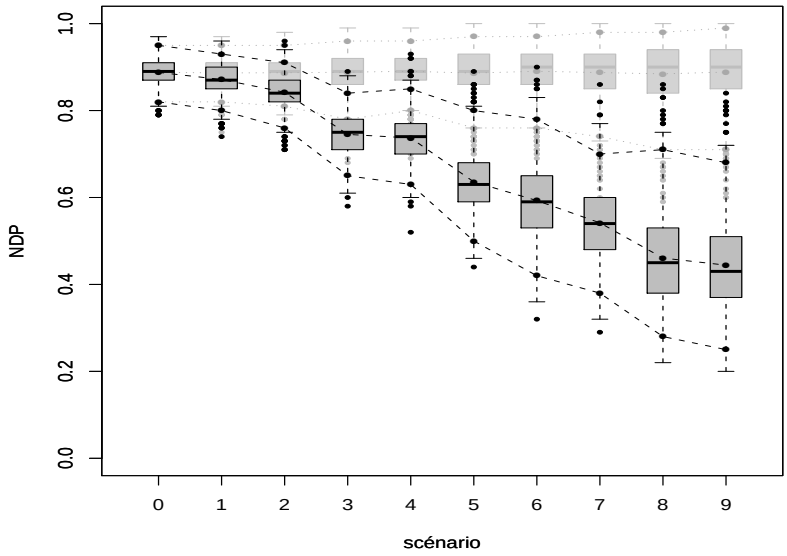
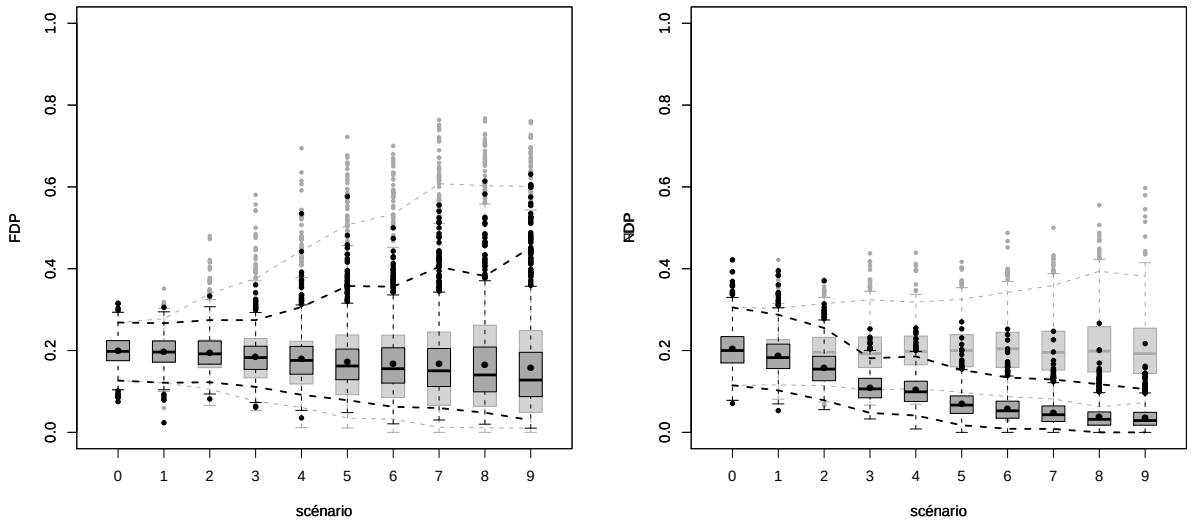


FIG. 3.7 – Nombre de non-découvertes (Non Discovery Proportion) en fonction du niveau de dépendance pour les 10 scénarios - Procédure de Sidak sur les probabilités critiques ajustées. En gris : résultats obtenus à partir des tests de Student (FIGURE 2.14)



(a) FDP (b) NDP

FIG. 3.8 – Proportion de faux-positifs (FDP) et de faux-négatifs (NDP), en fonction du niveau de dépendance pour les 10 scénarios - Procédure BH sur les probabilités critiques ajustées. En gris : résultats obtenus à partir des tests de Student (FIGURE 2.16) - niveau α fixé à $0,2$ pour le risque de type-I

FDR La FIGURE 3.8 représente pour chaque scénario la distribution de la proportion de faux-positifs (FDP) et de faux-négatifs (NDP) suite à l'application de la procédure BH sur les probabilités critiques ajustées (DÉFINITION 3.1.1) pour les données simulées de l'EXEMPLE 4. En gris clair sont représentées les distributions obtenues pour la FDP et la NDP obtenus à partir des probabilités critiques usuelles (voir FIGURE 2.16). Le seuil α pour le contrôle du FDR est ici de $0,20$. La distribution de la FDP est clairement stabilisée. La variabilité en présence de dépendance est donc fortement réduite par rapport à l'application de la méthode aux probabilités critiques usuelles (FIGURE 3.8(a)). Comme pour le FWER dans la section précédente, l'utilisation des probabilités critiques ajustées permet un gain en puissance en présence de dépendance (FIGURE 3.8(b)).

La dépendance entre variables n'est donc pas uniquement un facteur perturbateur des propriétés des procédures de tests multiples, mais également une information sur laquelle on peut s'appuyer pour améliorer la règle de décision.

3.2. Estimateurs conditionnels

Les expressions explicites des propriétés des estimateurs de π_0 et du FDR en présence de dépendance permettent d'envisager également des approches conditionnelles.

3.2.1 Estimation conditionnelle de π_0

Le biais de l'estimateur empirique est donné d'après la PROPOSITION 2.2.2 par : $\mathcal{B}(\hat{\pi}_0) = (1-\pi_0)\frac{1-\bar{G}_1(\lambda)}{1-\lambda}$ où $G_1^k(\lambda)$ désigne la fonction de répartition de la probabilité critique p_k pour $k \in \mathcal{M}_1$ et $\bar{G}_1(\lambda) = \frac{1}{m_1}\sum_{k\in\mathcal{M}_1}\left(G_1^k(\lambda)\right)$. On s'intéresse ici au biais conditionnel de l'estimation de π_0. Pour cela, on réécrit l'espérance conditionnelle de $\hat{\pi}_0(\lambda)$:

$$\begin{aligned}
\mathbb{E}\left(\hat{\pi}_0(\lambda)|Z\right) &= \frac{\mathbb{E}\left(W_\lambda|Z\right)}{m(1-\lambda)} = \frac{\sum_{k\in\mathcal{M}}\left(1-G^k(\lambda,Z)\right)}{m(1-\lambda)} = \frac{\sum_{k\in\mathcal{M}}\left([G^k(\lambda)-G^k(\lambda)]+[1-G^k(\lambda,Z)]\right)}{m(1-\lambda)} \\
&= \frac{\sum_{k\in\mathcal{M}}\left(1-G^k(\lambda)\right)}{m(1-\lambda)} + \frac{\sum_{k\in\mathcal{M}}\left(G^k(\lambda)-G^k(\lambda,Z)]\right)}{m(1-\lambda)} \\
&= \mathcal{B}(\hat{\pi}_0) + \frac{\sum_{k\in\mathcal{M}}\left(G^k(\lambda)-G^k(\lambda,Z)]\right)}{m(1-\lambda)} \qquad (3.4)
\end{aligned}$$

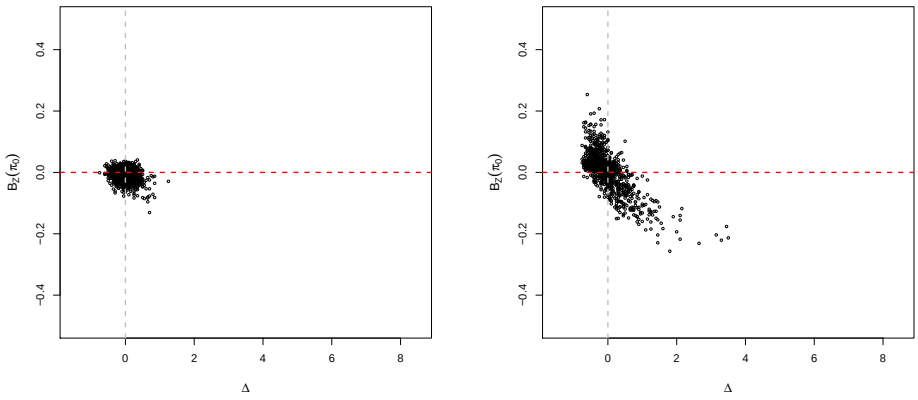
On note $\mathcal{B}_Z(\hat{\pi}_0) = \frac{\sum_{k\in\mathcal{M}}\left(G^k(\lambda)-G^k(\lambda,Z)]\right)}{m(1-\lambda)} = \frac{\bar{G}(\lambda,Z)}{(1-\lambda)}$, la partie aléatoire du biais de l'estimateur empirique conditionnel. La variable aléatoire $\bar{G}(\lambda,Z)$ définie précédemment est d'espérance nulle.

Nous proposons donc de corriger l'estimateur empirique de ce biais conditionnel :

DÉFINITION 3.2.1 (Estimateur conditionnel de π_0)**.**

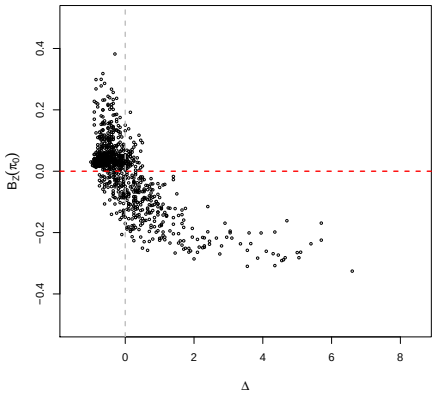
$$\check{\pi}_0(\lambda) = \hat{\pi}_0(\lambda) - \mathcal{B}_Z(\hat{\pi}_0)$$

où $\mathcal{B}_Z(\hat{\pi}_0) = \frac{\bar{G}(Z,\lambda)}{(1-\lambda)}$.

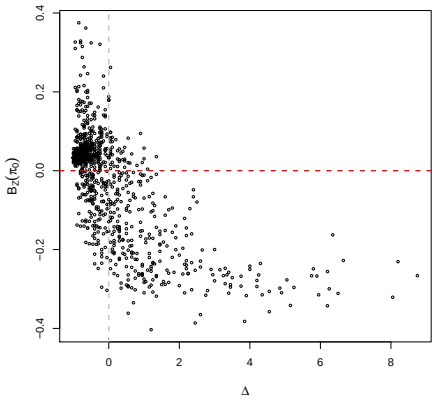


(a) Scénario 1

(b) Scénario 3



(c) Scénario 6



(d) Scénario 9

FIG. 3.9 – Représentation graphique de $\mathcal{B}_Z(\hat{\pi}_0)$ en fonction du critère Δ qui caractérise l'impact de la dépendance sur la forme de l'histogramme des probabilités critiques sous l'hypothèse nulle, pour quatre niveaux de dépendance (scénarios 1, 3, 6 et 9). Le paramètre λ est celui obtenu par bootstrap pour l'estimation de π_0

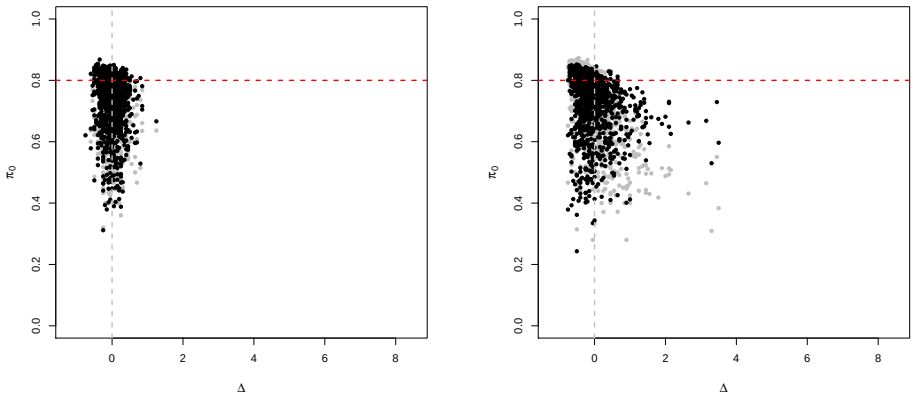
En pratique, l'estimation du biais conditionnel $\mathcal{B}_Z(\hat{\pi}_0)$ nécessite de connaitre $\mathcal{M}_0$. L'approximation de cet ensemble est un problème récurent dans les procédures de tests multiples. Plusieurs stratégies sont proposées (voir remarque dans la SECTION 3.2.2). Nous proposons simplement de prendre les indices des variables dont la probabilité critique ajustée est inférieure à $0,05$.

3.2.2 Estimateur conditionnel du FDR

Dans le cadre de la PROPOSITION 2.0.1, on peut définir l'espérance conditionnelle du nombre de faux-positifs :

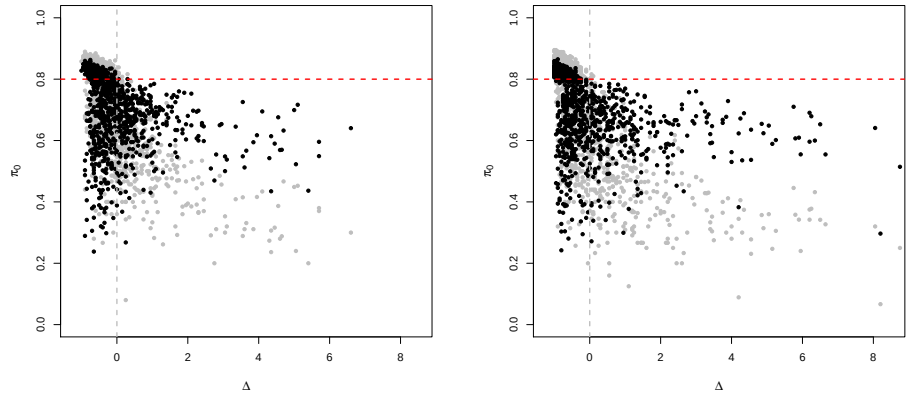
$$\mathbb{E}(V_t|Z) = \sum_{k\in\mathcal{M}_0} \mathbb{P}\left(p_k \leq t|Z\right) = \sum_{k\in\mathcal{M}_0} G^k(t,Z) \tag{3.5}$$

$$\tag{3.6}$$



(a) Scénario 1

(b) Scénario 3



(c) Scénario 6

(d) Scénario 9

FIG. 3.10 – Estimateur conditionnel $\breve{\pi}_0$ en fonction du critère Δ qui caractérise l'impact de la dépendance sur la forme de l'histogramme des probabilités critiques sous l'hypothèse nulle, pour quatre niveaux de dépendance (scénarios 1, 3, 6 et 9). Le paramètre λ est celui obtenu par bootstrap pour l'estimation de π_0

Cette expression conditionnelle permet de définir un estimateur conditionnel pour le taux de faux-positifs :

DÉFINITION 3.2.2 (FDR conditionnel)**.**

$$\begin{aligned} FDR_t^Z &= \frac{\mathbb{E}(V_t|Z)}{R_t} = \frac{\sum_{k\in\mathcal{M}_0}(G^k(t,Z))}{R_t} \\ &= \frac{m_0 t}{R_t} + \frac{\sum_{k\in\mathcal{M}_0}(G^k(t,Z))}{R_t} - \frac{m_0 t}{R_t} \\ &= \widehat{FDR}_t.\left[1 + \frac{\sum_{k\in\mathcal{M}_0}(G^k(t,Z)-t)}{m_0 t}\right] \end{aligned}$$

où $\widehat{FDR}_t = \frac{m_0 t}{R_t}$ *représente l'estimateur standard du taux de faux-positifs.*

Cette définition est à mettre en relation avec un estimateur proposé par Efron [2007] : $FDR_t^A = \mathbb{E}(V_t|A)/R_t = \widehat{FDR}_t\left[1 + A\frac{u_t\phi(u_t)}{\sqrt{2}(1-\Phi(u_t))}\right]$. ϕ et Φ représentent respectivement la densité et la fonction de répartition de la loi Normale centrée-réduite. A est un critère[1] quantitatif caractérisant la distribution des statistiques de test. C'est un indicateur de la dispertion des Z-scores : il est nul si les statistiques sont indépendantes (Z-scores distribués selon une loi Normale centrée réduite), et varie autour de zéro en présence de corrélation (rétrécissement ou aplatissement de la distribution des Z-scores par rapport à la loi Normale centrée réduite).

En notant $Y_0 = \#\{Z_k \in [-x_0; x0]\}$, $P_0 = 2\Phi(x_0) - 1$ et $Q_0 = \sqrt{2}x_0\phi(x_0)$, A est estimé par :

$$\hat{A} = \frac{P_0 - Y_0/n}{Q_0}$$

Remarque : Lorsque m_0 est inconnu, il est alors remplacé par m, comme pour l'estimateur FDR_t^A ou dans la procédure BH, ou bien estimé par une des méthodes présentées précédemment (voir SECTION 1.2).

L'estimateur conditionnel, ainsi que celui définit par Efron [2007], apparaissent tous deux comme des corrections de l'estimateur usuel, prenant en compte la dépendance des données. Cette prise en compte se fait grâce aux facteurs Z du noyau de dépendance du modèle (2.1), et à travers le critère A pour FDR_t^A.

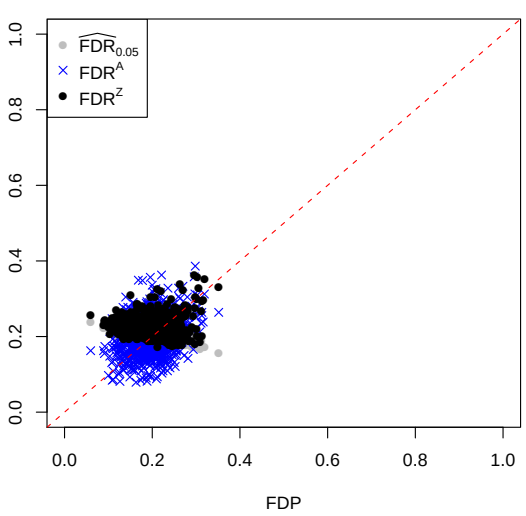
Remarque : Ces deux estimateurs nécessitent de connaitre $\mathcal{M}_0$, ou du moins une bonne approximation de cet ensemble. Efron [2007] suggère pour cela de calculer le critère A à partir des variables pour lesquelles le *Z-score* prend une valeur dans $[-1; 1]$ ($x_0 = 1$). De manière similaire, on propose de prendre $\widehat{\mathcal{M}}_0 = \{k : p_k \geq 0,05\}$, où p_k est la probabilité critique usuelle.

scénario	0	1	2	3	4	5	6	7	8	9
var. commune (%)	0	4,65	15,56	28,56	40,31	50,76	57,90	65,45	70,09	75,19
$\widehat{FDR}_t$	-0,197	-0,196	-0,189	-0,176	-0,187	-0,171	-0,175	-0,182	-0,169	-0,171
Fdr_t^A	0,155	0,335	0,748	0,966	1,062	1,129	1,139	1,09	1,134	1,180
Fdr_t^Z	-0,197	0,025	0,578	0,815	0,844	0,886	0,907	0,889	0,915	0,913

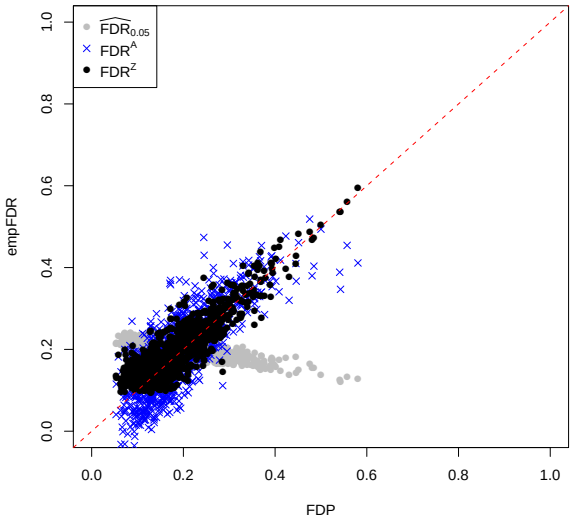
TAB. 3.2 – Coefficients de pente dans la régression entre différents estimateurs du FDR et la vraie proportion de faux-positifs (FDP) - $t = 0,05$

Les deux estimateurs conditionnels sont étudiés sur les données simulées de l'EXEMPLE 4 et comparés avec l'estimateur usuel. Afin d'éviter toute interférence, et n'étudier que l'impact de la dépendance sur l'estimation des taux d'erreurs, on suppose que m_0 est connu ($m_0 = 400$ pour l'EXEMPLE 4). La FIGURE 3.2.2 montre l'estimateur empirique et les estimateurs conditionnels du FDR en fonction de la vraie proportion de faux-positifs FDP_t avec $t = 0,05$, pour quatre scénarios de simulations caractérisés par différents niveaux de dépendance. Ces graphiques, ainsi que le TABLEAU 3.2, indiquent

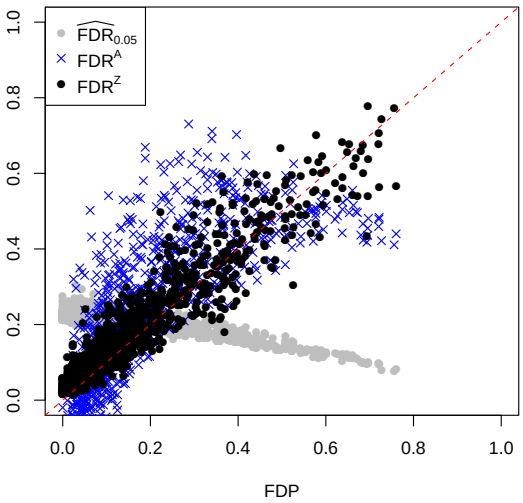
[1] L'interprétation de ce critère est à rapprocher de l'indicateur Δ introduit à la SECTION 2.2 pour caractériser l'écart à la loi $U[0; 1]$ de l'histogramme des probabilités critiques.



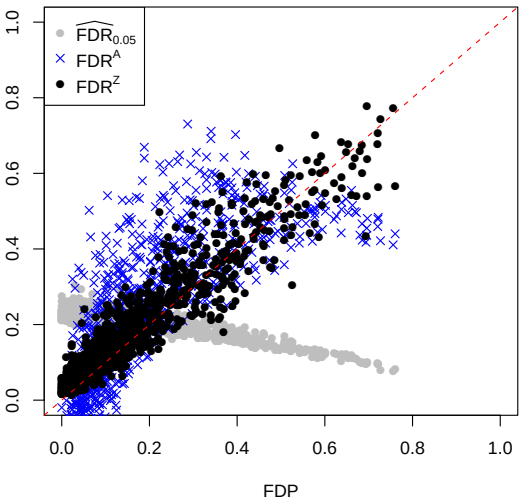
(a) scénario 1



(b) scénario 3



(c) scénario 6



(d) scénario 9

FIG. 3.11 – Estimateur empirique et estimateurs conditionnels du FDR en fonction de la vraie proportion de faux-positifs FDP_t avec $t = 0,05$, pour quatre scénarios de simulations caractérisés par différents niveaux de dépendance

que lorsque la structure de dépendance est marquée, les deux estimateurs conditionnels sont corrélés positivement à la vraie proportion de faux-positifs. L'estimateur FDZ_t^Z présente une dispersion moindre par rapport à l'estimateur FDR_t^A. Cette forte dispersion est également reportée dans les simulations de Efron [2007].

3.3. Analyse en Facteurs pour les Tests Multiples : FAMT

Nous présentons dans cette section la méthodologie d'Analyse en Facteurs pour les Tests Multiples, dénommée FAMT, du nom de package qui a été développé pour l'implémenter dans R (FAMT : Factor Analysis for Multiple Testing). Ce package est accessible sur le site du projet R (`http://cran.r-project.org/`) et un site internet lui est dédié (`http://famt.free.fr`). Un tutoriel est également disponible [Causeur et al., 2010].

Les propriétés conditionnelles étudiées dans ce chapitre permettent de définir un ensemble d'étapes constituant la procédure FAMT. Ces étapes sont décrites en détail ci-après. Il s'agit d'une procédure adaptative.

1. **Estimation de $\mathcal{M}_0$** : Comme la plupart des méthodes de tests multiples en grande dimension, nous avons besoin d'une première estimation de $\mathcal{M}_0$, l'ensemble des indices des variables pour lesquelles on ne rejette pas H_0^k. Efron [2007] propose de considérer les variables pour lesquelles le Z-score prend une valeur dans $[-1;1]$. On peut également utiliser les probabilités à postériori d'être sous H_0 pour chaque variable k : $\eta_0(p_k) = \frac{\pi_0 g_0(p_k)}{g(p_k)}$ [Efron et al., 2001] et choisir un seuil sur ces probabilités. Dans SVA, Leek and Storey [2007] considèrent les $m(1-\hat{\pi}_0)$ plus petites probabilités critiques.

   D'une façon plus générale, nous proposons de prendre $\widehat{\mathcal{M}_0} = \{k : p_k \geq 0,05\}$.

   Quelle que soit la méthode choisie, nous avons besoin soit de statistiques de test (pour le calcul des z-scores), soit de probabilités critiques. Pour cette première étape, l'estimation de $\mathcal{M}_0$ s'appuie sur les tests usuels.

2. **Choix du nombre de facteurs** : La SECTION 4.3 présente un certain nombre de méthodes habituellement mises en œuvre pour définir le nombre de facteurs à inclure dans le modèle d'Analyse en Facteurs. Dans le cadre des tests multiples, nous proposons une méthode qui détermine le nombre optimal de facteurs, permettant de minimiser la variance du nombre de faux-positifs (voir SECTION 4.3).

3. **Estimation des paramètres du modèle** : Connaissant le nombre de facteurs à inclure dans le modèle, il s'agit ensuite d'estimer les paramètres du modèle, à savoir les *loadings* (coefficients de la matrices B), les variances spécifiques (éléments diagonnaux de Ψ) et les scores (facteurs communs, Z). L'algorithme EMFA présenté à la SECTION 4.2.1 permet d'obtenir les estimateurs du Maximum de Vraisemblance de ces paramètres. Dans la plupart des implémentations de l'Analyse en Facteurs, il n'y a pas de covariables dans le modèle ($\forall k, m_k(x) = \mu_k$, modèle (4.2)). Dans notre contexte, le modèle s'écrit $Y_k = m_k(x) + b_k' Z + \varepsilon_k$ (4.2). Il faut donc centrer les variables par rapport à x, ce qui ne se fait pas de la même manière selon que $k \in \mathcal{M}_0$ ($Y_k - m_k^{(0)}(x)$) ou $k \notin \mathcal{M}_0$ ($Y_k - m_k(x)$). Cette étape de centrage préalable à la mise en œuvre de l'algorithme EMFA demande une attention particulière, et utilise l'estimation de $\mathcal{M}_0$ obtenue à l'ÉTAPE 1.

4. **Calcul des statistiques de test et des probabilités critiques** : Une fois le modèle ajusté, les statistiques de test ajustées par rapport aux facteurs (DÉFINITION 3.1.1) sont calculées. La

loi sous l'hypothèse nulle de ces statistiques de test est une loi de Student, ce qui permet de déduire les probabilités critiques (3.1.1).

5. **Mise à jour de $\widehat{\mathcal{M}}_0$** : Nous proposons de mettre à jour l'estimation de $\mathcal{M}_0$ (ÉTAPE 1) à partir des probabilités critiques ajustées : $\widehat{\mathcal{M}}_0 = \{k : \tilde{p}_k \geq 0,05\}$. A partir de cette nouvelle estimation de $\mathcal{M}_0$, les ÉTAPES 2 à 4 de la procédure sont reproduites.
6. **Estimation de π_0** : Les procédures déterminant le seuil sur les probabilités critiques pour le rejet des hypothèses nulles nécessitent pour la plupart l'estimation de la proportion d'hypothèses nulles π_0. Plusieurs approches ont été proposées à la SECTION 3.1.2 pour estimer ce paramètre en présence de dépendance, notamment à partir des probabilités critiques ajustées.
7. **Règles de décision** : Une méthode de seuillage est ensuite appliquée sur les probabilités critiques ajustées, permettant de contrôler le taux d'erreurs à un seuil α fixé à priori.

Conclusion

Ce chapitre présente plusieurs stratégies introduisant les facteurs communs de variabilité, définis au CHAPITRE 2, dans l'estimation des taux d'erreurs et de π_0

Les statistiques de test déduites de ces stratégies sont conditionnellement indépendantes. Ainsi, ce cadre étend donc les résultats obtenus en matière de contrôle de taux d'erreurs au cas de dépendance générale. En particulier, nous montrons que, en diminuant la corrélation entre les tests, les procédures d'inférence simultanées gagnent à la fois en puissance et en stabilité.

De la même manière, les répercussions de la dépendance sur l'estimation de π_0 sont également diminuées, puisque la variabilité de la distribution des probabilités critiques sous l'hypothèse nulle est fortement stabilisée.

Par ailleurs, cette modélisation de la dépendance par un ensemble de variables latentes permet de définir l'expression exacte de la variance du nombre de faux-positifs et de celle de π_0 au CHAPITRE 2. Nous nous appuyons sur ces propriétés pour définir des estimateurs conditionnels du FDR et de π_0, qui apparaissent comme des correction des estimateurs empiriques en présence de dépendance.

Enfin, nous présentons une procédure, appelée FAMT, qui décrit les différentes étapes pour la mise en œuvre de la méthodologie proposée dans ce chapitre. Cette procédure est illustrée au CHAPITRE 5 par deux études de cas. Au préalable, le CHAPITRE 4 s'intéresse à l'estimation des paramètres du modèle et aux stratégies de choix du nombre de facteurs à inclure dans le modèle.

CHAPITRE 4

ANALYSE EN FACTEURS EN GRANDE DIMENSION

Résumé Ce chapitre est dédié à la modélisation des données sous la forme du modèle d'Analyse en Facteurs introduit au CHAPITRE 2. Deux approches sont envisagées pour estimer les paramètres de ce modèle. Après les avoir décrites, nous justifions le choix de l'utilisation de l'estimation du modèle d'Analyse en Facteurs par maximum de vraisemblance, qui procure des propriétés intéressantes dans une optique d'utilisation du modèle dans le cadre des tests multiples. L'implémentation de cette méthode en grande dimension est décrite dans ce chapitre.

Sommaire

Introduction

La décomposition de la variabilité du modèle (2.1) s'apparente à celle d'un modèle d'Analyse en Facteurs [Mardia et al., 1979]. A l'origine, cette analyse a été développée dans le domaine de la psychométrie pour mesurer et définir un facteur dit "d'intelligence générale" : Spearman [1904] a été le premier à utiliser l'Analyse en Facteurs en psychologie et est parfois même considéré comme son créateur. Ses travaux ont abouti à la théorie selon laquelle les scores obtenus à des tests faisant appel à des processus cognitifs très disparates reflètent en fait une seule forme générale de performance intellectuelle ("g theory"). Ces recherches ont été poursuivies par R.B. Cattell [Cattell, 1966, 1978], à la fois dans le domaine de l'intelligence humaine et sur la théorie de l'Analyse en Facteurs. Connue des psychométriciens et des sociologues comme une technique de réduction de la dimension, l'Analyse en Facteurs est par ailleurs apparue récemment comme une technique d'analyse de la dépendance des données en grande dimension issues des expérimentations à haut-débit comme les biopuces par exemple [Kustra et al., 2006, Hsu and Chang, 2006, Pournara and Wernisch, 2007].

C'est une méthode d'analyse multivariée dont l'objectif est de décrire les relations de covariance, ou de corrélation, entre variables par l'intermédiaire d'un petit nombre de variables latentes appelées facteurs communs. Si on suppose, comme en Analyse en Facteurs dite exploratoire [Mardia et al., 1979] que $\mathbb{V}(Z) = \mathbb{I}_Q$, alors on obtient la décomposition suivante pour la matrice de variance-covariance des données $\mathbb{V}(Y) \equiv \Sigma$:

$$\Sigma = BB' + \Psi \tag{4.1}$$

où B représente la matrice de taille $m \times Q$ des coefficients b_k. BB' caractérise l'information commune à l'ensemble des variables et Ψ est une matrice diagonale caractérisant l'information spécifique à chacune d'elle.

Ici, le modèle (2.1) suppose trois hypothèses :

Hypothèse 2. *[Hypothèses sur le modèle d'Analyse en Facteurs]*

1. $\mathbb{E}(ZZ') = \mathbb{I}_Q$ *: les facteurs communs Z_q sont non corrélés et de variance unité. Cela favorise l'interprétation des facteurs, chacun portant une information non-redondante.*
2. $\mathbb{V}(E) = \Psi$*, diagonale : les facteurs spécifiques ε_k ne sont pas corrélés, et possèdent une variance spécifique ψ_k^2, k^{eme} élément diagonal de Ψ.*
3. $Cov(\varepsilon_k, Z_q) = 0, \forall k \in \mathcal{M}, \forall q \in [1; Q]$ *: le facteur spécifique ε_k représente l'information propre à la variable Y_k, non modélisée par les facteurs communs Z. Les facteurs communs et spécifiques ne sont donc pas corrélés.*

L'Analyse en Facteurs à été très largement étudiée dans la littérature au cours du XXème siècle, notamment dans les revues de psychométrie, de sciences du comportement et de marketing. Sa mise en œuvre passe par deux étapes principales :

1. le choix du nombre de facteurs à considérer dans le modèle (Q)
2. l'estimation des paramètres du modèle à Q facteurs : B, Ψ et Z

De nombreuses méthodes existent pour la réalisation des différentes étapes. Les nouveaux champs d'application de l'Analyse en Facteurs comme méthode de modélisation de la dépendance sont caractérisés par la grande dimension des données. L'analyse de ces données est un problème complexe, étant donné que dans la plupart des cas, l'augmentation du nombre d'expériences est impossible, pour des raisons de coûts par exemple. La prise en compte de la dimension particulière des données implique l'adaptation des méthodes usuelles ou le développement de nouvelles méthodes, satisfaisant les contraintes numériques et algorithmiques, tout en assurant une bonne précision des résultats obtenus.

Ce contexte introduit des conditions inhabituelles pour l'utilisation de l'Analyse en Facteurs et remets en cause les règles subjectives données par certains auteurs. En effet, la question du nombre d'individus minimum pour l'analyse est une question fréquemment posée [Preacher and MacCallum, 2002]. De nombreuses règles approximatives sont souvent avancées, comme un minimum de 100 à 250 individus, ou alors un ratio de 3 à 10 entre le nombre d'individus et celui de variables.

Ces conditions sont loin de ce qu'on observe dans le contexte de la grande dimension des données issues de technologies "haut-débit", où des milliers de variables sont mesurées sur un nombre modéré d'individus, de l'ordre de quelques dizaines.

Cependant, de nombreux auteurs s'accordent aussi pour dire que les règles basées sur le nombre minimum d'individus ou sur un ratio n/m ne sont pas celles qu'il faut considérer. Il faut en effet tenir compte de la structure des données qu'on veut analyser [MacCallum et al., 2008] : plus la communalité augmente, c'est-à-dire lorsque la structure commune est importante, moins le nombre d'individus nécessaire est important. De même, ce nombre va diminuer si le nombre de variables contribuant à la construction d'un facteur est important.

D'autre part, les avantages apportés par une décomposition de la matrice de variance-covariance Σ comme en (2.3) sont nombreux. Le nombre de paramètres inconnus à estimer dans les matrices B, Z et Ψ est beaucoup plus faible que ceux de la matrice de variance-covariance empirique S. En analyse de données génomiques par exemple, S est composé d'un nombre de paramètres inconnus de l'ordre de m^2, alors que la décomposition (2.3) implique seulement l'estimation de $m \times (Q+1)$. De plus, (2.3) offre la possibilité d'inverser facilement la matrice de variance-covariance[1], ce qui peut s'avérer utile pour des analyses ultérieures, en particulier celles impliquant l'analyse de la matrice des corrélations partielles.

L'Analyse en Facteurs est une méthode statistique éprouvée et de nombreux algorithmes sont disponibles pour sa mise en œuvre. On note essentiellement deux méthodes d'estimation, basées sur une

[1] $(A+BC)^{-1} = A^{-1} - A^{-1}B(\mathbb{I}_Q + CA^{-1}B)^{-1}CA^{-1}$, avec $A = \Psi$ et $C = B'$. Il s'agit donc d'inverser uniquement une matrice de taille $Q \times Q$ et une matrice diagonale.

approche factorielle d'une part et par Maximum de Vraisemblance d'autre part. Tout d'abord, nous présentons ces deux approches d'estimation des paramètres du modèle et justifions notre choix pour l'estimation par Maximum de Vraisemblance. Dans la deuxième partie de ce chapitre, nous étudions l'implémentation de cette méthode en grande dimension.

L'estimation des paramètres du modèle nécessite auparavant de connaitre le nombre de facteurs Q à inclure dans le modèle. Cette question est abordée dans la troisième section du chapitre. Nous supposons donc tout d'abord que Q est connu.

Notons que les différentes méthodes identifient les facteurs Z du modèle à partir des données centrées par rapport à la covariable d'intérêt x :

$$Y_k^* \quad = \quad Y_k - m_k(x) = b_k' Z + \varepsilon_k \tag{4.2}$$

4.1. Estimation du modèle par Analyse en Facteurs

Nous nous intéressons ici à l'extraction des scores Z, les facteurs communs, et à l'estimation des paramètres de variances B et Ψ. Plusieurs méthodes existent : Moindres Carrés Ordinaires ou Généralisés, méthodes factorielles (SVD, Principal Factoring, Alpha Factoring), Maximum de Vraisemblance. Nous présentons ci-après celles qui sont généralement implémentées dans les principaux logiciels et utilisées dans la plupart des études statistiques.

4.1.1 Méthode factorielle

Estimation du modèle par SVD (Décomposition en Valeurs Singulières) La Décomposition en Valeurs Singulières (SVD) permet de factoriser la matrice Y^*, en un produit de matrices simples possédant des caractéristiques interprétables. Si r désigne le rang de la matrice Y^*, on écrit :

$$Y^* = UDV' \tag{4.3}$$

où $U_{n\times r}$ est la matrice des vecteurs propres de $Y^*Y'^*$, $V_{m\times r}$ est la matrice des vecteurs propres de $Y' * Y*$ et $D_{r\times r}$ est une matrice diagonale dont les éléments diagonaux sont les racines carrées des valeurs propres, en général rangées par ordre décroissant. Par identification avec le modèle (2.1), on peut estimer Z par $\frac{1}{n}U_Q$ et B par $V_Q D_Q$, où U_Q, V_Q et D_Q représentent les Q premières colonnes des matrices U, V et D. E correspond alors aux résidus de cette approximation.

La matrice de variance-covariance se décompose alors de la façon suivant :

$$\Sigma = B\mathbb{V}(Z)B' + \mathbb{V}(E) = \frac{1}{n}V_Q D_Q^2 V_Q' + \frac{1}{n}E'E \tag{4.4}$$

car $U_Q' U_Q = \mathbb{I}$.

C'est sur ce principe qu'est basée la méthode SVA (pour Surrogate Variable Analysis), développée par J. Storey et J. Leek [Leek and Storey, 2007, 2008] initialement pour l'analyse de données d'expressions géniques, et programmée dans R dans le package `sva` [Leek, 2008]. Le principe de cette méthode et l'algorithme proposé par ses auteurs pour estimer les paramètres du modèle sont détaillés à l'ANNEXE C.

Principal Factoring (PF) Cette méthode itérative se base sur la décomposition en valeurs propres de la matrice de corrélation R. Les *loadings* B sont estimées à partir des Q premiers vecteurs propres. On en déduit les variances spécifiques puis cette procédure est répétée, jusqu'à ce qu'un certain critère d'arrêt ait atteint une limite définie à priori.

On suppose dans un premier temps que Ψ est connue, ou du moins qu'on en possède une estimation initiale notée Ψ_0. Certains logiciels proposent une valeur calculée automatiquement par régression multiple de chaque variable sur toutes les autres. En effet, le coefficient de corrélation multiple r_i^2 de la i^{eme} régression correspond à la proportion de variance associée à la i^{eme} variable, donc ce coefficient peut servir à estimer la i^{eme} communalité. Une estimation initiale de Ψ_0 est alors $diag\left(1-r_i^2\right)$. D'autres méthodes sont utilisées : choix d'une valeur par défaut (0 ou 0,5 par exemple). Le principe de l'algorithme est le suivant :

1. **Initialisation** de Ψ : $\hat{\Psi}_0$
2. **Itération**
   - Analyse en Composante Principale de $R-\hat{\Psi}_0$: décomposition en valeurs propres et vecteurs propres de $R-\hat{\Psi}_0$
   - Estimation des loadings B : $\hat{B}$ correspond aux Q premières composantes calculées à l'étape précédente
   - Estimation de la variance spécifique : $\hat{\Psi}$ correspond à la diagonale de $R^* = R - \hat{B}\hat{B}'$.
3. **Critère d'arrêt** : on itère l'étape 2, en considérant la nouvelle estimation de Ψ pour $\hat{\Psi}_0$, jusqu'à ce que deux estimations successives $\hat{\Psi}$ et $\hat{B}$ soient identiques, à un seuil ϵ près. Le critère considéré peut être un écart quadratique entre les matrices R et $\hat{B}\hat{B}'+\hat{\Psi}$ (par exemple : $tr[(R-(\hat{B}\hat{B}'+\hat{\Psi}))^2]$)

On peut définir des indices d'adéquation pour le modèle obtenu. Par exemple, on peut comparer la somme des carrés résiduels de la matrice estimée par rapport à la matrice d'origine.

$$GF_Q \quad = \quad 1 - \frac{\mathbb{1}\ R^{*2}\mathbb{1}'}{\mathbb{1}R^2\mathbb{1}'}$$

On peut également construire un test, en considérant la somme des carrés des résidus $r*_{ij}^2$, éléments de R^{*2}. Cette somme sera négligeable si le modèle est bien ajusté, ce qui conduira à la non-significativité du test [Chavent et al., 2007] :

$$(n-1)\sum_{i<j} r*_{ij}^2 \sim \chi^2 \frac{m(m-1)}{2}$$

Cette méthode peut servir à donner une estimation initiale de B et Ψ pour la méthode du Maximum de Vraisemblance présentée ci-après. L'inconvénient majeur de cette méthode est numérique car elle nécessite le stockage et l'inversion de matrice de (très) grande dimension ($m \times m$), notamment lors de la décomposition SVD de la matrice de variance-covariance dans l'ACP à l'étape 2.

4.1.2 Estimation par Maximum de Vraisemblance

Cette méthode se base sur le calcul du Maximum de Vraisemblance dans le cas d'une distribution de Wishart (loi Normale multivariée). On suppose donc que les données ont pour distribution :

$$Y \sim \mathcal{N}_m(\mu, \Sigma)$$

Cela a pour conséquence de supposer que Z et E dans (2.1) sont distribués selon des loi normales Q- et m-multivariées, respectivement. On rappelle dans un premier temps, dans le cadre de la multinormalité, la fonction de vraisemblance L en fonction de B et Ψ (les données Y sont supposées centrées) :

$$\begin{aligned} L(B,\Psi) &= (2\pi)^{-\frac{n.m}{2}} |\Sigma|^{-\frac{n}{2}} exp\left(-\frac{1}{2}\sum_{i=1}^{n} Y_i'\Sigma^{-1}Y_i\right) \\ &= (2\pi)^{-\frac{n.m}{2}} |\Sigma|^{-\frac{n}{2}} exp\left(-\frac{1}{2}tr(\Sigma^{-1}S)\right) \end{aligned}$$

où $\Sigma = BB' + \Psi$.

Les estimateurs du Maximum de Vraisemblance de B et Ψ, notés $\hat{B}$ et $\hat{\Psi}$, maximisent L. Il s'agit de trouver une solution permettant de résoudre :

$$\frac{\partial L}{\partial B} = 0 \qquad \text{et} \qquad \frac{\partial L}{\partial \Psi} = 0 \tag{4.5}$$

Il n'existe pas de solution analytique à (4.5). La résolution de cette équation se fait de manière itérative. Des algorithmes de type Newton-Raphson sont en général implémentés. Cependant, ces méthodes convergent souvent lentement, et mènent parfois à des solutions aberrantes dite *Heywood cases* (valeurs négatives pour les termes de la diagonale de $\hat{\Psi}$ par exemple). Connaissant une estimation de B et Ψ, nous pouvons alors trouver celle de Z par la méthode de Thomson [1951]. Le principe de cette méthode est de réaliser une régression des scores Z (inconnus) par rapport aux données Y (observées). On rappelle que :

$$\begin{bmatrix} y \\ z \end{bmatrix} = \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix} ; \begin{bmatrix} \Sigma & B \\ B & \mathbb{I}_Q \end{bmatrix} \right)$$

D'où : $\mathbb{E}(z|y) = y\Sigma^{-1}B$ et donc

$$\hat{z} = y[\hat{B}\hat{B}' + \hat{\Psi}]^{-1}\hat{B}$$

En forme matricielle, on a l'estimation des scores suivantes :

$$\hat{Z} = Y[\hat{B}\hat{B}' + \hat{\Psi}]^{-1}\hat{B} = Y\hat{\Psi}^{-1}\hat{B}[\mathbb{I}_Q + \hat{B}'\hat{\Psi}^{-1}\hat{B}]^{-1}$$

4.1.3 Choix de la méthode

L'avantage de l'approche factorielle est qu'elle ne suppose pas de distribution particulière pour les scores. En effet, ceux-ci sont considérés comme des effets fixes dans le modèle (2.1). Néanmoins, la décomposition en valeurs singulières présente un inconvénient. L'hypothèse d'indépendance des résidus $\hat{E}$ (HYPOTHÈSE 2) n'est pas vérifiée : estimer B et Z par SVD ne permet pas de garantir que $\hat{E}\hat{E}'$ est diagonale. En pratique, cette méthode est pourtant fréquemment utilisée, et programmée dans les principaux logiciels de statistique. Les résultats obtenus pour B et Z sont donnés comme satisfaisants, à condition que les termes extradiagonnaux de $\hat{E}\hat{E}'$ soient négligeables, ce qui passe en particulier par une estimation correcte de Q. La méthode PF nécessite la manipulation et le stockage de la matrice de variance-covariance empirique, ce qui peut être un obstacle numérique à sa mise en œuvre.

L'autre approche présentée consiste à chercher les estimateurs du Maximum de Vraisemblance des paramètres, sous la contrainte que Ψ est diagonale. On suppose alors que les données suivent une loi normale multivariée, de matrice de variance covariance $\Sigma = BB' + \Psi$. Les facteurs communs obtenus ici ne sont pas forcément ordonnés en fonction de leur inertie et le modèle à Q_1 facteurs n'est pas imbriqué dans le modèle à Q_2 facteurs ($Q_1 < Q_2$) comme c'est le cas en ACP.

Cette méthode d'estimation du modèle (2.1) permet de vérifier les HYPOTHÈSES 2 du modèle d'Analyse en Facteurs, et en particulier d'assurer l'indépendance des résidus $\hat{E}$ par construction.

Dans le cadre d'inférence définit au CHAPITRE 3, nous donnons une importance particulière à la variabilité spécifique dans la définition de la statistique de test (DÉFINITION 3.1.1). L'analyse des principales différences entre l'estimation du modèle (2.1) par ACP et par Analyse en Facteurs penche alors en faveur de la deuxième méthode.

Néanmoins, l'estimation des paramètres de ce modèle par Maximum de Vraisemblance doit être adaptée au contexte de la grande dimension : la convergence des algorithmes usuels n'est pas assurée et leurs propriétés asymptotiques mal définies. Cette étude fait l'objet de la section suivante.

4.2. Analyse en Facteurs en grande dimension

4.2.1 Algorithme EMFA

Parmi les méthodes d'estimation du modèle, la méthode du Maximum de Vraisemblance a des propriétés asymptotiques intéressantes, notamment en terme d'efficacité, d'invariance d'échelle, de possibilité de test du modèle. Elle satisfait de plus à l'objectif inférentiel de notre démarche. Cette méthode s'appuie néanmoins sur la normalité multivariée des données.

Robertson and Symons [2007] précisent les conditions d'utilisation de l'estimation par Maximum de Vraisemblance des paramètres de l'Analyse en Facteurs dans le cas où la matrice de variance-covariance n'est pas de plein rang. Leur étude montre que le nombre de facteurs Q ne doit pas être

trop important par rapport au nombre de variables. En particulier, leurs travaux identifient les cas où la vraisemblance est bornée, et peut donc être maximisée.

On suppose ici que les données suivent une loi normale multivariée, de matrice de variance covariance Σ. Lawley [1940] a présenté la méthode d'estimation par maximisation directe de la vraisemblance lorsque la matrice est de plein rang. De nombreux algorithme existent pour obtenir les estimateurs du Maximum de Vraisemblance en Analyse en Facteurs (par exemple [Jöreskog, 1977]). Néanmoins, pour éviter, entre autre, la génération de *Heywood cases* (variances spécifiques négatives) lors de la recherche des EMV par un algorithme de type Newton-Raphson, nous nous tournons vers un algorithme de type Expectation-Maximisation (EM) [Dempster et al., 1977]. Il s'agit d'une classe d'algorithmes qui permettent de trouver les Estimateurs du Maximum de Vraisemblance (EMV) des paramètres de modèles probabilistes lorsque le modèle dépend de variables latentes non observables. On alterne des étapes d'évaluation de l'espérance (E), où on calcule l'espérance de la vraisemblance en tenant compte des variables observées, et une étape de maximisation (M), où on estime les EMV des paramètres en maximisant la vraisemblance trouvée à l'étape (E). On utilise ensuite les paramètres trouvés en (M) comme point de départ d'une nouvelle phase d'évaluation de l'espérance, et on itère ainsi. Son implémentation est assez simple, et c'est un algorithme stable en terme de convergence (augmentation de la vraisemblance à chaque itération).

Cet algorithme a été implémenté dans le cadre de l'Analyse en Facteurs [Rubin and Thayer, 1982]. Il est utile lorsqu'on travaille avec des données en (très) grande dimension ($m \gg n$). En effet, la matrice de variance-covariance ($m \times m$) n'est pas calculée explicitement. Cela représente un atout numérique pour cette méthode (pas de stockage ni d'inversion d'une matrice aux dimensions très élevées), qui nécessite seulement l'inversion de matrices de taille $Q \times Q$, $Q << m$.

Remarque : De nombreuses extensions de l'algorithme EM ont été développées, notamment lorsque des formes exactes de la vraisemblance ne peuvent être obtenues, ou pour améliorer les performances de convergence, par exemple [Liu and Rubin, 1998, McLachlan et al., 2004]. Ici, on s'intéresse à la précision de l'estimation des paramètres, et les problématiques de vitesse de convergence, bien qu'essentielles, ne sont pas abordées.

Mise en œuvre de l'algorithme EMFA Dans un premier temps, comme la plupart des algorithme itératifs, l'algorithme EM considère des valeurs initiales pour B et Ψ. Plusieurs méthodes ont été proposées pour l'étape d'initialisation. On peut par exemple réaliser la décomposition en valeurs singulières de la matrice des données (centrées) Y^* (ou plus exactement de sa transposée puisque $m >> n$) et on en déduit les Q premières valeurs propres et vecteurs propres. Il s'agit d'une étape de l'estimation des paramètres par PF. Les *loadings* B sont estimés par les Q premiers vecteurs propres, multipliés par la racine des Q premières valeurs propres, et l'estimation de la variance spécifique correspond à la diagonale de $S - \hat{B}\hat{B}'$.

On note $Y^{(i)}$ la i^{eme} ligne de Y, associée à l'observation i, vecteur de taille $1 \times m$, et $Z^{(i)}$ le vecteur de taille $1 \times Q$ des scores associés à l'observation i. L'algorithme EM s'appuie sur la (log-)vraisemblance

du modèle (2.1) et a pour but de déterminer les paramètres qui la maximisent. L'ANNEXE B fournit le détail de la construction des équations de mise-à-jour de l'algorithme.

Le principe de l'algorithme EMFA est le suivant :

1. On part des données centrées $Y = Y^*$
2. **Initialisation** : On dispose d'une première estimation pour B et Ψ (B_0 et Ψ_0)
3. **Itération** : On alterne des étapes d'estimation de l'espérance de la vraisemblance et des étapes de maximisation, à partir des valeurs B_0 et Ψ_0 pour trouver les nouvelles estimations B_1 et Ψ_1 :
   (a) **étape E** : On estime l'espérance de la (log-)vraisemblance du modèle.

$$\begin{aligned}\mathbb{E}(Z^{(i)}|Y^{(i)}) &= Y^{(i)}\Psi_0^{-1}B_0(\mathbb{I}_Q + B_0'\Psi_0^{-1}B_0)^{-1} \\ \mathbb{E}(Z^{(i)}Z'^{(i)}|Y^{(i)}) &= (\mathbb{I}_Q + B_0'\Psi_0^{-1}B_0)^{-1} + \mathbb{E}(Z^{(i)}|Y^{(i)})'\mathbb{E}(Z^{(i)}|Y^{(i)})\end{aligned}$$

   D'après (B.8) et (B.9).

   (b) **étape M** : On définit alors les nouvelles estimations des paramètres (EMV).

$$\begin{aligned}B_1 &= \sum_{i=1}^{n}\left[Y'^{(i)}\mathbb{E}(Z^{(i)}|Y^{(i)})\right]\left[\sum_{i=1}^{n}\mathbb{E}(Z'^{(i)}Z^{(i)}|Y^{(i)})\right]^{-1} \\ \Psi_1 &= diag\left[S - \frac{1}{n}\sum_{i=1}^{n}Y'^{(i)}\mathbb{E}(Z^{(i)}|Y^{(i)})B_1'\right]\end{aligned}$$

   D'après les équations de mise-à-jour définies en (B.3) et (B.4).
4. **Critère d'arrêt** : on itère les étapes 3a et 3b en prenant les estimations de B_1 et Ψ_1 obtenues avec une étape E et une étape M comme point de départ de l'itération suivante. On itère jusqu'à ce que deux estimations successives de Ψ et B soient identiques, à un seuil ϵ près. Le critère considéré peut être la trace de l'écart quadratique entre les matrices Ψ, obtenues lors de 2 étapes successives : $\frac{1}{m}tr\left(\Psi_1 - \Psi_0\right)^2 < \epsilon$.

4.2.2 Rotations

Le modèle d'Analyse en Facteurs n'est identifiable qu'à une rotation près de la matrice des coefficients B. En effet, si on écrit $\Sigma = BB' + \Psi$ et qu'on multiplie B par une matrice orthogonale T tel que $M = BT$, on a alors $MM' = BTT'B' = BB'$, car $TT' = \mathbb{I}$ par définition. Ainsi, le choix judicieux d'une rotation peut définir des facteurs qui ajustent le modèle avec la même précision et facilitent l'interprétation des résultats.

Une façon arbitraire de définir une matrice des *loadings* unique est d'ajouter une contrainte : il s'agit par exemple de prendre la matrice orthogonale T telle que $B'\Psi^{-1}B$ soit diagonale [McLachlan et al., 2003]. Cela ajoute alors $\frac{1}{2}Q(Q-1)$ contraintes pour les paramètres.

Les rotations les plus courantes ont pour but de minimiser certains *loadings* et d'en maximiser d'autres pour mieux mettre en évidence une structure hétérogène des variables les plus corrélées à chacun des facteurs communs Z_q.

La rotation "Varimax" [Kaiser, 1958] est la plus utilisée et permet d'obtenir cette configuration pour les *loadings* : elle minimise la confusion entre les facteurs, tout en conservant l'orthogonalité des vecteurs propres. La rotation "Varimax" rend ainsi l'interprétation plus aisée en maximisant la variance du carré des coordonnées des variables par colonne (*loadings*). Il s'agit de maximiser le critère suivant pour déterminer les nouvelles coordonnées b_{kqj} de la variable k pour le facteur q :

$$R_v = \sum_{q=1}^{Q} \sum_{k=1}^{m} \left(b_{kq}^2 - \frac{1}{m} \sum_{r=1}^{m} \beta_{rq}^2 \right)^2$$

On vérifie les propriétés du modèle d'Analyse en Facteurs après la rotation T ($Q \times Q-$matrice orthogonale). On a : $Y = ZB' + E = ZTT'B' + E = GA' + E$. G représente la matrice des scores et A la matrice des *loadings* après la rotation T.

- $\mathbb{V}(G) = 1/nG'G = 1/nT'Z'ZT = I_Q$
- $\mathbb{V}(E) = \Psi$ car E n'est pas affecté par la rotation
- $\text{Cov}(E; G) = \text{Cov}(E; ZT) = 0$ car $\text{Cov}(E; Z) = 0$

4.2.3 Degrés de libertés

Par analogie avec les méthodes de lissage, le prédicteur du modèle (2.1) s'écrit aussi sous la forme $\hat{Y} = PY$. Pour tenir compte de la compléxité du modèle, de nombreux auterus suggèrent d'associer à l'ajustement un nombre de degrés de liberté égal à $ddl = tr(2P - PP')$.

Ces définitions des degrés de liberté dans le cadre du modèle linéaire se transposent au cas des modèles non paramétriques, c'est-à-dire lorsque P n'est plus un projecteur mais dépend de la distribution des données Y : $P = P(Y)$ [Hastie and Tibshirani, 1990].

Dans le cas de l'Analyse en Facteurs, on considère le modèle sous forme matricielle $Y = ZB' + E$. Les prédictions du modèle sont $\hat{Y} = ZB'$. On note $G = [\mathbb{I}_Q + B'\Psi^{-1}B]^{-1}$. On a par ailleurs : $B = Y'Z[\mathbb{E}(Z'Z)]^{-1}$ et $\mathbb{E}(Z'Z) = nG + Z'Z$ (d'après (B.3) et(B.9)), d'où :

$$\hat{Y} \quad = \quad ZB' = \left(Z[nG + Z'Z]^{-1}Z'\right) Y = HY \tag{4.6}$$

Ainsi, .dans le cadre de l'Analyse en Facteurs, les degrés de libertés des résidus seront :

$$ddl_r \quad = \quad n - tr(2H - HH') \tag{4.7}$$

4.2.4 Validation de l'estimation des paramètres en grande dimension

Les propriétés des estimateurs issus de l'algorithme EM sont étudiées en détail dans de nombreux articles et ouvrages, dans le cadre classique pour la dimension des données (n<m). On s'intéresse à la conservation de ces propriétés dans le contexte de la grande dimension.

Nous nous appuyons sur l'EXEMPLE 6 pour étudier dans un premier temps l'estimation de B puis dans un second temps l'estimation des variances spécifiques, à la fois en situation asymptotique et dans le cas de petits échantillons.

EXEMPLE 6. On considère 10 scénarios, caractérisés par le niveau de dépendance des données simulées, de faible (quasi-indépendance) à élevé. Il s'agit ici du même protocole de simulation que lors de l'EXEMPLE 3.

Chaque jeu de données est composé de $m = 500$ variables et de $n = 10, 50, 500$ observations. Le nombre de facteurs est connu, et le même pour tous les jeux de données. On considère différentes situations ($Q = 2; 5; 10$).

Dans chaque configuration (niveau de dépendance / taille d'échantillon / nombre de facteurs) 1000 jeux de données sont simulés. L'algorithme EMFA est appliqué à chacun des jeux de données pour obtenir les estimations des paramètres.

L'EXEMPLE 6 permet d'étudier l'estimation des *loadings* et des variances spécifiques en situation de petits échantillons et en situation asymptotique. Nous présentons les résultats pour $Q = 5$. Les résultats sont similaires pour les autres valeurs de Q.

Estimation de B La relation entre deux matrices $X_{n\times m_x}$ et $Y_{n\times m_y}$ peut être évaluée à travers le coefficient RV [Escouffier, 1973]. Ce coefficient prend des valeurs entre 0 et 1, une valeur nulle indiquant l'absence de liaison entre les deux matrices, et une valeur de 1 signifiant que les deux matrices sont des représentations d'une même configuration. On suppose X et Y centrées. Le coefficient RV est définit par :

$$RV(X,Y) = \frac{< XX'; YY' >}{\|XX'\| \, \|YY'\|} = \frac{tr(XX'YY')}{\sqrt{tr\left((XX')^2\right)) \, tr\left((YY')^2\right)}} \tag{4.8}$$

scénario	1	2	3	4	5	6	7	8	9
var. commune (%)	4,65	15,56	28,56	40,31	50,76	57,90	65,45	70,09	75,19
$n = 10$									
min	0,01	0,08	0,10	0,21	0,28	0,43	0,42	0,52	0,43
moyenne	**0,11**	**0,32**	**0,47**	**0,56**	**0,67**	**0,70**	**0,9774**	**0,78**	**0,81**
max	0,37	0,59	0,71	0,77	0,85	0,84	0,89	0,90	0,90
écart-type	**0,080**	**0,096**	**0,103**	**0,097**	**0,079**	**0,071**	**0,071**	**0,069**	**0,069**
$n = 50$									
min	0,11	0,54	0,71	0,75	0,83	0,85	0,85	0,88	0,89
moyenne	**0,48**	**0,72**	**0,85**	**0,89**	**0,93**	**0,94**	**0,95**	**0,96**	**0,96**
max	0,66	0,81	0,91	0,93	0,96	0,96	0,97	0,97	0,98
écart-type	**0,066**	**0,042**	**0,028**	**0,022**	**0,016**	**0,014**	**0,012**	**0,011**	**0,011**
$n = 500$									
min	0,833	0,924	0,975	0,982	0,988	0,984	0,989	0,991	0,991
moyenne	**0,863**	**0,936**	**0,984**	**0,988**	**0,992**	**0,993**	**0,995**	**0,995**	**0,996**
max	0,881	0,944	0,987	0,991	0,995	0,996	0,996	0,997	0,997
écart-type	**0,0068**	**0,0032**	**0,0016**	**0,0013**	**0,0010**	**0,0011**	**0,0009**	**0,0009**	**0,0008**

TAB. 4.1 – Statistiques descriptives des coefficients RV pour chaque scénario de dépendance

Il existe un test de significativité de ce coefficient [Josse et al., 2008] qui permet de valider l'estimation de B. En effet, cette matrice étant définie à une rotation près, nous allons vérifier si l'espace estimé

par les *loadings* obtenu par Maximum de Vraisemblance (algorithme EMFA) sur chacun des jeux de données simulés est le même que la matrice B théorique qui a servi pour les simulations à partir de l'EXEMPLE 6.

Les résultats sont donnés dans le TABLEAU 4.1 pour les différents scénarios, et pour les 3 tailles d'échantillons. Les résultats du tableau sont donnés pour $Q = 5$.

D'après le TABLEAU 4.1, plus la structure est forte, plus l'estimation de l'algorithme EMFA et la vraie matrice B sont homothétiques. Pour des structures plus faibles, les coefficients RV sont plus proches de 0. Ce phénomène est d'autant plus marqué que la taille d'échantillon est faible. Néanmoins, les tests de significativité des coefficients sont toujours positifs (probabilités critiques inférieures à 10^{-6}, voir de l'ordre de 10^{-100} pour les scénarios avec un niveau de dépendance élevé et/ou une taille d'échantillon importante ; probabilités critiques inférieures à 10^{-2} pour le scénario 1, quel que soit n, sauf pour $n = 10 : 45$ tests négatifs au seuil 10^{-2}).

Estimation de Ψ Dans le cas de petits échantillons, il existe un biais dans l'estimation de Ψ, qui peut être non négligeable. Il apparait particulièrement important de corriger ce biais, étant donné l'utilisation de la matrice de variances spécifiques (voir CHAPITRE 3) dans le cadre des tests multiples.

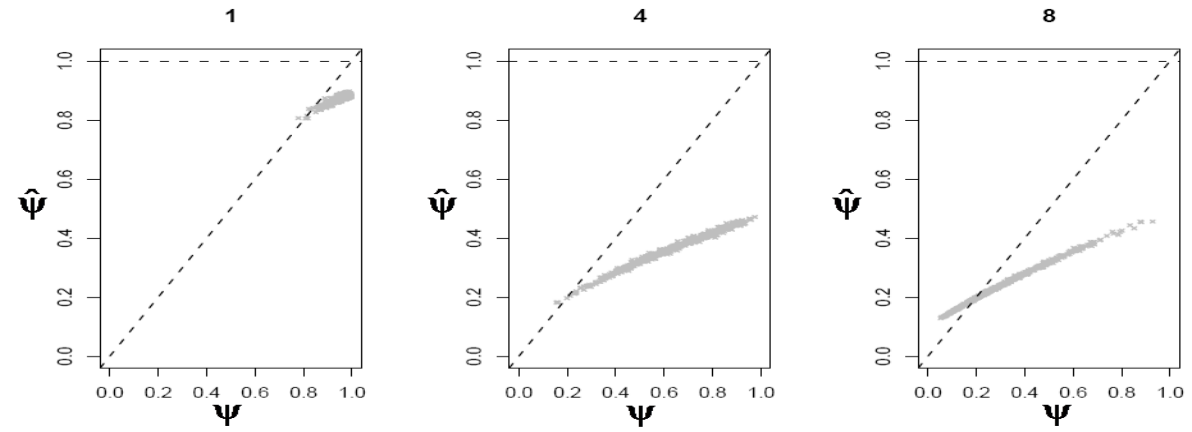
Les FIGURES 4.1(a), 4.1(b) et 4.1(c) montrent l'estimation de Ψ par rapport à la vraie matrice de variances spécifiques qui a servi à simuler le jeu de données, pour trois des scénarios (variance commune : niveau 1 : 4%, niveau 4 : 38%, niveau 8 : 68%), et pour les trois tailles d'échantillons ($n = 10$, $n = 50$ et $n = 500$).

Pour des petits échantillons, les variances spécifiques proches de 1 ont tendance à être sous-estimées. Ce phénomène est d'autant plus marqué que la dépendance entre les données est importante. A mesure que la taille d'échantillon augmente, ce biais disparaît, confirmant ainsi les propriétés asymptotiques des EMV.

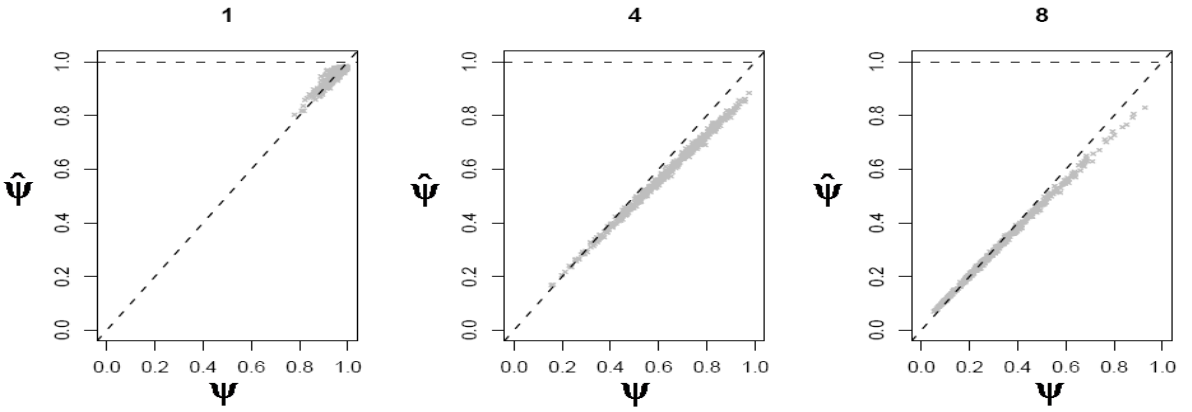
Nous proposons de corriger le biais observé dans l'estimation de Ψ à l'aide du nombre de degrés de libertés des résidus ddl_r du modèle (2.1), définit en (4.7) :

$$\hat{\Psi}_c = \frac{n}{ddl_r}$$

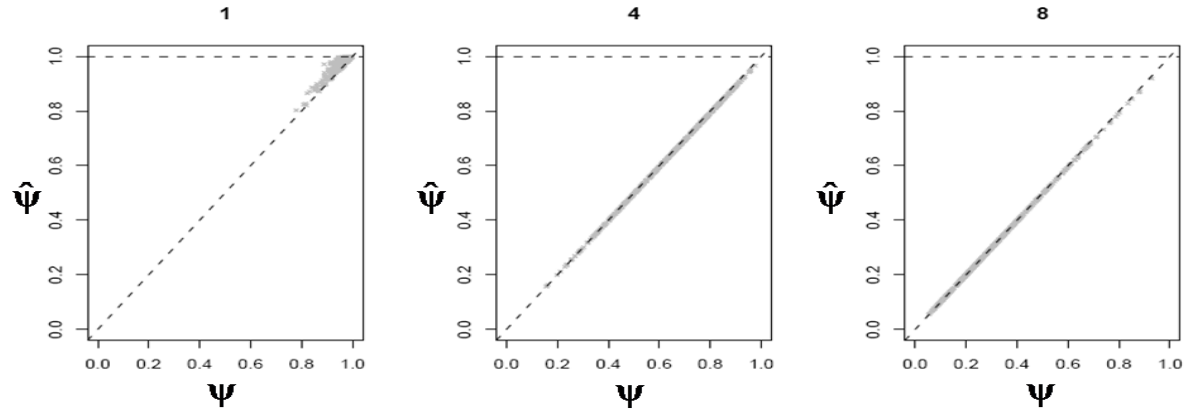
La FIGURE 4.2 montre les résultats de cette estimation corrigée sur les mêmes données que précédemment.



(a) $n = 10$

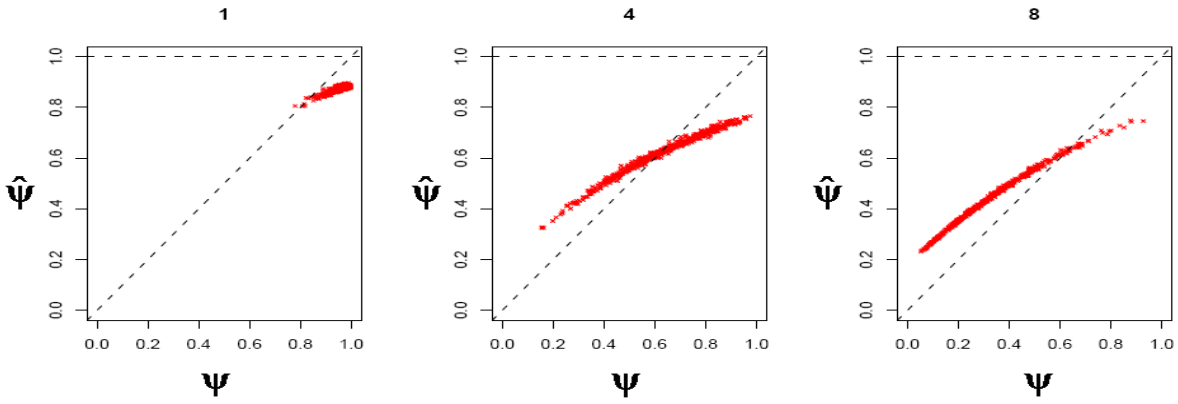


(b) $n = 50$

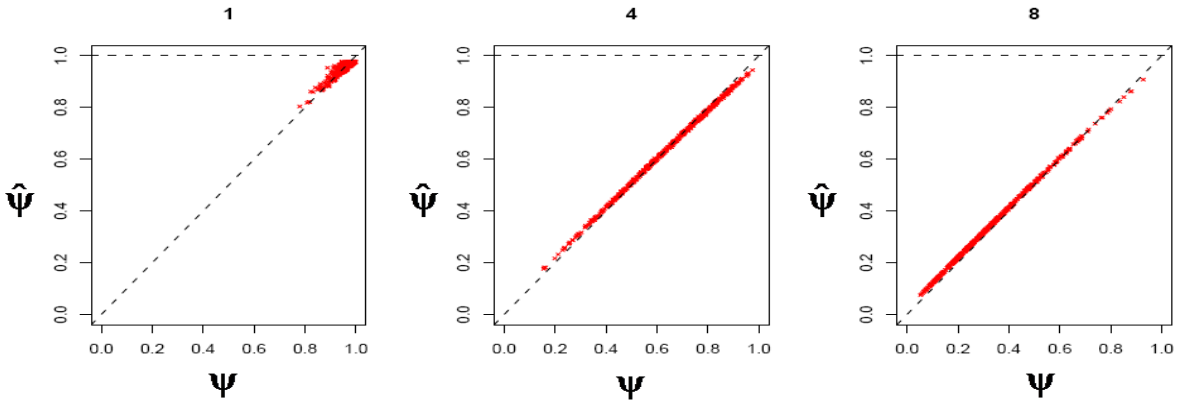


(c) $n = 500$

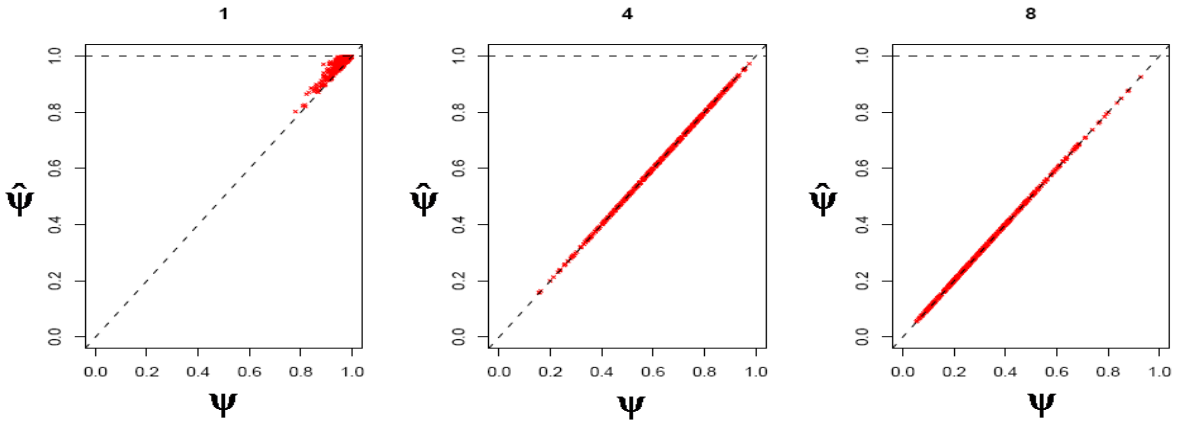
FIG. 4.1 – Estimation de Ψ par rapport à la valeur théorique ayant servie pour les simulations, pour 3 scénarios de dépendance : faible (1), intermédiaire (4) et élevé (8) et une taille d'échantillon $n = 10, 50, 500$ - 1000 jeux de données sont simulés pour chaque scénario - le graphique représente les moyennes des estimations obtenues sur les 1000 jeux de données



(a) $n = 10$



(b) $n = 50$



(c) $n = 500$

FIG. 4.2 – Estimation corrigée de Ψ par rapport à la valeur théorique ayant servie pour les simulations, pour 3 scénarios de dépendance : faible (1), intermédiaire (4) et élevé (8) et une taille d'échantillon $n = 10, 50, 500$ - 1000 jeux de données sont simulés pour chaque scénario - le graphique représente les moyennes des estimations obtenues sur les 1000 jeux de données

4.3. Choix du nombre de facteurs

Les données ont parfois des structures de dépendance complexes, et le nombre Q de facteurs communs est inconnu en pratique. La première étape de l'Analyse en Facteurs consiste à choisir le nombre approprié de facteurs qui rend compte des relations entre les variables du jeu de données.

Il est nécessaire de faire preuve de parcimonie afin de n'avoir à estimer que peu de paramètres dans le modèle mais également de garantir un bon ajustement de la matrice de covariance. Sous-estimer Q conduit à une perte d'information, le modèle ignorant un facteur commun, ou recombinant plusieurs facteurs en un seul. Les *loadings* sont alors mal estimés, et leur interprétation erronée. Sur-estimer Q est en général considéré comme moins grave. Cependant, considérer un trop grand nombre de facteurs augmente le nombre de paramètres à estimer d'une part, et met en avant des facteurs avec seulement quelques *loadings* élevés.

Le choix du nombre de facteurs joue un rôle important à la fois du point de vue de l'estimation des paramètres et de l'interprétation au niveau des données. Pour toutes ces raisons, cette question est très fréquemment étudiée dans les revues dédiées, débouchant sur des règles plus ou moins subjectives.

Nombre maximum de facteurs dans le modèle Le nombre de termes de la matrice de variance-covariance Σ de m variables est de $\frac{1}{2}m(m+1)$. Le nombre de *loadings* (éléments de B) est $m.Q$, et le nombre de valeurs de variances spécifique est m. Il s'agit donc d'identifier $m(Q+1)$ éléments. Par ailleurs, le nombre de contraintes du modèle est de $\frac{1}{2}Q(Q-1)$, en supposant $B'\Psi^{-1}B$ diagonale afin d'assuer l'unicité de la solution. En prenant en compte ces contraintes, il s'agit d'estimer $m(Q+1) - \frac{1}{2}Q(Q-1)$ éléments, à partir de $\frac{1}{2}m(m+1)$ valeurs. On peut donc en déduire une relation pour le nombre maximum de facteurs identifiable en fonction du nombre de variables :

$$\begin{aligned} \frac{1}{2}m(m+1) - [m(Q+1) - \frac{1}{2}Q(Q-1)] &\geq 0 \\ &\Updownarrow \\ (m-Q)^2 &\geq m+Q \\ &\Updownarrow \\ Q &\leq \frac{2m+1-\sqrt{8m+1}}{2} \end{aligned}$$

Méthodes existantes Les critères les plus connus sont basés sur les valeurs propres de la matrice de variance-covariance empirique (valeurs propres supérieures à 1 [Kaiser, 1960], critère du coude [Cattell, 1966]). La facilité de mise en œuvre de ces méthodes en ont fait des méthodes populaires, d'après de nombreuses études sur les méthodes de choix du nombre de facteurs utilisées dans les domaines de la psychologie ou du marketing [Stewart, 1981, Ford et al., 1986, Fabrigar et al., 1999, Norris and Lecavalier, 2009]. L'analyse parallèle [Montanelli and Humphrey, 1976] est souvent décrite comme performante pour déterminer le nombre de facteurs à conserver dans un modèle [Lance et al., 2006]. Elle est basée sur du rééchantillonnage. La méthode détermine le nombre de facteurs comparant deux courbes, une associée aux valeurs propres de la matrice de variance-covariance empirique, l'autre

associée aux valeurs propres moyennes de matrices de variance-covariance générées aléatoirement sous l'hypothèse d'absence d'une structure factorielle (de même dimensions). On retiendra alors le nombre de facteurs associés aux valeurs propres des données observées qui sont supérieures aux valeurs propres moyennes des matrices simulées.

D'autres critères sont également proposés dans la littérature, basés sur des critères de qualité d'ajustement de modèle notamment [Revelle and Rocklin, 1979]. En particulier, si les paramètres sont estimés par Maximum de Vraisemblance, on dispose d'un test validité du modèle à q_0 facteurs. Si on note $\hat{\Sigma}_{q_0} = \hat{B}_{q_0}\hat{B}'_{q_0} + \hat{\Psi}_{q_0}$ et S la matrice de variance-covariance empirique, la statistique $T_{q_0} = n - \frac{11+2m+4q_0}{6} log\left(\frac{|\hat{\Sigma}_{q_0}|}{|S|}\right)$ suit une loi du χ^2 à $1/2[(p-q_0)^2 - p - q_0]$ degrés de liberté sous l'hypothèse nulle selon laquelle q_0 est le bon nombre de facteurs. En pratique, on teste successivement les modèles à $q_0 = 1, \ldots, Q_{max}$ facteurs. Il existe également des critères portant sur la vraisemblance du modèle pénalisée (AIC d'Akaïké : $-2ln(L) + 2(q(m+1)-1)$ ou BIC de Schwartz : $-2ln(L)+ln(n)(q(m+1)-1)$) qui une fois minimisés permettent d'obtenir des modèles parcimonieux.

Cependant, il n'y a pas de consensus parmi le grand nombre de critères sur celui qui est le plus approprié. De nombreuses études ont comparé les performances des différents critères [Ford et al., 1986, Velicer et al., 2000]. La conclusion générale est que le critère de Kaiser (K1) ou le critère du coude, malgré leur popularité, sur-estiment le nombre de facteurs, en particulier lorsque la taille d'échantillon est petite. $K1$ doit en effet être considéré comme un indicateur de borne supérieure pour Q, puisque'il s'agit d'une borne inférieure pour le rang de la matrice de covariance. Le critère du coude souffre également d'une part de subjectivité, la démarcation pouvant être peu évidente en cas de petits échantillons ou si la structure de dépendance est faible. Les tests de qualité d'ajustement, basés sur des tests de χ^2, sont eux sensibles à la taille d'échantillon et leur utilisation en grande dimension est compromise. Pour de nombreux auteurs [Hayton et al., 2004, Velicer et al., 2000], l'analyse parallèle donc est la méthode à privilégier. C'est d'ailleurs un algorithme basé sur ce principe qui est utilisé dans SVA [Leek, 2008], à travers l'algorithme de Buja and Eyuboglu [1992]. Le principal problème de l'analyse parallèle en grande dimension est lié au temps de calcul.

Critère basé sur l'inflation de variance du nombre de faux-positifs Nous proposons un critère adapté au contexte des tests multiples. Développé initialement dans ce cadre, ce critère est néanmoins également utilisable pour l'Analyse en Facteurs en général.

L'Analyse en Facteurs a pour objectif de modéliser la structure de covariance, ou de façon équivalente la structure de corrélation. Nous considérons le modèle suivant pour la variable Y_k centrée, déjà défini au préalable en (2.1) et qui suppose l'existence de q facteurs : $Y_k = Zb_k'^{(q)} + \varepsilon_k^{(q)}$.

On définit alors la corrélation résiduelle, en supposant un modèle à q facteurs communs, par :

$$\rho_{kk'}^{(q)} = \frac{\sigma_k\sigma_{k'}\rho_{kk'} - b_k^{(q)}b_{k'}'^{(q)}}{\sqrt{\Psi_k^{(q)}\Psi_{k'}^{(q)}}} \tag{4.9}$$

Si l'ensemble de la structure de dépendance est correctement modélisée par le modèle d'Analyse en Facteurs à q facteurs, alors la corrélation résiduelle de ce modèle est nulle. On propose une stratégie

de comparaison des ajustements à $q = 1, 2, \ldots, q_{max}$ facteurs et choisir le nombre de facteurs qui minimise un certain critère de qualité d'ajustement, comme la somme des carrés des écarts entre S et $S_{FA} = \hat{B}^{(q)}\hat{B}'^{(q)} + \hat{\Psi}^{(q)}$ par exemple.

Toutefois, dans le cadre des tests multiples, une fonction, appelée $D^{kk'}(t)$ (DÉFINITION 2.2.1), permet de mettre en relation la dépendance et la variance du nombre de faux-positifs. Le critère que nous proposons ici consiste à déterminer le nombre de facteurs qui minimise l'inflation de variance due à la dépendance. En particulier, on montre que la variance minimale est atteinte pour l'indépendance.

PROPOSITION 4.3.1 (Choix du nombre Q de facteurs du modèle d'Analyse en Facteurs)**.** *$\hat{D}^{kk'}(t)^{(q)}$ correspond à la valeur de la fonction définie en 2.2.1 dans le cas où un modèle à q facteurs est ajusté, en considérant les corrélations empiriques entre les variables k et k'. On considère la somme des valeurs de cette fonction obtenues à partir des corrélations empiriques résiduelles $\hat{\rho}^{(q)}_{kk'}$ entre chaque paire de variable Y_k et $Y_{k'}$, $k \neq k' \in \mathcal{M}$ définies en* (4.9)*.*

On estime alors Q tel que :

$$\hat{Q} = \underset{q\in[0;q_{max}]}{\operatorname{argmin}} \left\{ \frac{1}{m(m-1)} \sum_{k\neq k'\in\mathcal{M}} D^{kk'}(t)^{(q)}) \right\}$$

Cette technique est maintenant illustrée à partir des données simulées à l'EXEMPLE 6. La FIGURE 4.3 montre un exemple de l'algorithme de recherche du nombre de facteurs pour un tableau de données simulé de chacun des scénarios. Pour chaque tableau de données, on applique l'algorithme EMFA en considérant successivement des modèles de 0 à 10 facteurs. A chaque fois, on détermine Q à partir des corrélations résiduelles. Sur la FIGURE 4.3, la ligne verticale rouge correspond au minimum du critère, la ligne verticale verte correspond à un "saut relatif" inférieur à $0,001$ pour le critère entre 2 modèles.

Pour ces exemples, les deux règles de décision (minimisation du critère ou saut relatif entre deux valeurs du critère pour deux valeurs de q successives) donnent des résultats similaires. Pour le scénario 2, la décision est différente. Cependant, la variation du critère entre les deux modèles proposés (3 ou 5 facteurs) est très faible. En pratique, on privilégierait le modèle le plus simple, celui à 3 facteurs, étant donné que la complexité du modèle est plus faible que pour le modèle à 5 facteurs tout en ayant une inflation de variance du nombre de faux-positifs du même ordre de grandeur. Pour le scénario 0 (indépendance), la valeur du critère augmente avec q. Dans ce cas, aucun facteur n'est à retenir dans le modèle. Pour les scénarios 1 et 2, la structure de dépendance est faible et le critère varie peu en fonction de q. Néanmoins, le critère proposé conduit à considérer un modèle à 1 ou 3 facteurs. Quand la structure de dépendance devient plus marquée, pour les scénarios 3 à 9, la décision du choix du nombre de facteurs devient plus évidente et on retient un modèle qui permet de minimiser l'inflation de variance tout en ne conservant qu'un petit nombre de facteurs.

Ces remarques se généralisent lorsqu'on applique cette méthode à l'ensemble des 1000 tableaux de données simulées pour chacun des scénarios. La FIGURE 4.4 présente les histogrammes des distribu-

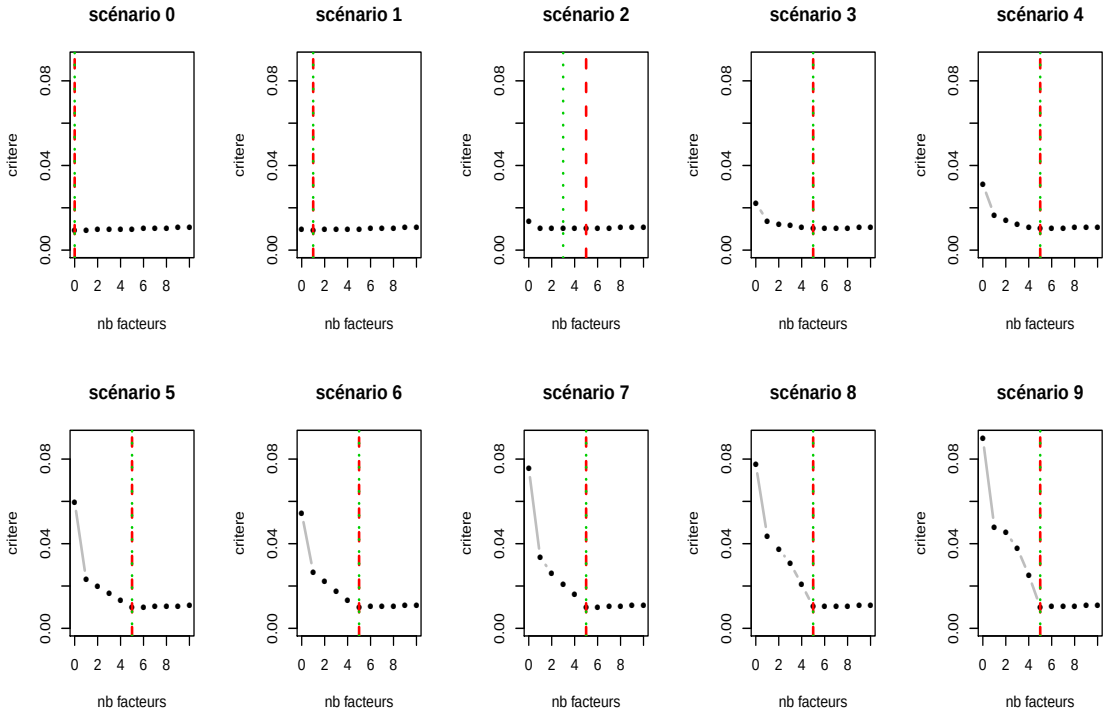


FIG. 4.3 – Choix du nombre de facteurs : exemples de 10 tableaux de données issus de chacun des 10 scénarios

tions des estimations du nombre de facteurs pour les 10 scénarios, dans le cas où le vrai nombre de facteurs, pour chacun des scénarios, est $Q = 5$ et $n = 50$. Lorsque la structure de dépendance est bien marquée (à partir du scénario 3), le nombre de facteurs déterminé par le critère est toujours 5. Ce résultat est à nuancer légèrement pour le scénario 3 car dans ce cas, le bon nombre de facteurs a été trouvé pour 98% des jeux de données et a été légèrement sous-estimé ($\hat{Q} = 4$) pour les 2% restant.

Pour les scénarios où les jeux de données sont simulés avec une faible part de variabilité commune (scénarios 0 à 3), la méthode tend à sous-estimer Q. Cette observation est un point positif de la méthode. En effet, lorsque la variabilité commune est faible, il est cohérent que cette information soit concentrée sur un petit nombre de facteurs, cela permet d'avoir des communalités (ou des *loadings*) plus élevées.

Pour les mêmes données simulées, le nombre de facteurs a ensuite été estimé par l'analyse parallèle [Montanelli and Humphrey, 1976], méthode basée sur du rééchantillonnage qui est souvent présentée comme performante par un grand nombre d'auteurs. Cette méthode met également en évidence le bon nombre de facteurs lorsque la structure de dépendance est forte, et à tendance à sous-estimer Q pour les premiers scénarios. Néanmoins, cette méthode présente deux inconvénients. Le premier est lié à la méthode elle-même : le rééchantillonnage en grande dimension demande un temps de calcul important. Estimer le nombre de facteurs nécessite les simulations de d'un grand nombre de tableaux de données "aléatoires" de même dimensions et d'en calculer les valeurs et vecteurs propres.

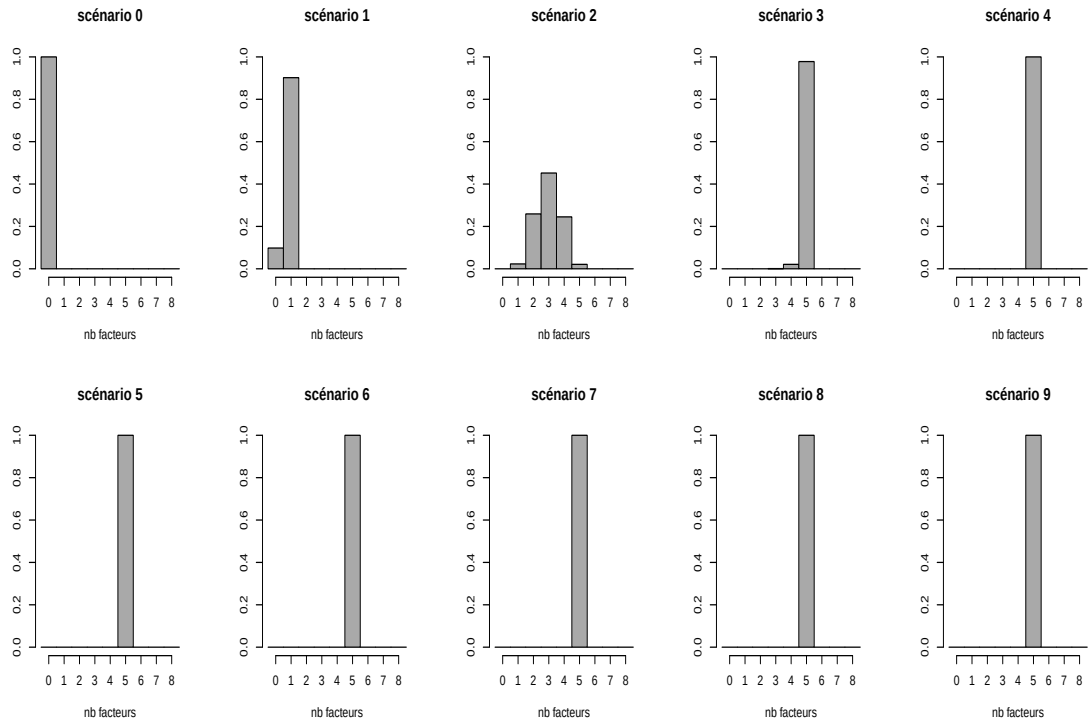


FIG. 4.4 – Distributions des estimations du nombre de facteurs pour les 10 scénarios

Le second inconvénient est que, pour les données des scénarios 0 et 1, cette méthode met parfois en évidence une structure à 10 facteurs communs. Cela est incohérent avec la structure réelle des données, proche de l'indépendance. Néanmoins, il est important ici de remarquer que ces cas extrêmes sont principalement dû à l'automatisation de la méthode.

Ainsi, la méthode proposée, basée sur la minimisation du critère d'inflation de variance (PROPOSITION 4.3.1) permet d'obtenir le bon nombre de facteurs lorsque la structure de dépendance est marquée, et un nombre de facteurs cohérent pour l'interprétation et la pertinence du modèle (communalités élevées) en général.

D'un point de vue pratique, la mise en oeuvre de ce critère nécessite le calcul de la somme de $m \times (m-1)$ termes. La matrice de corrélation étant symétrique, il ne s'agit en fait que de calculer $m \times (m-1)/2$ termes. Néanmoins, ce nombre reste en général très élevé (m est de l'ordre de plusieurs milliers) et nous proposons une approximation pour le critère. On considère un ensemble de η valeurs équi-réparties sur $[0;1]$, puisque la fonction $D^{kk'}(t)$ est symétrique autour de $\rho = 0$. En pratique, considérer un pas de $0,01$ semble être un bon compromis entre une précision suffisante et un nombre de valeurs faible pour augmenter l'efficacité numérique des calculs. On compte alors le nombre d'occurrence de chacune de ces η valeurs de ρ parmi les corrélations arrondies $\rho_{kk'}$, à la précision choisie :

$$\sum_{k \neq k' \in \mathcal{M}} D^{kk'}(t) \approx \sum_{j=1}^{\eta} n_j D(t; \rho_j) \tag{4.10}$$

où n_j représente le nombre de paires de variables $\{k; k'\}$ dont la corrélation est arrondie, en valeur absolue, à ρ_j. Owen [2005] propose une approximation similaire. Si la répartition des corrélations est symétrique, ce qui est le cas de la plupart des applications étudiées, alors l'approximation (4.10) est précise et permet de gagner en vitesse de calculs.

Une autre approximation intéressante est celle implémentée dans le package `FAMT` de `R` [Causeur et al., 2010] :

$$\frac{1}{m(m-1)} \sum D^{kk'}(t) \approx \int_{[0;1]} D^{kk'}(t) f(\rho) d\rho$$

où $f(\rho)$ est une fonction de densité de la distribution des corrélations. On l'estime alors par échantillonnage dans la série des corrélations observées.

Conclusion

Parmi les méthodes à disposition pour estimer les paramètres du modèle à variables latentes étudié dans cette thèse, nous montrons l'intérêt d'utiliser les estimateurs du Maximum de Vraisemblance. La mise en œvre de l'Analyse en Facteurs sur de grands volumes de données doit néanmoins se faire par des méthodes adaptées. A cet effet, nous proposons un algorithme de type *EM*.

Le choix du nombre de facteurs à inclure dans le modèle est également une étape cruciale de l'Analyse en Facteurs. Nous proposons de minimiser un critère définit comme l'inflation de variance du nombre de faux-positifs (V_t). L'utilisation de ce critère dans le cadre de la procédure FAMT présentée au CHAPITRE 3 prend alors un intérêt particulier.

Une propriété importante résultant de cette étude est que plus la dépendance est marquée, plus il sera facile de la modéliser et meilleure sera l'estimation des paramètres du modèle, en particulier en présence de petits échantillons. La conséquence pour l'utilisation du modèle d'Analyse en Facteurs dans le cadre des tests multiples est que lorsque la structure de dépendance est faible, il vaut mieux utiliser les procédures développées sous l'hypothèse d'indépendance. Le critère de choix du nombre de facteurs proposé dans ce chapitre permet cette stratégie.

La méthode FAMT est maintenant illustrée au CHAPITRE 5 par deux études de cas.

CHAPITRE 5

ÉTUDES DE CAS : MISE EN ŒUVRE DE FAMT POUR L'ANALYSE DE DONNÉES GÉNOMIQUES

Résumé Ce chapitre propose l'application de FAMT pour l'analyse de données génomiques, à travers deux études de cas. Ces analyses sont menées à l'aide du package `FAMT` [Causeur et al., 2010] intégrant les outils méthodologiques présentés aux chapitres précédents. Nous montrons ici l'intérêt de la procédure FAMT pour les biologistes en termes d'identification et d'interprétation des facteurs de confusion, de recherche de QTL ou encore d'inférence sur réseaux géniques.

Sommaire

Introduction

La génétique s'intéresse au lien entre le génotype d'un individu, c'est-à-dire son patrimoine génétique, et son phénotype, c'est-à-dire l'ensemble des caractères (non-)observables de cet individu. Les caractères constituant le phénotype d'un individu sont très généralement liés à la présence de protéines, qui sont les constituants essentiels des êtres vivants et qui sont caractérisées par leur séquence d'acides aminés. Cette séquence est codée génétiquement : la synthèse d'une protéine est le résultat de l'expression d'un gène. Le phénotype s'interprète donc comme l'expression des gènes, ou la réalisation du génotype, mais aussi par les effets du milieu et de l'environnement.

Depuis quelques années, on note un regain d'intérêt pour des problématiques liées à l'étude de ce lien entre génotype et phénotype, suite à l'apparition de nouvelles biotechnologies à la fin des années 1990, qui permettent d'accéder aux séquences génomiques de chaque individu. On a vu en effet le développement des puces à ADN, ou biopuces. Utilisant les propriétés d'hybridation de l'ADN et des marqueurs fluorescents, elles permettent de mesurer simultanément le niveau d'expression de plusieurs dizaines de milliers de gènes, si on considère l'étude de l'ensemble du génome d'un organisme. Les données qui résultent de cette technologie sont utilisées de plusieurs façons, notamment à des fins de diagnostics médicaux, pour mesurer l'effet d'un traitement sur différentes populations, ou encore dans l'optique de définir des groupes de gènes ayant des profils d'expression similaires.

Les données issues des biopuces sont une énorme source d'information pour les biologistes : un tel volume de données ouvre des perspectives d'analyses nouvelles, en proposant une vue d'ensemble du génome. L'accès à cette information est essentiel en vue de mieux comprendre le rôle et la fonction de chacun des gènes. Cependant, de ces nouvelles expériences résultent des problèmes statistiques complexes. Parmi les challenges à relever, le problème de la grande dimension des données biologiques est le premier qui vient à l'esprit : il s'agit d'analyser simultanément un très grand nombre de gènes, souvent de l'ordre de plusieurs milliers. Dans le cas des biopuces, les données recueillies sont sous la forme classique de tableaux individus (puces) / variables (expressions de gènes), mais leur particularité est de décrire un petit nombre d'individus (quelques dizaines) à l'aide de très nombreuses variables (plusieurs milliers), le niveau d'expression de chacun des gènes. Les données génomiques sont par nature organisées selon une structure très complexe : de nombreux liens existent entre les gènes, de par leurs fonctions biologiques, les processus biologiques dans lesquels ils sont impliqués, les réseaux auxquels ils appartiennent, etc. Cette interdépendance entre gènes se retrouve au niveau des mesures d'expressions. La modélisation de la structure de dépendance des données génomiques est également une étape clé de l'analyse de ce type de données. De plus, l'information biologique, le signal d'intérêt, est observée en même temps qu'un certain nombre de facteurs de confusion. Le lien déduit entre la variable d'intérêt et la condition expérimentale étudiée peut alors s'avérer erroné, et les effets observés devraient être attribués à ces facteurs. De plus, les facteurs de confusions ne sont pas toujours contrôlés par le plan d'expérience et sont parfois non observables.

Un des objectifs pour les biologistes qui mènent des expériences à l'aide de biopuces est d'identifier les gènes qui sont différemment exprimés lorsque les conditions expérimentales sont modifiées, d'un

régime alimentaire à un autre, d'un état sain à un état malade ou selon la dose de médicament administrée par exemple : on parle d'analyse différentielle. Notons que cette étape est en général la première menée lors de l'analyse de données génomiques, le sous ensemble de gène identifiés servant ensuite pour des analyses ultérieures.

Cette problématique biologique relève en terme statistique des tests multiples. Ce chapitre est présenté comme une étude de cas, avec l'objectif de présenter l'utilisation du package `R` appelé `FAMT` [Causeur et al., 2010] intégrant l'ensemble des outils développés dans cette thèse. Il s'agit en particulier de montrer l'intérêt de la démarche proposée pour des biologistes, en termes d'identification et d'interprétation de facteurs de confusion, de recherche de QTL, ou encore d'inférence sur des réseaux géniques. Les graphiques illustrant les résultats sont des sorties de ce logiciel et les différentes analyses présentées sont issues de fonctions programmées dans `FAMT`.

Dans un premier temps, on présente l'analyse de données publiques, fréquemment utilisées dans la littérature pour illustrer les méthodes d'analyse différentielle issue d'une étude menée par Hedenfalk et al. [2001] dans le cadre de la recherche sur le cancer du sein. Dans un deuxième temps, nous nous appuierons sur un exemple de problématique d'analyse différentielle à travers une étude menée à l'INRA de Rennes par l'unité de Génétique Animale. Ces données sont analysées par les méthodes classiques de tests multiples tout d'abord, et par FAMT, présentée au CHAPITRE 3. Des éléments de validation biologique des résultats sont présentés également pour le deuxième exemple, réalisés par les généticiens de l'unité de Génétique Animale INRA/Agrocampus. En effet, l'analyse différentielle n'est dans ce cas qu'une première étape de l'analyse des données expérimentales. Ces deux exemples ont été choisis car ils permettent d'illustrer les deux configurations de distribution des probabilités critiques en présence de dépendance identifiés au CHAPITRE 2.

5.1. Étude 1 : identification de gènes impliqués dans le développement de tumeurs de cancer du sein

5.1.1 Présentation du jeu de données

La première application présentée s'appuie sur une étude menée dans le cadre de la recherche sur le cancer à travers l'analyse de données d'expressions géniques. La plupart des cancers du sein héréditaires sont dus à une mutation du gène BRCA1 ou du gène BRCA2. L'étude réalisée par Hedenfalk [Hedenfalk et al., 2001] vise à mettre en évidence les gènes différemment exprimés entre les tumeurs liées à ces deux types de mutation (BRCA1 ou BRCA2). Les expressions géniques de $m = 6384$ gènes pour 7 tumeurs ayant une altération pour le gène BRCA1 et 8 tumeurs ayant une altération pour le gène BRCA2 ont été analysées par puces à ADN de type lame de verre. Les données originales sont déposées sur un site public (`http://research.nhgri.nih.gov/microarray/NEJM_Supplement/`). Outre les données relatives aux tumeurs BRCA1 et BRCA2, les données d'un troisième groupe de tumeurs dites sporadiques sont disponibles, mais nous ne les étudions pas ici.

Nous avons également exclu de notre analyse les variables pour lesquels des valeurs étaient inférieures à $0,1$ ou supérieures à 10, valeurs que nous supposons aberrantes comme lors de l'analyse initiale de Hedenfalk. Les valeurs sont ensuite transformée en logarithme base-2. Notre analyse porte donc sur $m = 3030$ gènes. Les données Y se présentent sous la forme d'une matrice $n \times m$ où $m = 3030$ (rapports d'intensités du niveaux d'expression tumeur/référence, passés au logarithme) et $n = 15$ (prélèvements), le type de mutation (BRCA1 ou BRCA2) étant la variable explicative X de l'étude.

Les données sont préalablement normalisées.

5.1.2 Analyse statistique : identification des gènes différentiellement exprimés

Nous nous intéressons au lien entre $m = 6384$ variables d'intérêt, mesurant les expressions géniques des m gènes, et une variable explicative qualitative, le type de tumeur, à 2 modalités. L'objectif est d'identifier les gènes pour lesquels le niveau d'expression génique est lié au type de tumeur. Il s'agit alors, pour chaque variable, de réaliser un test de Student de comparaison de moyennes. L'histogramme des probabilités critiques est donné par la FIGURE 5.1.

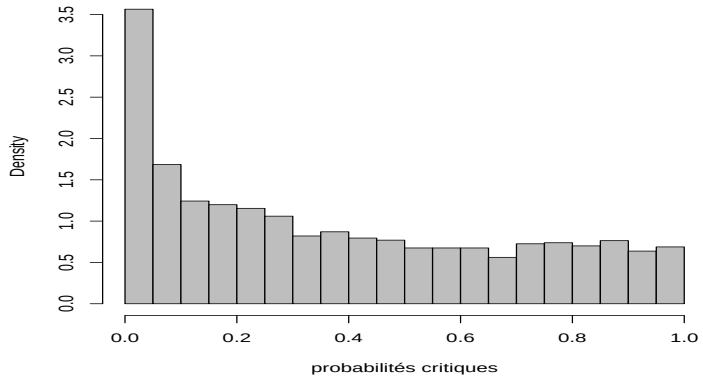
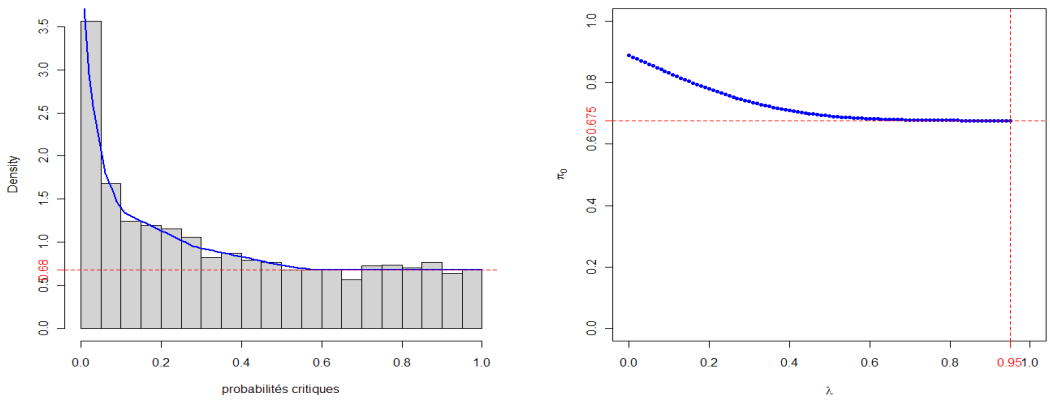


FIG. 5.1 – Histogramme des probabilités critiques des tests de Student - jeu de données *Hedenfalk*

La configuration de l'histogramme de la FIGURE 5.1 nous place dans une situation telle que décrite lors de l'étude de l'impact de la dépendance sur la distribution des probabilités critiques sous l'hypothèse nulle (FIGURE 2.3 - gauche). Sur cet exemple, nous comparons les approches usuelles des tests multiples et la méthode FAMT.

Méthodes usuelles Les approches dites usuelles considérées ici sont les méthodes BH avec estimation de π_0 et la méthode BY. En supposant la densité des probabilités critiques convexe, l'estimation de π_0 est de $0,68$. L'estimateur empirique avec choix du paramètre λ par lissage spline donne $\hat{\pi}_0 = 0,675$ (voir FIGURE 5.2). Nous choisissons la méthode d'estimation basée sur l'estimateur empirique pour la suite de l'étude soit $\hat{\pi}_0 = 0,675$.

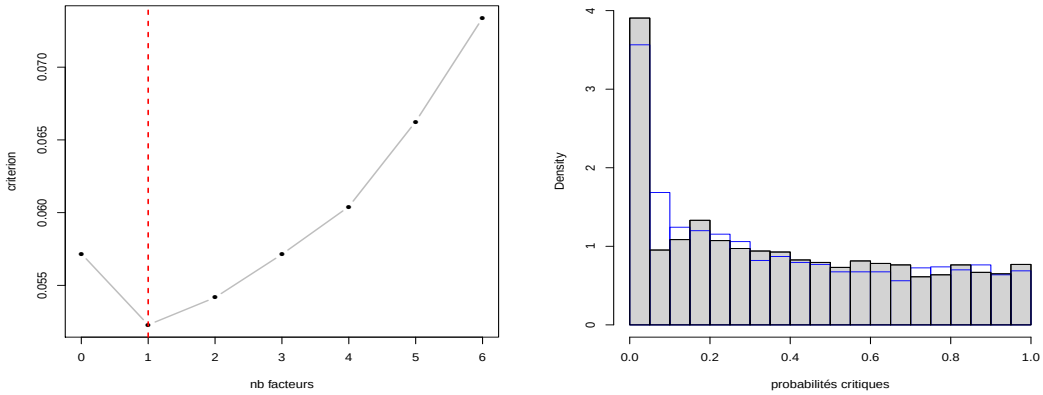


(a) Densité estimée des probabilités critiques de Student (densité supposée convexe, [Langaas et al., 2005])

(b) Estimation de π_0 par l'estimateur empirique, choix de λ par lissage spline [Storey and Tibshirani, 2003]

FIG. 5.2 – Estimation de π_0 pour les données Hedenfalk

Méthode FAMT Les données de l'étude présentent une structure de dépendance à 1 facteur (FIGURE 5.3(a)). L'information commune représente $17,8\%$ de la variabilité de l'ensemble de données. La distribution des probabilités critiques, ajustées de l'effet de ce facteur, sont représentées sur la FIGURE 5.3(b). L'estimation de π_0 à partir des probabilités critiques ajustées donne des valeurs respectivement de $0,69$ et $0,681$ pour les deux méthodes utilisées (voir FIGURE 5.4).

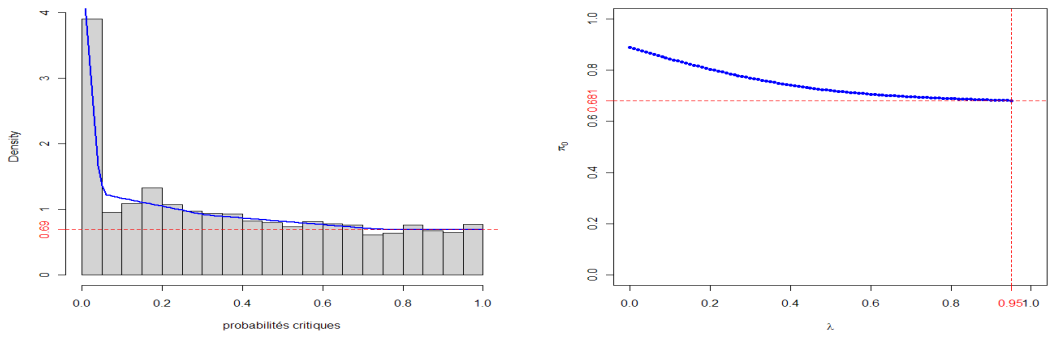


(a) Choix du nombre de facteurs

(b) Distribution des probabilités critiques ajustées

FIG. 5.3 – Mise en oeuvre de FAMT - jeu de données Hedenfalk : choix du nombre de facteurs et calculs des probabilités critiques ajustées - en bleu : probabilités critiques de Student

Nous appliquons ensuite une procédure de tests multiples pour identifier les gènes différentiellement exprimés en fonction du type de tumeur. A partir des probabilités critiques usuelles, les démarches retenues sont la détermination du seuil de rejets des hypothèses assurant un contrôle du FDR au niveau α par la méthode BH avec estimation de π_0 par l'estimateur empirique (Student BH + π_0)



(a) Densité estimée des probabilités critiques ajustées (densité supposée convexe, [Langaas et al., 2005])

(b) Estimation de π_0 par l'estimateur empirique, choix de λ par lissage spline [Storey and Tibshirani, 2003]

FIG. 5.4 – Estimation de π_0 à partir des probabilités critiques ajustées pour les données Hedenfalk

et par la méthode BY (Student + BY). D'autre part, à partir des probabilités critiques ajustées, nous appliquons la méthode BH avec l'estimation de π_0 par la même méthode que précédemment (FAMT+BH+π_0). La FIGURE 5.5 présente le nombre de tests déclarés positifs en fonction du seuil α pour le contrôle du FDR. La procédure BY est très conservatrice, et ne permet de déclarer comme différentiels qu'un petit nombre de gène (1 au seuil $0,05$, 5 au seuil $0,15$). La procédure BH est beaucoup plus puissante. Néanmoins, l'application de cette méthode pour la détermination du seuil permet de déclarer comme différentiels un plus grand nombre de gènes lorsque la méthode est appliquée sur les probabilités critiques ajustées par rapport aux probabilités critiques usuelles (au seuil $0,05$: 191 contre 96 et au seuil $0,15$: 560 contre 435). Pour les deux niveaux de contrôle du FDR pris en exemple, l'ensemble des gènes déclarés différentiellement exprimés par l'approche usuelle font partie de ceux déclarés différentiellement exprimés par l'approche FAMT. Cela n'est pas une propriété de la méthode mais une spécificité de ces données.

En pratique, les biologistes ont souvent un raisonnement qui rejoint l'approche de Storey [2003] (q-value). Pour des raisons de coûts (financiers, humains, temps) l'analyse plus approfondie de gènes déclarés positifs par la procédure de tests multiples devra se limiter à un nombre donné de gènes, par exemple 100. Il s'agit alors de déterminer le risque associé (FDR) au seuil correspondant à la 100^{eme} probabilité critique ordonnée. Ainsi, c'est non pas le nombre de tests déclarés positifs qui importe mais le rang donné à chaque test par sa probabilité critique. Ce rang n'est pas nécessairement le même selon qu'on considère les probabilités critiques brutes ou celles issues des tests ajustés.

Données brutes *vs* données ajustées La FIGURE 5.6 représente les doubles-classifications respectivement des données brutes et des données ajustées de l'effet du facteur, restreintes aux gènes dont la probabilité critique est inférieure à $0,05$, soit 619 (fonction `heatmap` du package `gplots` de `R` [Warnes et al., 2009]). Chaque colonne correspond à l'expression d'un gène et chaque ligne à un individu (puce). La coloration indique le niveau d'intensité de l'expression (la couleur traduit la graduation du niveau d'expression, de rose foncé à clair pour les valeurs négatives, puis de vert clair à

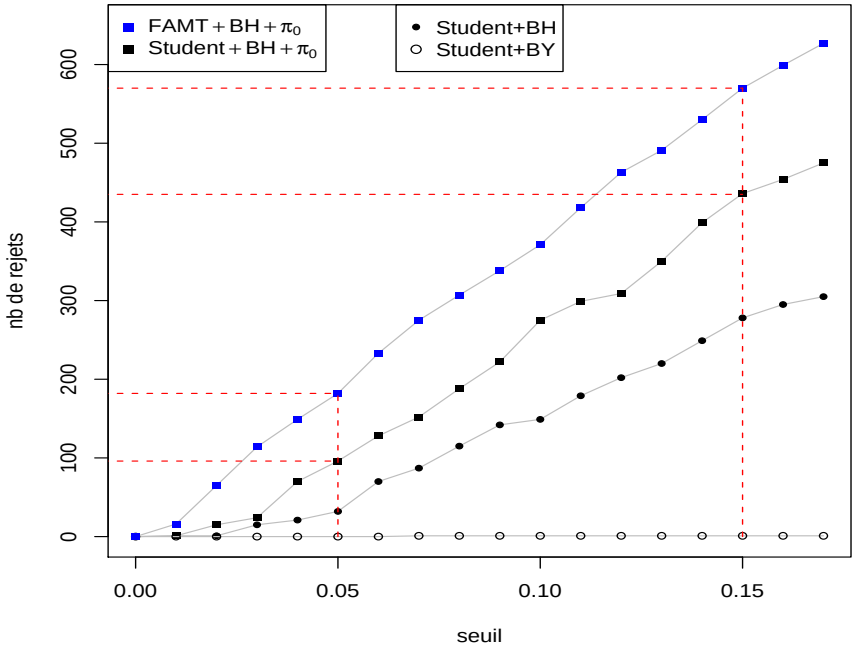
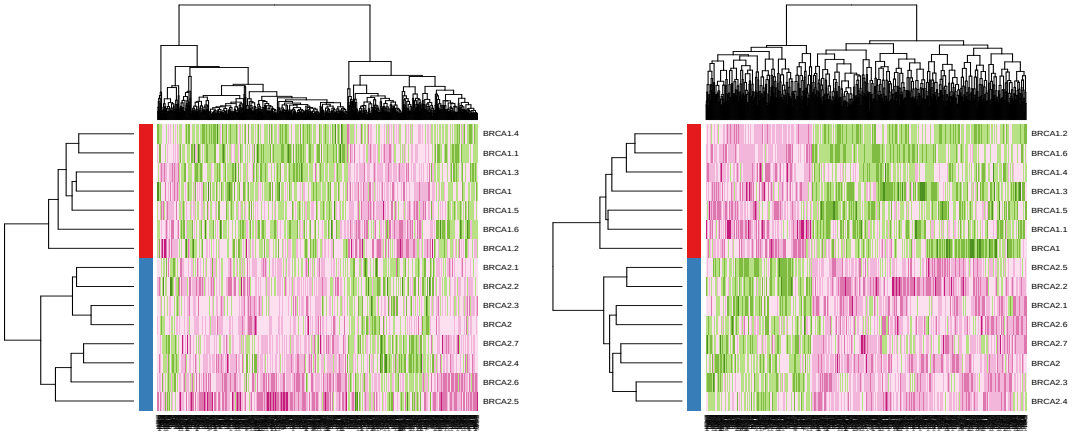


Fig. 5.5 – Nombre de rejets en fonction du seuil choisit pour le contrôle du FDR, pour différentes procédures de tests multiples - données Hedenfalk

foncé pour les valeurs positives). La représentation permet une double classification des profils d'expressions et des individus. Si la représentation des données brutes et des données ajustées regroupent bien les individus selon le type de tumeur (bleu/rouge), la structure des expressions génique apparait nettement plus claire dans le cas des données ajustées.



(a) Données brutes (b) Données ajustées

Fig. 5.6 – Double-classification des données brutes et corrigées

5.2. Étude 2 : identification de gènes impliqués dans le métabolisme des lipides

5.2.1 Présentation du jeu de données

Cet exemple s'appuie sur une étude menée dans le cadre de la recherche sur le métabolisme des lipides chez le poulet à l'INRA de Rennes, par l'unité de Génétique Animale (UMR GARen, INRA/Agrocampus OUEST).

Les données ont été générées initialement pour la détection de QTL (Quantitative Trait Loci), zone d'intérêt du génome contrôlant la variabilité d'un caractère phénotypique quantitatif particulier, ici le caractère gras. Un QTL a ainsi été précédemment détecté sur le chromosome 5 (GGA5), aux alentours de 175cM [LeMignon et al., 2009].

Le but de cette nouvelle analyse [Blum et al., 2010] est alors d'améliorer la caractérisation de la localisation du QTL sur le chromosome, en utilisant des données transcriptomiques. Les expressions de 11213 gènes hépatiques ont donc été mesurées sur 45 poulets à l'aide de biopuces à ADN (une biopuce/individu). La quantité de gras abdominal (poids en grammes) est mesurée également pour chaque individu.

La FIGURE 5.7 illustre le protocole de recueil des données. Les individus étudiés présentent la particularité d'être tous issus du même père, et de 8 mères différentes.

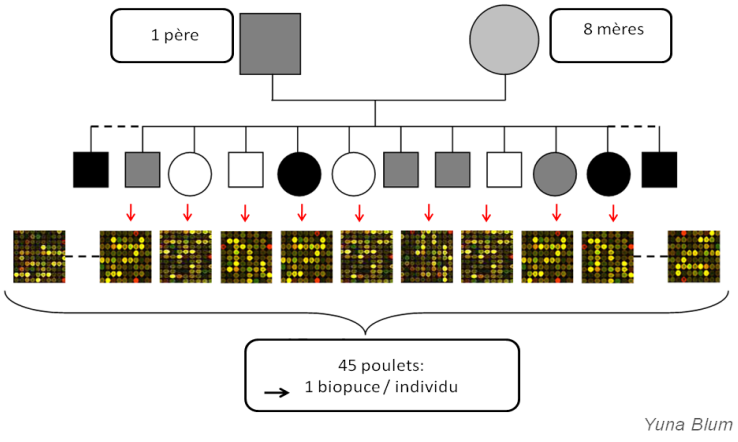


FIG. 5.7 – Protocole expérimental : recueil des données d'expressions géniques par biopuces

Après pré-traitement (normalisation, suppression des données aberrantes...), nous conservons pour notre étude les mesures d'expressions de $m = 9893$ gènes pour $n = 43$ poulets.

La première étape de l'analyse consiste à identifier les gènes différentiellement exprimés en fonction de la quantité de gras abdominal. Nous allons donc utiliser les procédures de tests multiples décrites dans cette thèse.

5.2.2 Analyse statistique : identification des gènes différentiellement exprimés

Nous nous intéressons au lien entre m variables d'intérêt, mesurant les expressions géniques de m gènes, et une variable explicative, le poids de gras abdominal, noté x. Dans l'expérience décrite précédemment, les individus sont issus de 8 mères différentes. Cela peut induire de la variabilité dans les expressions géniques des individus. Nous allons prendre en compte ce facteur de variabilité dans le modèle. Le lien se formalise pour chaque gène k par le modèle d'analyse de covariance suivant, pour les individus issus de la mère j :

$$Y_{k,j} = \beta_0 + \beta_1 x + \alpha_j + \epsilon_k \quad \forall k \in [1;m], j \in [1;8] \tag{5.1}$$

L'objectif est d'identifier les gènes pour lesquels le niveau d'expression génique est lié à la quantité de gras abdominal x. Il s'agit alors, pour chaque variable k, de réaliser un test de Student. L'histogramme des probabilités critiques est donné par la FIGURE 5.8.

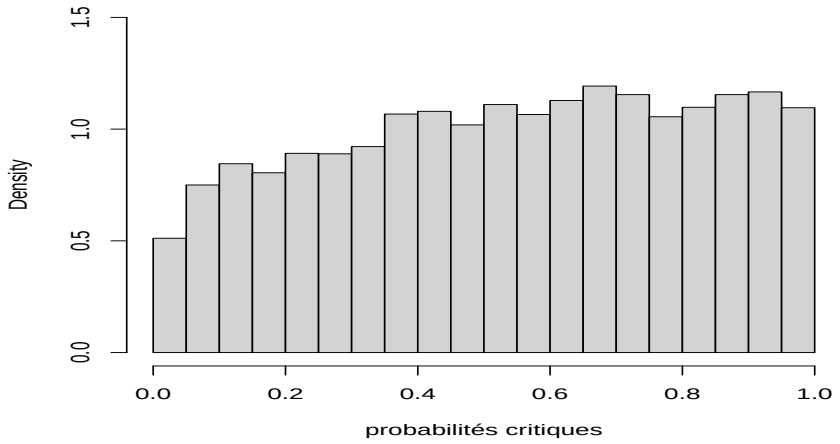


FIG. 5.8 – Histogramme des probabilités critiques des tests de Student - jeu de données *Famille*

La configuration de l'histogramme de la FIGURE 5.8 nous place cette fois dans une situation telle que décrite lors de l'étude de l'impact de la dépendance sur la distribution des probabilités critiques sous l'hypothèse nulle (FIGURE 2.3 - droite).

On obtient un ensemble de $p_1 = 253$ tests dont les probabilités critiques brutes sont inférieures à $0,05$, soit $2,56\%$ des variables : la proportion de probabilités critiques brutes inférieures à 5% est elle-même inférieure à 5%.

L'estimation de π_0 par les méthodes usuelles donne $\hat{\pi}_0 = 1$ et est utilisée dans les procédures BH et BY.

Les probabilités critiques ajustées obtenue par les méthodes BH ou BY sont toutes proches de 1. Aucun test n'est donc déclaré positif par la méthode usuelle, au seuil $\alpha = 0,05$. Le TABLEAU 5.1 donne les 255 plus petites probabilités critiques et leurs valeurs ajustées par les deux procédures.

id	raw p-values	BH	BY
1	8,710792e-05	0,4462354	1
2	9,021235e-05	0,4462354	1
3	1,689088e-04	0,5570049	1
4	5,872947e-04	0,9995822	1
5	8,130185e-04	0,9995822	1
6	1,371078e-03	0,9995822	1
7	1,696091e-03	0,9995822	1
8	2,053841e-03	0,9995822	1
9	2,599947e-03	0,9995822	1
10	2,714114e-03	0,9995822	1
...			
250	0,04951468	0,9995822	1
251	0,04977052	0,9995822	1
252	0,04979009	0,9995822	1
253	0,04995317	0,9995822	1
254	0,05002063	0,9995822	1
255	0,05040713	0,9995822	1
...			

TAB. 5.1 – 255 plus petites probabilités critiques issues des tests de Student et leurs valeurs ajustées par les procédures BH et BY

Nous mettons maintenant en œuvre la méthode FAMT, telle que décrite à la section 3.3 et implémentée dans le package `FAMT` de `R`. La FIGURE 5.9 représente la valeur du critère d'inflation de variance (PROPOSITION 4.3.1) en fonction du nombre de facteurs.

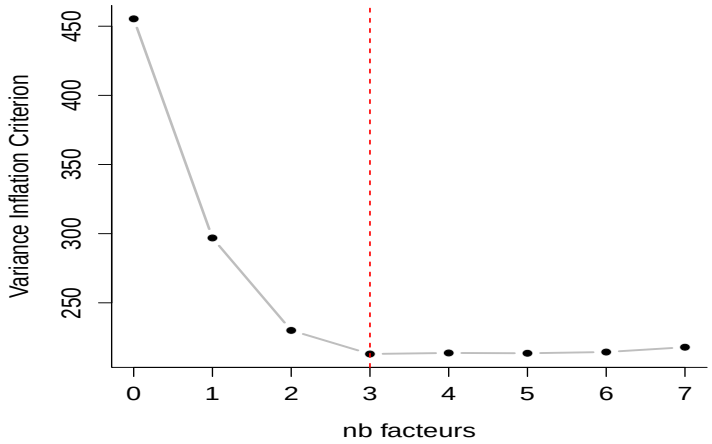


FIG. 5.9 – Choix du nombre de facteurs - jeu de données *Famille*

Le critère est minimisé pour $\hat{Q} = 3$ facteurs. Le modèle considéré est alors :

$$Y_{k,j} = \beta_0 + \beta_1 x_1 + \alpha_j + Z b'_k + \varepsilon_k \quad \forall k \in [1;m], j \in [1;8] \tag{5.2}$$

avec Z de dimension $n \times Q$ et b_k de dimension $1 \times Q$. La FIGURE 5.10 montre la distribution des probabilités critiques ajustées. On superpose (en bleu) l'histogramme des probabilités brutes.

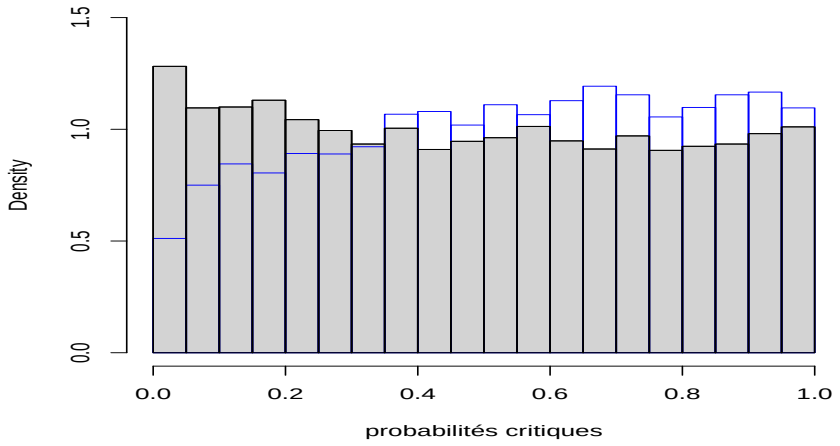


FIG. 5.10 – Histogramme des probabilités critiques des tests ajustés. En bleu : histogramme des probabilités critiques brutes (tests de Student) - jeu de données *Famille*

D'après le TABLEAU 5.2, 634 gènes ont une probabilité critique ajustée inférieure à $0,05$. Si on s'intéresse au FDR, un niveau de contrôle au seuil $0,15$ (procédure BH) permet de déclarer 6 gènes comme différentiellement exprimés.

5.2.3 Validation biologique

Parmi les 6 gènes identifiés, un a suscité l'intérêt des généticiens. Il s'agit d'un gène contrôlé par la zone d'intérêt du chromosome 5 étudiée, et qui intervient dans une réaction de méthylation dans le métabolisme des folates. Cependant, l'analyse différentielle menée ici s'inscrit dans une étude plus large que la simple identification de gènes différentiellement exprimés. On cherche en effet à mieux comprendre les mécanismes sous-jacents à la régulation génétique du caractère gras. La suite de l'étude nécessite de détenir une liste de gènes liés au caractère d'intérêt assez conséquente.

Dans un premier temps, on confirme par l'analyse exploratoire des données d'expressions la pertinence d'un point de vue biologique de la liste de gènes corrélés au caractère gras trouvée par FAMT. Les individus ont été répartis dans trois groupes par les biologistes, en fonction de leur masse de gras abdominal : individus maigres (L : *lean*), individus intermédiaire (I) et individus gras (F : *fat*). La FIGURE 5.11 représente les premiers plans factoriels des Analyse en Composantes Principales menées sur les $p = 634$ gènes détectés par FAMT. Le graphique de gauche correspond à l'ACP des données brutes, celui de droite à l'ACP menée sur les données corrigées de l'effet des facteurs : $\tilde{Y} = Y - \hat{Z}\hat{B}'$.

id	raw p-values	BH	BY
1	8,429627e-06	0,0567365	0,554705
2	1,147004e-05	0,0567365	0,554705
3	3,176313e-05	0,1047442	1
4	4,611245e-05	0,1076170	1
5	6,142819e-05	0,1076170	1
6	6,526859e-05	0,1076170	1
7	1,719728e-04	0,2306635	1
8	1,865266e-04	0,2306635	1
9	2,534048e-04	0,2642470	1
10	2,839211e-04	0,2642470	1
11	2,938155e-04	0,2642470	1
12	4,514420e-04	0,3721763	1
13	5,209183-04	0,3964188	1
14	6,272664e-04	0,4135403	1
15	6,423067e-04	0,4135403	1
...			
630	0,04877718	0,76595659	1
631	0,04917085	0,76897412	1
632	0,04919935	0,76897412	1
633	0,04920253	0,76897412	1
634	0,04931387	0,76949863	1
635	0,05003253	0,77948321	1
...			

TAB. 5.2 – 635 plus petites probabilités critiques issues des tests ajustés par rapport aux facteurs et leurs valeurs ajustées par les procédures BH et BY

L'ACP des données corrigées montre nettement une meilleure séparation des individus selon leur quantité de gras selon le premier axe factoriel, ce qui illustre l'intérêt des données ajustées pour des analyses ultérieures, comme l'identification de QTL [Blum et al., 2010].

D'autre part, les tests d'enrichissement basés sur un test exact de Fisher comme proposé par Man et al. [2000] permettent de caractériser les propriétés fonctionnelles d'un ensemble de gènes, en utilisant les informations disponibles dans les bases de données de termes fonctionnels (Gene Ontology, KEGG). Nous avons comparé de ce point de vue les deux listes de gènes déclarés différentiellement exprimés, obtenues respectivement par l'approche classique et par FAMT. Ici, seule la liste obtenue à partir des données ajustées permet de mettre en évidence des termes biologiques liés au métabolisme des lipides (processus de biosynthèse des hormones stéroïdiennes).

Interprétation des facteurs communs Les facteurs permettent de modéliser des sources possibles de variabilités non observés lors de l'expérience. De l'information externe disponible (concernant les individus ou les variables de l'étude) peut permettre d'interpréter les facteurs communs. Concernant les données présentées ici, cette étude a été menée par Blum et al. [2010] et est développée également dans le tutoriel du package `R FAMT` [Causeur et al., 2010]. Les principaux résultats sont décrits ci-après. On sait en particulier que les individus étudiés dans cette expérience sont issus de 4 lots d'éclosions différents. On a mesuré également leur poids à 9 semaines. Pour chaque facteur,

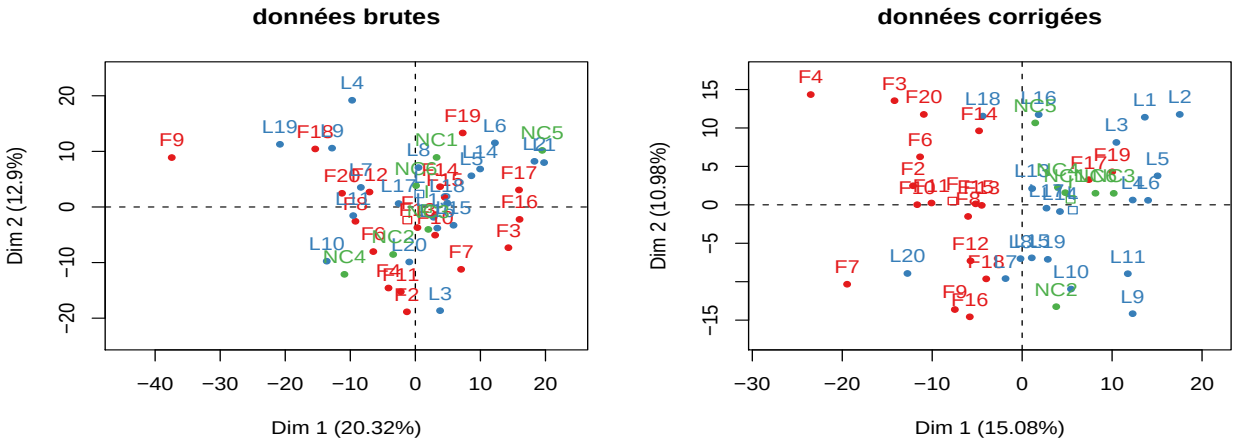


FIG. 5.11 – Premiers plans factoriels des ACP menées sur les $p = 634$ gènes détectés par FAMT (nuages des individus) - Bleu : individus maigres (L) - Vert : individus intermédiaires - Rouge : individus gras (F)

nous étudions donc son lien avec ces variables. Les trois analyses de covariance donnent les résultats suivant (au seuil $\alpha = 0,05$).

	Lot	Pds9s
Factor 1	0.006437319	0.27847793
Factor 2	0.258859549	0.00124608
Factor 3	0.271648846	0.96813322

TAB. 5.3 – Probabilités critiques des tests de l'effet du Lot et du Poids à 9 semaines pour chacun des trois facteurs communs mis en évidence par FAMT

Pour le Facteur 1, les scores varient en fonction du lot d'éclosion de façon significative ($p = 1,43.10^{-2}$). Les poulets du lot 4 ont plutôt des scores positifs élevés, tandis que ceux des autres lots ont des scores moyens (voir FIGURE 5.12. Le coefficient R^2 du modèle est de $26,8\%$. Le Facteur 2 quand à lui, est significativement lié au poids à 9 semaines ($p = 1,25.10^{-3}$, $R^2 = 22,7\%$) et les individus ayant un poids faible à 9 semaines ont un score significativement plus élevé pour ce facteur.

	Block	Column	Row
Loadings 1	8.148075e-25	0.2368072	0.21767892
Loadings 2	3.328477e-19	0.9323152	0.01889079
Loadings 3	0.000000e+00	0.1030426	0.68201616

TAB. 5.4 – Probabilités critiques des tests de l'effet de caractéristiques techniques de l'expérience chacun des trois facteurs communs mis en évidence par FAMT

D'autre part, des biais techniques semble avoir un impact également sur la structure de corrélation des mesures d'expressions géniques. La localisation du gène sur la puce lors de l'expérience a un effet

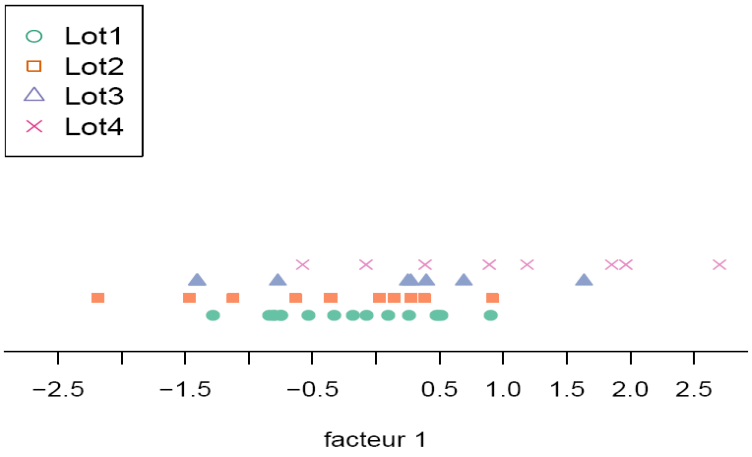


FIG. 5.12 – Scores obtenus pour le Facteur 1 en fonction du lot d'éclosion

significatif (au seuil $\alpha = 0,05$) sur les *loadings* (voir TABLEAU 5.4). L'effet du "bloc" est capturé par l'ensemble des facteurs, et celui de la "ligne" par le Facteur 2. Ces effets peuvent être induits par les pré-traitements de normalisation [Qiu et al., 2005].

La pertinence de la méthode FAMT pour la détection de gènes différentiellement exprimés dans un contexte biologique donné est confirmée. En particulier, l'utilisation de ce modèle a déjà permis d'améliorer la précision de la localisation de la région contrôlant le caractère gras sur le chromosome 5 [Blum et al., 2010].

Conclusion et perspectives

Dans ce chapitre, deux exemples d'application de la méthode FAMT ont été présentés. Le premier s'appuie sur des données publiques. Ces données sont fréquemment utilisées dans de nombreux articles pour illustrer l'apport d'une méthode et la comparer avec celles existantes. C'est également l'objectif de cette étude. Nous nous sommes appuyés sur cet exemple pour comparer les différentes méthodes à disposition pour répondre à la problématique biologique posée dans l'étude initiale de Hedenfalk, et illustrer l'apport de notre approche dans ce cadre.

Néanmoins, cet exemple reste illustratif et les conclusions obtenues ne sont pas approfondies en terme d'interprétation.

Le second exemple se place dans le cadre d'une collaboration avec le laboratoire de Génétique Animale de l'INRA à Agrocampus OUEST. Le contexte général de l'analyse de données génomiques est ce qui a motivé au départ les travaux réalisés au cours cette thèse. Les données expérimentales de cette deuxième étude ont été générées dans le cadre d'une étude aux objectifs plus large que la simple identification de gènes différentiellement exprimés. Les résultats obtenus par FAMT ont été prometteurs, et les avantages pour l'analyse biologique de ces données ont été évoquée dans la SECTION 5.2.3 de ce chapitre. En particulier, les généticiens ont pu améliorer la précision de la localisation d'une zone d'intérêt (QTL) sur le chromosome 5 dans l'étude des mécanismes biologiques

impliqués dans la régulation du caractère gras [Blum et al., 2010].

Par ailleurs, un des enjeux majeurs en génomique est la recherche de liens fonctionnels entre les gènes : on cherche à comprendre la logique des régulations entre les gènes déclarés comme étant en lien avec le caractère d'intérêt, et par extension, celle des processus biologiques associés. On formalise les régulations par un réseau de gènes dont les noeuds représentent les gènes et les arêtes les régulations entre les gènes. Les modèles graphiques permettent la représentation des relations de dépendances conditionnelles entre les nœuds du graphe. Cette problématique s'inscrit dans la continuité des premiers résultats de cette thèse, notamment autour de la modélisation de la structure de dépendance sous forme de graphes et de l'inférence sur les réseaux biologiques.

Plus particulièrement, la modularité dans les réseaux de régulation constitue une hypothèse intéressante tant au niveau de la biologie qu'au niveau de la modélisation statistique. Dans ce cadre, d'un point de vue statistique, nous nous intéressons à l'apport d'une modélisation sous forme d'une structure en facteurs pour l'estimation de la matrice de variance-covariance d'une part et pour la matrice des corrélations partielles d'autre part. Cette dernière permet de représenter les interactions directes entre gènes. Des premiers travaux à partir de simulations donnent déjà des résultats prometteurs en terme d'estimation de cette matrice par un modèle d'Analyse en Facteurs. Nous pouvons d'ores et déjà nous appuyer également sur des travaux similaires réalisés récemment dans le domaine de l'économie et des finances [Fan et al., 2008].

CONCLUSION

D'une manière générale, ce travail de thèse porte sur l'étude de l'impact de la dépendance sur les propriétés des procédures de tests multiples en grande dimension. Motivé par des applications dans le domaine de l'analyse de données génomiques, il vise à développer un outil statistique qui permet de prendre en compte l'hétérogénéité de données en grande dimension en inférence simultanée.

Cette hétérogénéité peut être due à l'effet de facteurs techniques ou environnementaux par exemple qui sont non-observables ou mal contrôlés par le plan d'expérience. La méthode proposée consiste à identifier un espace linéaire engendré par un ensemble de variables latentes qui modélise cette structure hétérogène en capturant la variabilité partagée par l'ensemble des variables, dans une approche de type Analyse en Facteurs [Mardia et al., 1979].

Les facteurs latents identifiés sont associés à des sources de variation ignorées par l'espérance du modèle et donc assimilées à de l'hétérogénéité. Leek and Storey [2008] donnent également l'exemple de facteurs traduisant une dépendance spatiale, comme dans le cas de l'analyse d'image en médecine. Cette modélisation est donc finalement assez générale.

La plupart des procédures de tests multiples reposent sur l'analyse du processus empirique des probabilités critiques associées aux tests individuels sous l'hypothèse d'indépendance. La règle de décision visant à définir un seuil pour les probabilités critiques permet de tenir compte de la multiplicité des tests dans ce cadre. Il y a alors principalement deux enjeux dans les procédures de tests multiples : le contrôle des taux d'erreurs et indirectement l'estimation de la proportion d'hypothèses nulles, π_0. L'impact de la dépendance sur la stabilité des procédures de tests multiples est un des résultats majeurs de cette thèse. La dépendance induit en effet une variabilité qui perturbe notamment la distribution empirique des probabilités critiques sous l'hypothèse nulle, celle-ci pouvant s'éloigner fortement de la distribution théorique. Les conséquences directes sont d'une part l'augmentation de la variabilité du nombre de faux-positifs (V_t) et d'autre part un biais dans l'estimation de π_0. Plus particulièrement, on montre que les variances du nombre de faux-positifs d'une part, et de π_0 d'autre part, dépendent toutes deux explicitement d'un terme fonction de la corrélation entre les variables. La dépendance a donc des répercussions sur l'estimation des taux d'erreurs, entraînant une instabilité des procédures de tests multiples.

On définit une procédure d'ajustement des variables d'intérêt par rapport aux variables latentes, conditionnellement auxquelles les données sont indépendantes. La dépendance n'est plus seulement prise en compte par une correction des procédures de contrôle du risque de première espèce, mais en amont par l'intégration d'un modèle de dépendance dans le calcul des statistiques de test. Par conséquent, le principe à l'origine de cette méthodologie remet en cause la transposition de l'optimalité de la théorie univariée des tests d'hypothèses (Neymann-Pearson) dans le cas des tests multiples.

Dans cette procédure, la propriété d'indépendance conditionnelle entre les statistiques de test permet d'étendre au cas de dépendances générales les résultats en matière de contrôle de taux d'erreurs de la méthodologie développée initialement sous l'hypothèse de l'indépendance. Ainsi, le cadre proposé procure à la fois une augmentation de la puissance des tests et une meilleure stabilité des procédures d'inférence simultanée.

D'autre part, dans un contexte de grande dimension, la représentation de la matrice de variance-covariance comme en Analyse en Facteurs, à travers un espace linéaire de dimension restreinte, peut être comparée à d'autres d'approches comme l'estimation par pénalisation conduisant à un rétrécissement des coefficients [Schäfer and Strimmer, 2005]. Cette propriété s'avère alors être particulièrement intéressante pour l'estimation de la matrice des corrélations partielles.

Enfin, au delà des considérations sur le modèle lui même, nous avons évoqué de nombreux points concernant la mise en œuvre de cette méthode en pratique. En particulier, l'estimation des paramètres du modèle par un algorithme EM, l'estimation de la proportion d'hypothèses nulles et le choix du nombre de facteurs ont fait l'objet d'études approfondies. L'algorithme EMFA permet d'obtenir de bonnes estimations des paramètres de variances par maximisation de la vraisemblance du modèle, y compris en situation de grande dimension (contexte non asymptotique). De plus, nous proposons un critère pour déterminer le nombre de facteurs à inclure dans le modèle par minimisation de la variance du nombre de faux-positifs, ce qui garantit une meilleure stabilité pour les procédures de tests multiples. Cependant, certaines questions restent à explorer, mais qui sont essentiellement d'ordre numérique. En particulier, l'estimation de $\mathcal{M}_0$ à l'étape 1 de la procédure `FAMT` et son impact sur l'estimation des scores sont des questions posées récemment également par Efron [2007] et par Leek and Storey [2007].

Finalement, la méthode a été appliquée à des données issues de biopuces, dans le cadre d'études visant à l'identification de gènes différentiellement exprimés en génomique et a montré un grand intérêt en terme d'interprétation biologique [Blum et al., 2010]. La représentation des interactions directes entre gènes dans les réseaux biologiques est une problématique qui s'inscrit dans le prolongement des analyses différentielles menées sur les données génomiques. Cette représentation sous forme de graphe, où un nœud correspond à un gène et une arrête correspond à une relation de dépendance, est basée sur la matrice des corrélations partielles. Des premiers travaux à partir de simulations donnent d'ailleurs déjà des résultats prometteurs en terme d'estimation de cette matrice à partir d'un modèle d'Analyse en Facteurs (42èmes Journées de la SFdS, 2010) et certains résultats analytiques d'articles récents [Fan et al., 2008, Ambroise et al., 2009] nous encouragent à approfondir cette approche.

La méthode est également mise à disposition des utilisateurs à travers un module informatique dans

le logiciel libre et gratuit `R`, facilitant son transfert et sa diffusion vers son domaine d'application privilégié. Le package `FAMT` [Causeur et al., 2010] est disponible sur le site de développement de `R` (`http://www.r-project.org/`) et un site internet (`http://famt.free.fr/`) présente le principe de la méthode et l'utilisation des fonctions associées à sa mise en œuvre.

ANNEXE

A. Simulations : code R

Exemples 3, 4 et 6

On fixe tout d'abord le nombre de variables m et le nombre de facteurs communs Q. On simule alors une matrice des loadings (B) et une matrice de variance spécifique (Ψ) pour chaque scénario, correspondant à différents niveaux de dépendance.

```
#### Simulation d'une matrice B et d'une matrice Psi par scénario
# m: nombre de variables
# Q: nombre de facteurs communs dans le modèle
#
i=1 # numéro du scénario: de 0 à 9
shrink = i/10
B = shrink*rmvnorm(m,mean=rep(0,Q),sigma=diag(Q))
Sigma = apply(B^2,1,sum)+ rep(1,m)
B = diag(1/sqrt(Sigma))%*%B
Psi = diag(1/Sigma)
```

Pour chaque scénario, on dispose d'une matrice de loadings B et d'une matrice de variance spécifique Ψ. Le coefficient `shrink` permet de donner plus ou moins de poids à la structure commune. On utilise ensuite la fonction suivante pour simuler $nbsim = 1000$ jeux de données par scénario. Pour les EXEMPLES 3 et 6, δ est fixé à 0. Pour l'EXEMPLE 4, δ est fixé en utilisant la fonction `power.t.test` de R. On fixe également la taille d'échantillon souhaitée (n_A et n_B).

```
#### Simulation d'un jeu de données à partir de ces matrices B et Psi
# nA, nB: nombre d'observations pour les groupes A et B
# m0: nombre de variables sous H0
# m1: nombre de variables sous H1
# B: matrice des loadings
# Psi: matrice des variances spécifiques
# delta: différence de moyenne entre les observations des groupes A et B,
#      calculé en pratique à partir de la fonction power.t.test
# @return donnees: jeu de données simulé; scores: facteurs communs Z
simuldata = function(nA,nB,m0,m1,B,Psi,delta) {
n=nA+nB
  Q = ncol(B)
  specific = rmvnorm(n,sigma=Psi) # facteurs spécifiques
  scores = rmvnorm(n,sigma=diag(Q)) # facteurs communs
  donnees = scores%*%t(B)+specific
  donnees[(nA+1):n,(m0+1):(m0+m1)] = donnees[(nA+1):n,(m0+1):(m0+m1)] + delta
  list(donnees=donnees,scores=scores)
}
```

Exemple 5

```
# mY, sY: moyenne et écart-type de la variable Y
# mZ, sZ: moyenne et écart-type de la variable Z
# rho: corrélation entre Y et Z
# delta: différence de moyenne entre les observations des groupes A et B
#   pour Y, calculé en pratique à partir de la fonction power.t.test
# nA, nB: nombre d'observations pour les groupes A et B
# @return donnees: tableau de données composé de 2 colonnes et de n=nA+nB lignes
simul2var= function(mY,sY,mZ,sZ,rho,delta,nA,nB){
  n=nA+nB
  EZ = rep(mZ,n)
  EY = c(rep(mY,nA),rep(mY+delta,nB))
  matsigma = matrix(c(sY^2,sY*sZ*rho,sY*sZ*rho,sZ^2),nrow=2,ncol=2)
  donnees = data.frame(c(EY,EZ)+ rmvnorm(n=n, mean=c(0,0), sigma= matsimga))
  colnames(donnees)=c("Y","Z")
  return(donnees=donnees)
}
```

On simule alors $nbsim = 10000$ tableaux de données pour chaque valeur de ρ allant de $\rho_{min} = 0$ à $\rho_{max} = 0.9$.

B. Algorithme EMFA

Cette annexe présente la construction des équations de mise-à-jour des étapes E et M de l'algorithme EMFA présenté en SECTION 4.2 qui permet d'obtenir les Estimateurs du Maximum de Vraisemblance des paramètres du modèle d'Analyse en Facteurs [Rubin and Thayer, 1982] en grande dimension.

Le principe est décrit ci-après, où $Y^{(i)}$ représente la ligne de la matrice des données Y correspondant à l'individu i, de dimension $1 \times m$. De même, $Z^{(i)}$ représente l'ensemble des scores associés à l'individu i, vecteur de dimension $1 \times Q$.

1. **On calcule la log-vraisemblance du modèle et son espérance**

$$\begin{aligned}
\mathcal{L}(B,\Psi) &= \sum_{i=1}^{n} ln\left\{(2\pi)^{-m/2}\left|\Psi\right|^{-1/2} exp\left[-\frac{1}{2}(Y^{(i)}-Z^{(i)}B')\Psi^{-1}(Y^{(i)}-Z^{(i)}B')'\right]\right\} \\
&= -\frac{nm}{2}ln(2\pi)+\frac{n}{2}ln\left|\Psi^{-1}\right|-\frac{1}{2}\sum_{i=1}^{n}\left[(Y^{(i)}-Z^{(i)}B')\Psi^{-1}(Y^{(i)}-Z^{(i)}B')'\right] \\
&= -\frac{nm}{2}ln(2\pi)+\frac{n}{2}ln\left|\Psi^{-1}\right|-\frac{1}{2}\sum_{i=1}^{n}\Big[Y^{(i)}\Psi^{-1}Y'^{(i)}-Z^{(i)}B'\Psi^{-1}Y'^{(i)} \\
&- Y^{(i)}\Psi^{-1}BZ'^{(i)}+Z^{(i)}B'\Psi^{-1}BZ'^{(i)}\Big] \\
&= -\frac{nm}{2}ln(2\pi)+\frac{n}{2}ln\left|\Psi^{-1}\right|-\frac{1}{2}\sum_{i=1}^{n}\left[Y^{(i)}\Psi^{-1}Y'^{(i)}-2Y^{(i)}\Psi^{-1}BZ'^{(i)}+Z^{(i)}B'\Psi^{-1}BZ'^{(i)}\right] \\
&= -\frac{nm}{2}ln(2\pi)+\frac{n}{2}ln\left|\Psi^{-1}\right|-\frac{1}{2}\sum_{i=1}^{n}\left[Y^{(i)}\Psi^{-1}Y'^{(i)}-2Y^{(i)}\Psi^{-1}BZ'^{(i)}+tr\left(B'\Psi^{-1}BZ'^{(i)}Z^{(i)}\right)\right] \quad (B.1)
\end{aligned}$$

En utilisant : $Z^{(i)}B'\Psi^{-1}Y'^{(i)} = (Z^{(i)}B'\Psi^{-1}Y'^{(i)})' = Y^{(i)}\Psi^{-1}BZ'^{(i)}$ puis $x'Ax = tr(Axx')$ avec $A = B'\Psi^{-1}B$ et $x = Z'^{(i)}$.

On en déduit l'espérance de la log-vraisemblance :

$$\begin{aligned}
\mathbb{E}(\mathcal{L}|Y) = cst+\frac{n}{2}ln\left|\Psi^{-1}\right| &- \frac{1}{2}\sum_{i=1}^{n}\Big[Y^{(i)}\Psi^{-1}Y'^{(i)}-2Y^{(i)}\Psi^{-1}B\mathbb{E}(Z'^{(i)}|Y^{(i)}) \\
&+ tr\left(B'\Psi^{-1}B\mathbb{E}(Z'^{(i)}Z^{(i)}|Y^{(i)})\right)\Big] \quad (B.2)
\end{aligned}$$

2. **On maximise l'espérance de la log-vraisemblance**
   – Maximisation par rapport à B : $\frac{\partial\mathbb{E}(\mathcal{L}|Y)}{\partial B}=0$

$$\begin{aligned}
\frac{\partial\mathbb{E}(\mathcal{L}|Y)}{\partial B} &= \frac{1}{2}\sum_{i=1}^{n}\left[-2\Psi^{-1}Y'^{(i)}\mathbb{E}(Z'^{(i)}|Y^{(i)})'-\Psi^{-1}B\mathbb{E}(Z'^{(i)}Z^{(i)}|Y^{(i)})-\Psi^{-1}B\mathbb{E}(Z'^{(i)}Z^{(i)}|Y^{(i)})'\right]=0 \\
&\Leftrightarrow \sum_{i=1}^{n}\left[\Psi^{-1}Y'^{(i)}\mathbb{E}(Z'^{(i)}|Y^{(i)})'\right]=\sum_{i=1}^{n}\left[\Psi^{-1}B\mathbb{E}(Z'^{(i)}Z^{(i)}|Y^{(i)})\right] \\
&\Leftrightarrow \sum_{i=1}^{n}\left[Y'^{(i)}\mathbb{E}(Z'^{(i)}|Y^{(i)})'\right]=\sum_{i=1}^{n}\left[B\mathbb{E}(Z'^{(i)}Z^{(i)}|Y^{(i)})\right] \\
&\Leftrightarrow B=\sum_{i=1}^{n}\left[Y'^{(i)}\mathbb{E}(Z^{(i)}|Y^{(i)})\right]\left[\sum_{i=1}^{n}\mathbb{E}(Z'^{(i)}Z^{(i)}|Y^{(i)})\right]^{-1} \quad (B.3)
\end{aligned}$$

On utilise $\frac{\partial A'XB}{\partial X}=AB'$ et $\frac{\partial tr(X'AXB)}{\partial X}=AXB+A'XB'$.

– Maximisation par rapport à Ψ^{-1} : $\frac{\partial \mathbb{E}(\mathcal{L}|Y)}{\partial \Psi^{-1}} = 0$

$$\begin{aligned}
\frac{\partial \mathbb{E}(\mathcal{L}|Y)}{\partial \Psi^{-1}} &= \frac{n}{2}\Psi - \frac{1}{2}\sum_{i=1}^{n}\left[Y'^{(i)}Y^{(i)} - 2Y'^{(i)}\mathbb{E}(Z'^{(i)}|Y^{(i)})'B' + B\mathbb{E}(Z'^{(i)}Z^{(i)}|Y^{(i)})'B'\right] \\
&= \frac{n}{2}\Psi - \frac{1}{2}\sum_{i=1}^{n}Y'^{(i)}Y^{(i)} + \sum_{i=1}^{n}Y'^{(i)}\mathbb{E}(Z'^{(i)}|Y^{(i)})'B' - \frac{1}{2}\sum_{i=1}^{n}B\mathbb{E}(Z'^{(i)}Z^{(i)}|Y^{(i)})'B' = 0 \\
\Leftrightarrow \quad \Psi &= \frac{2}{n}\left[\frac{1}{2}\sum_{i=1}^{n}Y'^{(i)}Y^{(i)} + \frac{1}{2}\sum_{i=1}^{n}B\mathbb{E}(Z'^{(i)}Z^{(i)}|Y^{(i)})'B' - \sum_{i=1}^{n}Y'^{(i)}\mathbb{E}(Z'^{(i)}|Y^{(i)})'B'\right]
\end{aligned}$$

On peut ensuite insérer l'expression obtenue en (B.3) pour "simplifier" cette équation :

$$\begin{aligned}
\Psi &= \frac{2}{n}\left[\frac{1}{2}\sum_{i=1}^{n}Y'^{(i)}Y^{(i)} + \frac{1}{2}\sum_{i=1}^{n}B\mathbb{E}(Z'^{(i)}Z^{(i)}|Y^{(i)})'B' - \sum_{i=1}^{n}Y'^{(i)}\mathbb{E}(Z'^{(i)}|Y^{(i)})'B'\right] \\
&= \frac{1}{n}\sum_{i=1}^{n}Y'^{(i)}Y^{(i)} + \frac{1}{n}\left\{\left[\sum_{i=1}^{n}Y'^{(i)}\mathbb{E}(Z'^{(i)}|Y^{(i)})'\left[\sum_{i=1}^{n}\mathbb{E}(Z'^{(i)}Z^{(i)}|Y^{(i)})\right]^{-1}\right]\sum_{i=1}^{n}\mathbb{E}(Z'^{(i)}Z^{(i)}|Y^{(i)})'B'\right. \\
&\quad \left. - 2\sum_{i=1}^{n}Y'^{(i)}\mathbb{E}(Z'^{(i)}|Y^{(i)})'B'\right\} \\
&= \frac{1}{n}\sum_{i=1}^{n}Y'^{(i)}Y^{(i)} - \frac{1}{n}\sum_{i=1}^{n}Y'^{(i)}\mathbb{E}(Z'^{(i)}|Y^{(i)})'B' \\
&= S - \frac{1}{n}\sum_{i=1}^{n}Y'^{(i)}\mathbb{E}(Z^{(i)}|Y^{(i)})B' \qquad \text{(B.4)}
\end{aligned}$$

On prend la diagonale de la matrice obtenue en (B.4) comme estimation de Ψ.

3. **Il reste alors à calculer $\mathbb{E}(Z'^{(i)}|Y^{(i)})$ et $\mathbb{E}(Z'^{(i)}Z^{(i)}|Y^{(i)})$.** Pour cela, on va d'abord s'intéresser à la densité de $Z|Y$, notée $f(z|y)$. On a :

$$\begin{aligned}
f(z|y) &= \frac{f(z,y)}{f(y)} \\
&= \frac{(2\pi)^{-(m+Q)/2}|C|^{-1/2}exp\left\{-\frac{1}{2}(y,z)C^{-1}(y,z)'\right\}}{(2\pi)^{-m/2}|\Sigma|^{-1/2}exp\left\{-\frac{1}{2}y\Sigma^{-1}y'\right\}} \\
&= (2\pi)^{-Q/2}\frac{|C|^{-1/2}}{|\Sigma|^{-1/2}}exp\left\{-\frac{1}{2}\left[(y,z)C^{-1}(y,z)' - y\Sigma^{-1}y'\right]\right\} \\
&= (2\pi)^{-Q/2}\frac{|C|^{-1/2}}{|\Sigma|^{-1/2}}exp\left\{-\frac{1}{2}\theta\right\} \qquad \text{(B.5)}
\end{aligned}$$

Où on définit $C = \mathbb{V}(Y,Z) = \begin{bmatrix} \Sigma & B \\ B' & \mathbb{I}_Q \end{bmatrix}$. On montre alors que

$$C^{-1} = \begin{bmatrix} (C^{-1})_{11} & (C^{-1})_{12} \\ (C^{-1})_{21} & (C^{-1})_{22} \end{bmatrix} = \begin{bmatrix} \Psi^{-1} & \Sigma^{-1}B'(B'\Sigma^{-1}B - \mathbb{I}_Q)^{-1} \\ (B'\Sigma^{-1}B - \mathbb{I}_Q)^{-1}B\Sigma^{-1} & \mathbb{I}_Q + B'\Psi^{-1}B \end{bmatrix} \qquad \text{(B.6)}$$

On étudie maintenant l'expression θ de l'exponentielle de (B.5).

$$\begin{aligned}
\theta &= y(C^{-1})_{11}y' + y(C^{-1})_{12}z' + z(C^{-1})_{21}y' + z(C^{-1})_{22}z' - y\Sigma^{-1}y' \\
&= y[(C^{-1})_{11} - \Sigma^{-1}]y' + 2y(C^{-1})_{12}z' + z(C^{-1})_{22}z'
\end{aligned}$$

En effet, $(C^{-1})_{12} = (C^{-1})'_{21}$. De plus, en remarquant que $(C^{-1})_{11} - \Sigma^{-1} = \Psi^{-1}B(\mathbb{I}_Q + B'\Psi^{-1}B)^{-1}B'\Psi^{-1}$, on a :

$$\begin{aligned}\theta &= y[\Psi^{-1}B(\mathbb{I}_Q + B'\Psi^{-1}B)^{-1}B'\Psi^{-1}]y' + 2y[\Sigma^{-1}B'(B'\Sigma^{-1}B - \mathbb{I}_Q)^{-1}]z' + z[\mathbb{I}_Q + B'\Psi^{-1}B]z' \\ &= [z - y\Sigma^{-1}B][\mathbb{I}_Q + B'\Psi^{-1}B][z' - B'\Sigma^{-1}y'] \\ &= [z - y\Sigma^{-1}B][\mathbb{I}_Q + B'\Psi^{-1}B][z - y\Sigma^{-1}B]' \end{aligned} \tag{B.7}$$

Car $[B'\Sigma^{-1}B - \mathbb{I}_Q]^{-1} = -[\mathbb{I}_Q + B'\Psi^{-1}B]$. Ainsi, $f(z|y)$ peut être vue comme un mélange de gaussienne, d'espérance $y\Sigma^{-1}B$ et de variance $G = [\mathbb{I}_Q + B'\Psi^{-1}B]^{-1}$.
D'où l'expression de $\mathbb{E}(Z^{(i)}|Y^{(i)})$:

$$\begin{aligned}\mathbb{E}(Z^{(i)}|Y^{(i)}) &= Y^{(i)}\Sigma^{-1}B = Y^{(i)}[\Psi^{-1} - \Psi^{-1}B(\mathbb{I}_Q + B'\Psi^{-1}B)^{-1}B'\Psi^{-1}]B \\ &= Y^{(i)}[\Psi^{-1}B - \Psi^{-1}B(\mathbb{I}_Q + B'\Psi^{-1}B)^{-1}B'\Psi^{-1}B] \\ &= Y^{(i)}\Psi^{-1}B[\mathbb{I}_Q - (\mathbb{I}_Q + B'\Psi^{-1}B)^{-1}B'\Psi^{-1}B] \\ &= Y^{(i)}\Psi^{-1}B[\mathbb{I}_Q - (\mathbb{I}_Q + B'\Psi^{-1}B)(\mathbb{I}_Q + B'\Psi^{-1}B)^{-1} + (\mathbb{I}_Q + B'\Psi^{-1}B)^{-1}] \\ &= Y^{(i)}\Psi^{-1}B(\mathbb{I}_Q + B'\Psi^{-1}B)^{-1} \end{aligned} \tag{B.8}$$

D'autre part, on a :

$$\begin{aligned}\mathbb{E}(Z'^{(i)}Z^{(i)}|Y^{(i)}) &= \mathbb{V}(Z^{(i)}|Y^{(i)}) + \mathbb{E}(Z^{(i)}|Y^{(i)})'\mathbb{E}(Z^{(i)}|Y^{(i)}) \\ &= [(C^{-1})_{22}]^{-1} + \mathbb{E}(Z^{(i)}|Y^{(i)})'\mathbb{E}(Z^{(i)}|Y^{(i)}) \\ &= (\mathbb{I}_Q + B'\Psi^{-1}B)^{-1} + \mathbb{E}(Z^{(i)}|Y^{(i)})'\mathbb{E}(Z^{(i)}|Y^{(i)}) \end{aligned} \tag{B.9}$$

L'algorithme EMFA est donc :

1. On part des données centrées $Y = Y_c$
2. **Initialisation** : On dispose d'une première estimation pour B et Ψ (B_0 et Ψ_0)
3. **Itération** : On va ensuite alterner des étapes d'estimation de l'espérance de la vraisemblance et des étapes de maximisation, à partir des valeurs B_0 et Ψ_0 pour trouver les nouvelles estimations B_1 et Ψ_1 :
   (a) **étape E** : On estime l'espérance de la vraisemblance du modèle.

$$\begin{aligned}\mathbb{E}(Z^{(i)}|Y^{(i)}) &= Y^{(i)}\Psi_0^{-1}B_0(\mathbb{I}_Q + B_0'\Psi_0^{-1}B_0)^{-1} \\ \mathbb{E}(Z^{(i)}Z'^{(i)}|Y^{(i)}) &= (\mathbb{I}_Q + B_0'\Psi_0^{-1}B_0)^{-1} + \mathbb{E}(Z^{(i)}|Y^{(i)})'\mathbb{E}(Z^{(i)}|Y^{(i)}) \end{aligned}$$

   D'après (B.8) et (B.9).
   (b) **étape M** : On définit alors les nouvelles estimations des paramètres (EMV).

$$\begin{aligned}B_1 &= \sum_{i=1}^{n}\left[Y'^{(i)}\mathbb{E}(Z^{(i)}|Y^{(i)})\right]\left[\sum_{i=1}^{n}\mathbb{E}(Z'^{(i)}Z^{(i)}|Y^{(i)})\right]^{-1} \\ \Psi_1 &= diag\left[S - \frac{1}{n}\sum_{i=1}^{n}Y'^{(i)}\mathbb{E}(Z^{(i)}|Y^{(i)})B_1'\right] \end{aligned}$$

   D'après les équations de mise-à-jour définies en B.3) et B.4.

4. **Critère d'arrêt** : on itère les étapes 3a et 3b en prenant les estimations de B_1 et Ψ_1 obtenues avec une étape E et une étape M comme point de départ de l'itération suivante. On itère jusqu'à ce que deux estimations successives de Ψ et B soient identiques, à un seuil ϵ près. Le critère considéré peut être la trace de l'écart quadratique entre les matrices Ψ, obtenues lors de 2 étapes successives : $\frac{1}{m} tr\left(\Psi_1 - \Psi_0\right)^2$.

C. Méthode SVA

Nous décrivons dans cette annexe une méthodologie proposée par J. Storey et J. Leek appelée SVA (pour Surrogate Variable Analysis) [Leek and Storey, 2007, 2008] et programmée dans R dans le package `sva` [Leek, 2008].

Le but de cette méthode est d'identifier et construire des variables dites de substitution (*surrogate variables*) au sein d'un jeu de données en grande dimension afin de modéliser la structure de dépendance au sein de ce jeu de données.

Les données d'expressions géniques, domaine d'application pour lequel cette méthode a été initialement développée, sont affectées par un certain nombre de facteurs environnementaux, biologiques ou encore techniques, non-observés (ou non observables) qui induisent de la variabilité dans les mesures (appelée *expression heterogeneity* par les auteurs). Les variables de substitution ont pour fonction de modéliser cette variabilité. Ces variables, construites à partir des variables observées, vont ensuite pouvoir être utilisées dans d'autres analyses, comme l'analyse différentielle, avec comme objectif de surmonter les problèmes dus à la dépendance.

Les vecteurs Z du modèle (4.2) sont les "surrogate variables" que la méthode SVA va estimer par décomposition en valeurs singulières (SVD).

Les L facteurs sont des sources de variabilité autres que x, mais ils ont aussi la possibilité d'être confondus avec x. L'algorithme SVA composé de deux étapes :

A **Choix du nombre de facteurs** (basé sur l'analyse parallèle [Buja and Eyuboglu, 1992])

1 Estimation de m_k en ajustant le modèle (1.1) et calcul des résidus $r_k = y_k - \hat{m}_k(x)$: construction de la matrice $R = [r_1; \ldots; r_m]$.

2 Décomposition en valeurs singulières de R : $R = UDV'$, avec $D = diag(d_1; \ldots; d_{n-ddl}; 0 \ldots; 0)$.

3 Construction $T_i = \frac{d_i^2}{\sum_i d_i^2}$, pour $i \in [1; n - ddl]$, qui correspond à la variance expliquée par le $i$ème vecteur propre.

4 Détermination par permutations de la distribution nulle de T_i et calcul des probabilités critiques associées : détermination du nombre Q de vecteurs propres significatifs (seuil α donné).

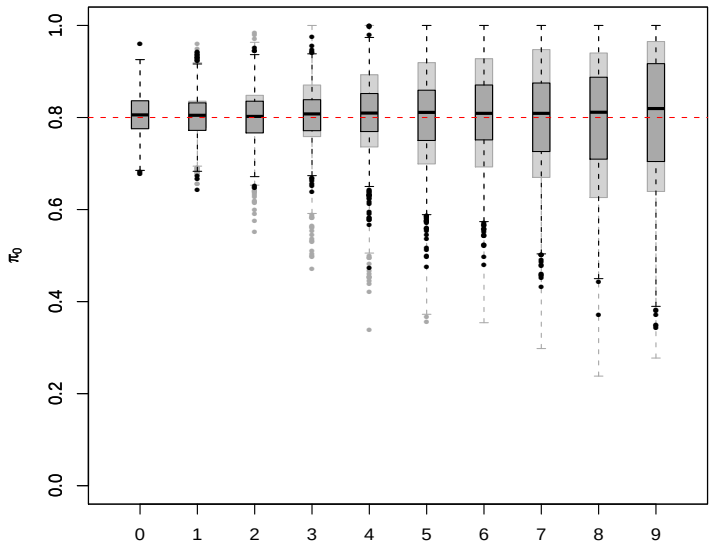
B **Estimation des facteurs** (*surrogate variables*)

1 Estimation de m_k en ajustant le modèle (1.1) et calcul des résidus $r_k = y_k - \hat{m}_k(x)$: construction de la matrice $R = [r_1; \ldots; r_m]$.

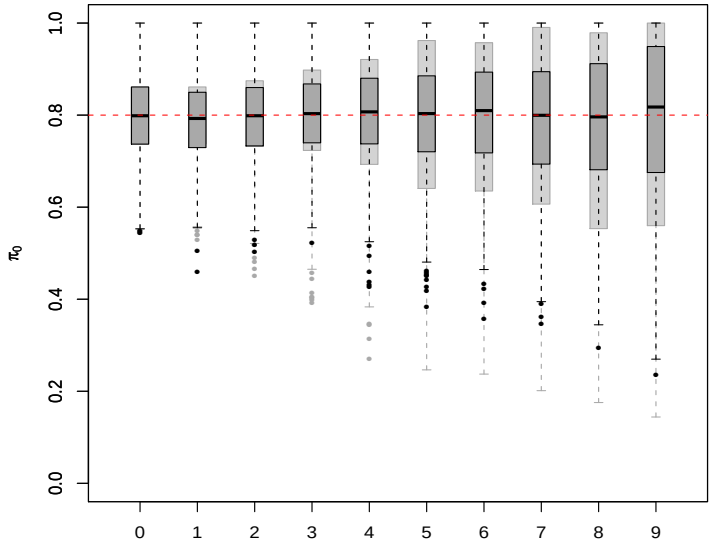
2 Décomposition en valeurs singulières de R : $R = UDV'$, avec $V = [v_1; \ldots; v_n]$. V représente la variabilité indépendante de x. On retient les Q premiers vecteurs de V, où Q a été déterminé à l'étape (A).

3 Pour chacun des Q vecteurs v_q :
- Calcul des probabilités critiques des tests du coefficient de régression entre v_q et $Y = [Y_1; \ldots; Y_m]$. On mesure ainsi la force du lien entre v_q et chacune des variables d'intérêt Y_k.
- Estimation de la proportion p_0/m de variables Y_k dont le lien avec v_q n'est pas significatif [Storey and Tibshirani, 2003]. On note $\mathcal{S} = \{s_1; \ldots; s_{\hat{p}_1}\}$ l'ensemble des indices des $\hat{p}_1 = m - \hat{p}_0$) plus petites probabilités critiques. $\tilde{Y} = Y[\mathcal{S}]$: contient les variables supposées contenir la variabilité représentée par une certaine variable z_q.
- Décomposition en valeurs singulières de $\tilde{Y}$: $e = [e_1; \ldots; e_n]$ sont les vecteurs propres de cette matrice.
- Estimation de la "*surrogate variable*" z_q : $\hat{z}_q = e_{i*}$, avec $i* = \text{argmax}_{i \in [1;n]}\{cor(v_q; e_i)\}$

D. Figures supplémentaires

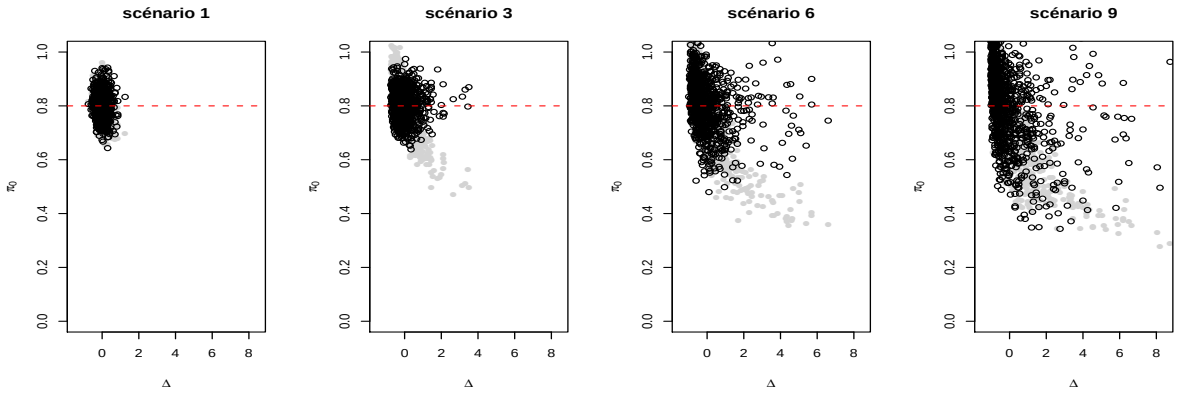


(a) Estimation de la densité par une méthode à noyau

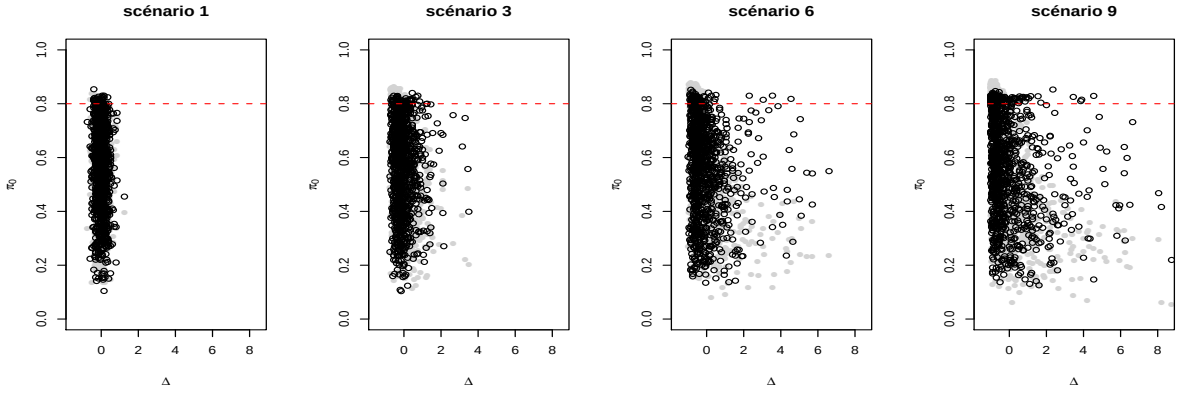


(b) Estimateur empirique avec choix de λ par lissage spline

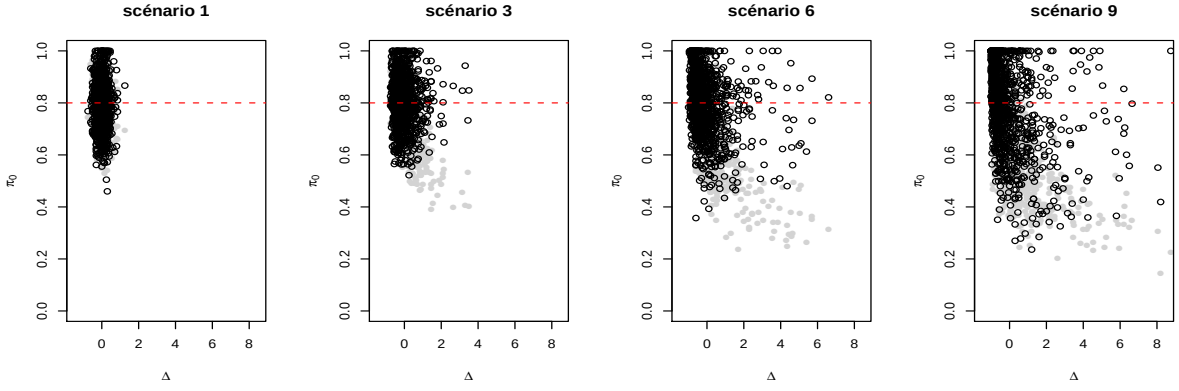
FIG. D.1 – Estimations de π_0 sur des données simulées selon différents scénarios de dépendance -$\pi_0 = 0,80$. Comparaison entre les probabilités critiques usuelles (tests de Student) : en gris, et les probabilités critiques ajustées : en noir.



(a) Estimation de la densité par un estimateur à noyau



(b) Estimation basée sur le plus long intervalle de la densité empirique



(c) Estimateur empirique avec choix de λ par lissage spline

FIG. D.2 – Estimations de π_0 à partir de probabilités critiques ajustées (en noir) sur des données indépendantes (scénario 1), modérément (scénario 3 et scénario 6) ou très corrélées (scénario 9) avec différentes méthodes (SECTION1.2) en fonction de la forme de l'histogramme de la distribution des probabilités critiques aux alentours de 0 - Mêmes méthodes d'estimations à partir des probabilités critiques usuelles en gris - $\pi_0 = 0,80$

BIBLIOGRAPHIE

D.B. Allison. A mixture model approach for the analysis of microarray gene expression data. *Computational Statistics and Data Analysis*, 39 :1–20, 2002.

C. Ambroise, J. Chiquet, and C. Matias. Inferring sparse gaussian graphical models with latent structure. *Electronic Journal of Statistics*, 3 :205–238, 2009.

Y. Benjamini and Y. Hochberg. Controlling the false discovery rate : a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57 :289–300, 1995.

Y. Benjamini and D. Yekutieli. The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics*, 29 :1165–1188, 2001.

Y. Benjamini, A. Krieger, and D. Yekutieli. Adaptive linear step-up procedures that control the false discovery rate. *Biometrika*, 93 :491–507, 2006.

M. A. Black. A note on the adaptative control of false discovery rates. *Journal of the Royal Statistical Society. Series B*, 66 :297–304, 2004.

G. Blanchard and E. Roquain. Two simple sufficient conditions for FDR control. *Electronic journal of Statistics*, 2 :963–992, 2008.

Y. Blum, G. LeMignon, S. Lagarrigue, and D. Causeur. A factor model to analyze heterogeneity in gene expression. *BMC bioinformatics*, 11 :368, 2010.

C. E. Bonferroni. Teoria statistica delle classi e calcolo delle probabilità. *Pubblicazioni del R Istituto Superiore si Scienze Economiche e Comerciali di Firenze*, pages 3–62, 1936.

A. Buja and N. Eyuboglu. Remarks on parallel analysis. *Multivariate behaviour*, 27, 1992.

R. B. Cattell. The scree test for the number of factors. *Multivariate Behavioural Research*, 1 : 245–276, 1966.

R. B. Cattell. The scientific use of factor analysis in behavioral and life sciences. *New York : Plenum*, 1978.

D. Causeur, C. Friguet, M. Houée, and M. Kloareg. Factor analysis for multiple testing (famt) : an r package for large-scale significance testing under dependence. *Journal of Statistical Software, submitted*, 2010.

M. Chavent, V. Kuentz, and J. Saracco. Analyse en facteurs : présentation et comparaison des logiciels sas, spad et spss. *La revue MODULAD*, 37, 2007.

Alan Dabney, John D. Storey, and with assistance from Gregory R. Warnes. *qvalue : Q-value estimation for false discovery rate control*, 2009. URL `http://CRAN.R-project.org/package=qvalue`. R package version 1.20.0.

M. Davidian, A. Tsiatis, and S. Leon. Semiparametric estimation of treatment effect in a pretest-posttest study with missing data. *Statistical Science*, 3 :261–301, 2005.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society, Series B*, 34 :1–38, 1977.

S. Dudoit and M.J. VanDerLaan. *Multiple testing procedures with application to genomics*. 2008.

S. Dudoit, J. Fridlyand, and T.P. Speed. Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American Statistical Association*, 97 :77–87, 2002.

S. Dudoit, J. Shaffer, and J. C. Boldrick. Multiple hypothesis testing in microarray experiments. *Statistical Science*, 18 :71–103, 2003.

B. Efron. Large-scale simultaneous hypothesis testing : The choice of a null hypothesis. *Journal of the American Statistical Association*, 99 :96–104, 2004.

B. Efron. Correlation and large-scale simultaneous testing. *Journal of the American Statistical Association*, 102 :93–103, 2007.

B. Efron, R. Tibshirani, J.D. Storey, and V. Tusher. Empirical bayes analysis of a microarray experiment. *Journal of the American Statistical Association*, 96 :1151–1160, 2001.

Y. Escouffier. Le traitement des variables vectorielles. *Biometrics*, 29 :751–760, 1973.

L.R. Fabrigar, R. MacCallum, D.T. Wegener, and E.J. Strahan. Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods*, 4 :272–299, 1999.

J. Fan, Y. Fan, and J. Lv. High dimensional covariance matrix estimation using a factor model. *Journal of Econometrics*, 147 :186–197, 2008.

J.K. Ford, R. MacCallum, and M. Tait. The application of exploratory factor analysis in applied psychology : a critical review and analysis. *Personnel Psychology*, 39 :291–314, 1986.

C. Genovese and L. Wasserman. Operarting characteristics and extensions of the false discovery rate procedure. *Journal of the Royal Statistical Society. Series B*, 64 :499–517, 2002.

T.R. Golub, D.K. Slonim, C. Tamayo, P.and Huard, M. Gaasenbeek, J.P. Mesirov, H. Coller, M.L. Loh, J.R. Downing, M.A. Caligiuri, C.D. Bloomfield, and E.S. Lander. Molecular classification of cancer : class discovery and class prediction by gene expression monitoring. *Science*, 286 :531–537, 1999.

M. Guedj, G. Nuel, S. Robin, and A. Celisse. *kerfdr : semi-parametric kernel-based approach to local* FDR *estimations*, 2007. URL `http://stat.genopole.cnrs.fr/sg/software/kerfdr`. R package version 1.0.1.

T.J. Hastie and R. J. Tibshirani. *Generalized Additive Models*. 1990.

J.C. Hayton, D.G. Allen, and V. Scarpello. Factor retention in exploratory factor analysis : a tutorial on parallel analysis. *Organisational research Methods*, 7 :191–205, 2004.

I. Hedenfalk, D. Duggan, Y. D. Chen, M. Radmacher, M. Bittner, R. Simon, P. Meltzer, B. Gusterson, M. Esteller, O. P. Kallioniemi, B. Wilfond, A. Borg, and J. Trent. Gene expression profiles in hereditary breast cancer. *New England Journal of Medicine*, 344 :539–548, 2001.

S. Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6 :65–70, 1979.

J. Hsu and T. Chang, J. ans Wang. Simultaneous confidence intervals for differential gene expressions. *Journal of Statistical Planning and Inference*, 136 :2182–2196, 2006.

J.C. Hsu. The factor analytic approach to simultaneous inference in the general linear model. *Journal of computational and graphical statistics*, 1 :151–168, 1992.

J.C. Hsu and B. Nelson. Multiple comparisons in the general linear model. *Journal of computational and graphical statistics*, 7 :23–41, 1998.

J. Josse, J. Pagès, and F. Husson. Testing the significance of the RV coefficient. *Computational Statististics and Data Analysis*, 53 :82–91, 2008.

K. G. Jöreskog. Factor analysis by least square and maximum likelihood methods. *In K. Enslein, A. Ralston, & H. S. Wilf (Eds.) - Statistical methods for digital computers*, 3 :125–165, 1977.

H. F. Kaiser. The varimax criterion for analytic rotation in factor analysis. *Psychometrika*, 23 : 187–200, 1958.

H. F. Kaiser. The application of electronic computers to factor analysis. *Educationnal and Psychological Measurment*, 20 :141–151, 1960.

C. Kendziorski, H. Newton, M.and Lan, and M. Gould. On parametric empirical bayes methods for comparing multiple groups using replicated gene expression profiles. *Statistics in Medicine*, 22 : 3899–3914, 2003.

K. I. Kim and M. Van de Wiel. Effects of dependence in high-dimensional multiple testing problems. *BMC Bioinformatics*, 9, 2008.

M. Kloareg and D. Causeur. *Improving type* II *error rates of multiple tests by use of auxiliary variables and application to microarray data.* Ed. Christos Skiadas, World Scientific Publishing, Co Pte Ldt., 2007.

E.L. Korn, J.F. Troendle, L.M. McShane, and R. Simon. Controlling the number of false discoveries : application to high-dimensional genomic data. *Journal of Statistical Planning and Inference*, 124 : 379–398, 2004.

R. Kustra, R. Shioda, and M. Zhu. A factor analysis model for functional genomics. *BMC Bioinformatics*, 7, 2006.

C. E. Lance, M. M. Butts, and L. C. Michels. The sources of four commonly reported cutoff criteria : what did they really say ? *Organizational Research Methods*, 9 :202–220, 2006.

M. Langaas, B. H. Lindqvist, and E. Ferkingstad. Estimating the proportion of true null hypotheses, with application to dna microarray data. *Journal of the Royal Statistical Society. Series B*, 67 : 555–572, 2005.

D.N. Lawley. The estimation of factor loadings by the method of maximum likelihood. *Proceedings of the Royal Society of Edinburgh (A)*, 60 :64–82, 1940.

J. T. Leek and J. Storey. Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genetics*, 3(9) :e161, 2007.

J. T. Leek and J. Storey. A general framework for multiple testing dependence. *Proceedings of the National Academy of Sciences*, 105 :18718–18723, 2008.

Jeffrey T. Leek. *sva : Surrogate Variable Analysis*, 2008. R package version 1.1.0.

G. LeMignon, C. Désert, F. Pite, S. Leroux, O. Demeure, G. Guernec, B. Abasht, M. Douaire, P. LeRoy, and S. Lagarrigue. Using transcriptome profiling to characterize qtl regions on chicken chromosome 5. *BMC Genomics*, pages 10–575, 2009.

S. Leon, A. Tsiatis, and M. Davidian. Semiparametric estimation of treatment effect in a pretest-posttest study. *Biometrics*, 59 :1046–1055, 2003.

C. Liu and D. Rubin. Maximum likelihood estimation of factor analysis using the ecme algorithm with complete and incomplete data. *Statistica Sinica*, 8 :729–747, 1998.

I. Lönnstedt and T.P. Speed. Replicated microarray data. *Statistica Sinica*, 12 :31–46, 2002.

R. MacCallum, K. Widaman, S. Zhang, and S. Hong. Sample size in factor analysis. *Psychological Methods*, 4 :84–99, 2008.

M. Man, X. Wang, and Y. Wang. Power-sage : comparing statistical tests for sage experiments. *Bioinformatics*, 16 :953–959, 2000.

K. V. Mardia, J. T. Kent, and J. M. Bibby. *Multivariate Analysis*. 1979.

G.J. McLachlan, D. Peel, and R.W. Bean. Modelling high-dimensional data by mixtures of factor analysers. *Computational Statistics and Data Analysis*, 41 :379–388, 2003.

G.J. McLachlan, T. Krishnan, and S.K. NG. The EM algorithm. http ://www.econstor.eu/bitstream/10419/22198/1/24_tk_gm_skn.pdf, 2004. Humboldt-Universitat Berlin, Center for Applied Statistics and Economics.

G.J. McLachlan, R.W. Bean, and L.B. Jones. A simple implementation of a normal mixture approach to differential gene expression in multiclass microarrays. *Bioinformatics*, 22 :1608–0615, 2006.

R. G. Montanelli and L. G. Humphrey. Latent roots of ranrom data correlatoin matrices with squared multiple correlations on the diagonal : a monte-carlo study. *Psychometrica*, 41 :341–348, 1976.

M. Norris and L. Lecavalier. Evaluating the use of exploratory factor analysis in developmental disability psychological research. *Journal of Autism and Developpement Disorders*, 2009.

A.B. Owen. Variance of the number of false discoveries. *Journal of the Royal Statistical Society. Series B*, 67 :411–426, 2005.

K. Pollard, Y. Ge, S. Taylor, and S. Dudoit. *multtest : Resampling-based multiple hypothesis testing*. R package version 1.23.3.

I. Pournara and L. Wernisch. Factor analysis for gene regulatory networks and transcription factor activity profiles. *BMC Bioinformatics*, 8 :61, 2007.

K. Preacher and R. MacCallum. Exploratory factor analysis in behavior genetics research : Factor recovery with small sample sizes. *Behavior Genetics*, 32 :153–161, 2002.

X. Qiu, A. Brooks, L. Klebaniv, and A. Yakovlev. The effects of normalisation on the correlation structure of microarray data. *BMC Bioinformatics*, 6 :120, 2005.

W. Revelle and T. Rocklin. Very simple structure : an alternative procedure for estimating the opimal number of interprétable factors. *Multivariate Behavioural Research*, 14 :403–414, 1979.

D. Robertson and J. Symons. Maximum likelyhood factor analysis with rank-deficient sample covariance matrix. *Journal of Multivariate Analysis*, 98 :813–828, 2007.

S. Robin, A. Bar-Hen, J-J. Daudin, and L. Pierre. A semi-parametric approach for mixture models : Application to local false discovery rate estimation". *Computational Statistics & Data Analysis*, 51 :5483 – 5493, 2007.

J. M. Robins, A. Rotnizky, and L. P. Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89 :846–866, 1994.

D. B. Rubin and D. T. Thayer. EM algorithms for ML factor analysis. *Psychometrika*, 47 :69–76, 1982.

D. Salsburg. The lady tasting tea : How statistics revolutionized science in the twentieth century. *ISBN 0-8050-7134-2*, 2002.

S. Sarkar. Two-stage stepup procedures controlling fdr. *Journal of Statistical Planning and Inference*, 138 :1072–1084, 2008.

S.K. Sarkar. Some results on false discovery rate in stepwise multiple testing procedures. *Annals of statistics*, 30 :239–257, 2002.

J. Schäfer and K. Strimmer. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical Applications in Genetics and Molecular Biology*, 4 :32, 2005.

T. Schweder and E. Spjotvoll. Plots of p-values to evaluate many tests simultaneously. *Biometrika*, 69 :493–502, 1982.

J. Shaffer. Multiple hypotheses testing : a review. *Annual review of psychology*, 46 :561–584, 1995.

Z. Sidak. Rectangular confidence regions for the means of multivariate normal distributions. *Journal of the American Statistical Association*, 62 :626–633, 1967.

B. W. Silverman. *Density estimation for statistics and data analysis*. 1986.

R.J. Simes. An improved bonferroni procedure for multiple tests of significance. *Biometrika*, 73 : 751–754, 1986.

G. Smyth. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology*, 3 :article 3, 2004.

C. Spearman. General intelligence, objectively determined and measured. *American Journal of Psychology*, 15 :201–293, 1904.

D.W. Stewart. The application and misapplication of factor analysis in marketing research. *Journal of Marketing Research*, 18 :51–62, 1981.

J. Storey and R Tibshirani. Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences*, 100 :9440–9445, 2003.

J. D. Storey. A direct approach to false discovery rates. *Journal of the Royal Statistical Society. Series B*, 64 :479–498, 2002.

J. D. Storey. The positive false discovery rate : a bayesian interpretation and the q -value. *Annals of Statistics*, 31 :2013–2035, 2003.

J. D. Storey, J. E. Taylor, and D. Siegmund. Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates : a unified approach. *Journal of the Royal Statistical Society. Series B*, 66 :187–205, 2004.

J. D. Storey, J.Y. Dai, and J. T. Leek. The optimal discovery procedure for large-scale significance testing, with application to comparative microarray experiments. *Biostatistics*, 8 :414–432, 2007.

J.D. Storey. The optimal discovery procedure : A new approach to simultaneous significance testing. *Journal of the Royal Statistical Society. Series B*, 69 :347–368, 2007.

G.H. Thomson. *The Factorial Analysis of Human Ability*. 1951.

A. Tsiatis, M. Davidian, M. Zhang, and X. Lu. Covariate adjustment for a two-sample treatment comparison in randomized clinical trials : a principled yet flexible approach. *statistics in medecine*, 25 :1–10, 2000.

V. Tusher, R. Tibshirani, and G. Chu. Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences*, 98 :5116–5121, 2001.

W. F. Velicer, C. A. Eaton, and J. L. Fava. Construct explication through factor or component analysis : A review and evaluation of alternative procedures for determining the number of factors or components. *in R. D. Goffin and E. Helmes, eds.*, pages 41–71, 2000.

Gregory R. Warnes, Ben Bolker, Lodewijk Bonebakker, Robert Gentleman, Wolfgang Huber Andy Liaw, Thomas Lumley, Martin Maechler, Arni Magnusson, Steffen Moeller, Marc Schwartz, and Bill Venables. *gplots : Various R programming tools for plotting data*, 2009. URL `http://CRAN.R-project.org/package=gplots`. R package version 2.7.4.

P. H. Westfall and S. S. Young. *Resampling-based multiple testing : examples and methods for p-value adjustment*. 1993.

H. Yifan, H. Xu, V. Calian, and J.C. Hsu. To permute or not to permute. *Bioinformatics*, 22 : 2244–2248, 2006.

LISTE DES PUBLICATIONS ET COMMUNICATIONS

Les résumés des articles sont disponibles en fin de ce document ainsi que le tutoriel du package `R`*. Les résumés et présentations des communications orales sont disponibles sur ma page web :* `http://friguetchloe.wordpress.com`*.*

Articles dans des revues internationales

- **C. Friguet**, M. Kloareg & D. Causeur (2009) - A factor model approach to multiple testing under dependence, *Journal of the American Statistical Association* - 104 :488,1406-1415
  DOI : 10.1198/jasa.2009.tm08332
- D. Causeur, M. Kloareg. & C. Friguet (2009) - Control of the FWER in Multiple Testing Under Dependence, *Communications in Statistics - Theory and Methods* - 38 :16,2733-2747
  DOI : 10.1080/03610910902936281
- **C. Friguet** & D. Causeur (2010) - Estimation of the proportion of true null hypotheses in high-dimensional data under dependence, *Article soumis - CSDA, 05/2010*
- D. Causeur, C. Friguet, M. Houée & M. Kloareg - Factor Analysis for Multiple Testing (FAMT) : an R package for large-scale significance testing under dependence, *Article soumis - JSS, 06/2010*

Congrès de statistiques nationaux et internationaux (avec actes)

- **C. Friguet**, M. Kloareg & D. Causeur - Multiple tests for high-throughput data assuming a factor modeling of dependence
  40èmes Journées de Statistiques Société Française de Statistiques/Société Statistique Canada
  Ottawa, Canada, 2008
- D. Causeur, M. Kloareg & C. Friguet - Impact of dependence on the stability of model selection in supervised classification for high-throughput data
  International Indian Statistical Association (IISA) Conference "Frontiers of Probability and Statistical Science"
  Storrs, Connecticut, USA, 2008

- **C.Friguet** & D. Causeur - Estimation conditionnelle de la proportion d'hypothèses nulles en grande dimension
  41èmes Journées de Statistiques, Société Française de Statistiques
  Bordeaux, 2009
- **C. Friguet** & D. Causeur - Estimation of the proportion of null p-values among dependent tests
  Workshop on Simulation
  St Petersbourg, Russia, 2009
- Y. Blum, C. Friguet, S. Lagarrigue & D. Causeur - Inférence sur réseaux géniques par Analyse en Facteurs
  42èmes Journées de Statistiques, Société Française de Statistiques
  Marseille, 2010

Congrès de statistiques internationaux dédiés à l'analyse de données génomiques

- **C. Friguet**, M. Kloareg & D. Causeur - Accounting for a factor structure in high-dimensional data to improve multiple testing procedures
  6th workshop Statistical methods for post-genomic data
  Rennes, 2008
- **C. Friguet**, M. Kloareg & D. Causeur - Factor Analysis for Multiple Testing : A general approach for differential analysis of genome-scale dependent data
  Workshop Statistical advances in Genome-scale Data Analysis
  Ascona, Suisse, 2009 (poster)

Séminaires de groupe de travail

- M. Kloareg, C. Friguet, Y. Blum & D. Causeur - Factor Analysis for Multiple Testing : large scale significance testing under dependance
  EMBL
  Heidelberg, Allemagne, 2009
- **C. Friguet** - Impact de la dépendance en analyse différentielle en grande dimension
  Séminaire du groupe de travail en biostatistique
  Institut Elie Cartan, Nancy, 2010

Séminaires de jeunes chercheurs

- **C. Friguet** - La Statistique : un outil pour comprendre les données génomiques
  Doctoriales® de Bretagne
  Brest, 2008 (exposé + poster)
- **C. Friguet** - Approche conditionnelle des tests multiples pour données biologiques à haut-débit
  7ème Journée Jeunes Chercheurs en Biométrie (Société Française de Biométrie)
  INSERM, Villejuif, 2008

Valorisation sous forme de package R

- M. Kloareg, C. Friguet & D. Causeur - Factor Analysis for Multiple Testing : an R-package to analyze a genome-scale dataset
  Workshop Statistical advances in Genome-scale Data Analysis
  Ascona, Suisse, 2009 (poster)
- M. Kloareg, C. Friguet & D. Causeur - Factor Analysis for Multiple Testing (FAMT) : an R package for simultaneous tests under dependence in high-dimensional data
  UseR !2009, conférence des utilisateurs de R
  Agrocampus OUEST, Rennes, 2009
- Site web du package `FAMT` : `http://famt.free.fr`