Intégration de données complexes et hétérogènes à partir de tableaux de tailles différentes

Alyssa Imbert

Université de Toulouse

THÈSE

En vue de l'obtention du

DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE

Délivré par : *l'Université Toulouse 1 Capitole (UT1 Capitole)*

Présentée et soutenue le *19/10/2018* **par :**

Alyssa Imbert

Intégration de données hétérogènes complexes à partir de tableaux de tailles déséquilibrées

---

JURY

JULIE JOSSE	Professeur	Rapportrice
JEAN-FRANÇOIS LANDRIER	Directeur de Recherche	Examinateur
ANDREA RAU	Chargée de Recherche	Rapportrice
ANNE RUIZ-GAZEN	Professeur	Examinatrice
NATHALIE VIALANEIX	Chargée de Recherche	Direcrice de thèse
NATHALIE VIGUERIE	Chargée de Recherche	Codirectrice de thèse

École doctorale et spécialité :
MITT : Domaine Mathématiques : Mathématiques appliquées
Unité de Recherche :
Unité de Mathématiques et Informatique Appliquées de Toulouse, 875 INRA
Directeur(s) de Thèse :
Nathalie VIALANEIX et *Nathalie VIGUERIE*
Rapportrices :
Julie JOSSE et *Andrea RAU*

Remerciements

Je tiens tout d'abord à remercier mes directrices de thèse : Nathalie Vialaneix et Nathalie Viguerie. Je vous remercie pour la qualité de votre encadrement et vos conseils avisés. Merci Nathalie d'avoir toujours trouvé du temps (malgré tous tes déplacements), ta rigueur et la confiance que tu m'as accordée. Merci Nathalie pour m'avoir initiée à la biologie et appris à vulgariser mes propos pour les biologistes. Ce fut un plaisir de travailler avec vous.

Je remercie également la région Languedoc-Roussillon Midi-Pyrénées et la société Methodomics qui ont permis, grâce à leurs financements, l'accomplissement de ce travail.

Je remercie Julie Josse et Andrea Rau pour avoir accepté de consacrer une partie de leur temps à la lecture de ce manuscrit ainsi qu'à l'évaluation de mon travail. Je remercie également Jean-François Landrier et Anne Ruiz-Gazen pour avoir accepté de faire partie de mon jury de thèse. Je tiens à remercier les membres de mon comité de thèse : Avner Bar-Hen, Pierre-Antoine Gourraud, Laurence Liaubet, Joost-Peter Schanstra et Armand Valsesia pour leurs conseils enrichissants.

Je souhaite exprimer ma gratitude à Jorg et à toute son équipe pour m'avoir accueillie un mois dans leur département Santé Métabolique de Nestlé Health Science, à Lausanne. Grâce à eux, j'ai passé un agréable séjour, enrichissant, en Suisse. Je remercie plus particulièrement Armand pour son encadrement, ses conseils lors de ce séjour et tout au long de ma thèse.

Un grand merci à Mélina pour son aide sur l'inférence de réseau pour des données RNA-Seq et les scripts. Cela m'a beaucoup aidée. Merci également à Jérôme pour son template Latex pour la mise en forme de la thèse.

Merci à l'unité MIAT. Je tiens à remercier chaleureusement les gestionnaires (Fabienne, Nathalie et Alain) pour tout le travail qu'ils fournissent. J'ai pu grâce à cette thèse rencontrer des personnes formidables. Merci pour tous les bons moments passés pendant les pauses et/ou soirées à discuter sur des sujets sérieux (ou pas) : Sylvain, Damien et Damien, Charlotte, Sara, Franck, Etienne, Sébastian, Clémence, Manon, Lise, Léo, Malo, Nesrine, Fulya, Gaëlle, Lina, Faustine, Nathanaël, Floréal (merci également pour tes soirées jeux de société) et Léonard. Merci aux stagiaires pour votre bonne humeur. Je souhaite bon courage à ceux de l'unité qui débutent cette année leur thèse et à ceux qui l'ont déjà commencée.

Merci à mes amis de Clermont-Ferrand : petit mouton (et son vieux) et Alexis (oui, je ne mettrai pas ici le surnom que tu te donnes. Il ne te convient pas.). Même s'il aura fallu trois ans pour vous faire visiter Toulouse, je sais que vous êtes là pour me soutenir.

Enfin, je remercie ma famille pour m'avoir soutenue, encouragée tout au long de mon cursus. J'en profite pour souhaiter un bon courage à mon petit frère qui se lance à son tour dans cette aventure qu'est la thèse. Merci grand-mère : même si tu n'auras pas vu la fin de cette histoire, tu m'as beaucoup aidée. Pour finir, merci à celui qui partage ma vie (et me supporte) depuis plusieurs années déjà, mon petit chat, qui est toujours là quand j'en ai besoin.

Résumé

Les avancées des nouvelles technologies de séquençage ont permis aux études cliniques de produire des données volumineuses et complexes. Cette complexité se décline selon diverses modalités, notamment la grande dimension, l'hétérogénéité des données au niveau biologique (acquises à différents niveaux de l'échelle du vivant et à divers moments de l'expérience), l'hétérogénéité du type de données, le bruit (hétérogénéité biologique ou données entachées d'erreurs) dans les données et la présence de données manquantes (au niveau d'une valeur ou d'un individu entier). L'intégration de différentes données est donc un défi important pour la biologie computationnelle.

Cette thèse s'inscrit dans un projet de recherche clinique sur l'obésité, DiOGenes, pour lequel nous avons fait des propositions méthodologiques pour l'analyse et l'intégration de données. Ce projet est basé sur une intervention nutritionnelle menée dans huit pays européens et vise à analyser les effets de différents régimes sur le maintien pondéral et sur certains marqueurs de risque cardio-vasculaire et de diabète, chez des individus obèses. Dans le cadre de ce projet, mes travaux ont porté sur l'analyse de données transcriptomiques (RNA-Seq) avec des individus manquants et sur l'intégration de données transcriptomiques (nouvelle technique QuantSeq) avec des données cliniques.

La première partie de cette thèse est consacrée aux données manquantes et à l'inférence de réseaux à partir de données d'expression RNA-Seq. Lors d'études longitudinales transcriptomiques, il arrive que certains individus ne soient pas observés à certains pas de temps, pour des raisons expérimentales. Nous proposons une méthode d'imputation multiple hot-deck (**hd-MI**) qui permet d'intégrer de l'information externe mesurée sur les mêmes individus et d'autres individus. **hd-MI** permet d'améliorer la qualité de l'inférence de réseau.

La seconde partie porte sur une étude intégrative de données cliniques et transcriptomiques (mesurées par QuantSeq) basée sur une approche réseau. Nous y montrons l'intérêt de cette nouvelle technique pour l'acquisition de données transcriptomiques et l'analysons par une approche d'inférence de réseau en lien avec des données cliniques d'intérêt.

Abstract

The development of high-throughput sequencing technologies has lead to a massive acquisition of high dimensional and complex datasets. Different features make these datasets hard to analyze : high dimensionality, heterogeneity at the biological level or at the data type level, the noise in data (due to biological heterogeneity or to errors in data) and the presence of missing data (for given values or for an entire individual). The integration of various data is thus an important challenge for computational biology.

This thesis is part of a large clinical research project on obesity, DiOGenes, in which we have developed methods for data analysis and integration. The project is based on a dietary intervention that was led in eight Europeans centers. This study investigated the effect of macronutrient composition on weight-loss maintenance and metabolic and cardiovascular risk factors after a phase of calorie restriction in obese individuals. My work have mainly focused on transcriptomic data analysis (RNA-Seq) with missing individuals and data integration of transcriptomic (new QuantSeq protocol) and clinic datasets.

The first part is focused on missing data and network inference from RNA-Seq datasets. During longitudinal study, some observations are missing for some time step. In order to take advantage of external information measured simultaneously to RNA-Seq data, we propose an imputation method, hot-deck multiple imputation (**hd-MI**), that improves the reliability of network inference.

The second part deals with an integrative study of clinical data and transcriptomic data, measured by QuantSeq, based on a network approach. The new protocol is shown efficient for transcriptome measurement. We proposed an analysis based on network inference that is linked to clinical variables of interest.

Contributions

Cette thèse est rédigée intégralement en français.

Articles

— Imbert, Alyssa et al. (2018). Multiple hot-deck imputation for network inference from RNA sequencing data. *Bioinformatics*, 34(10) : 1726-1732

— Imbert, Alyssa et Vialaneix Nathalie (2018). Décrire, prendre en compte, imputer et évaluer les valeurs manquantes dans les études statistiques : une revue des approches existantes (accepté par le journal de la Société Française de Statistique)

— un article est en cours de rédaction en collaboration avec Armand Valsesia, Nathalie Viguerie et Nathalie Vialaneix

Communications orales et poster dans des conférences

— Imbert, Alyssa et al. Imputation de données manquantes pour l'inférence de réseau à partir de données RNA-Seq, *Journées de Statitisque de la SFdS*, 2016, Montpellier, France

— Imbert, Alyssa and Villa-Vialaneix Nathalie. Outils pour l'analyse et la simulation de données RNA-seq, *Rencontres R*, 2016, Toulouse, France

— Imbert, Alyssa et al. Multiple hot-deck imputation for network inference from RNA sequencing data, *Statistical Methods for Postgenomic Data* (SMPGD), poster, 2017, Londres, Angleterre

— RNAseqNet : un package pour l'inférence de réseaux à partir de données RNA-seq, *Rencontres R*, 2018, Rennes, France

— Imbert, Alyssa et al. Multiple hot-deck imputation for network inferencefrom RNA sequencing data. *European Conference on Computational Biology* (ECCB), 2018, Athènes, Grèce

Autres communications orales

Les résultats de l'article Bioinformatics [107] ont également été présentés aux journées suivantes :

— *Netbio*, 2017, Paris, France

— *Young Statisticians and Probabilists (YSP)*, 2018, Paris, France

Package R

— **RNAseqNet** [108], implémentant la méthode d'imputation multiple hot-deck (hd-MI). Le package est disponible sur le CRAN.

Séjour au sein d'un laboratoire international

Séjour d'un mois au département Santé Métabolique du Nestlé Health Science à Lausanne, en Suisse, financé grâce à une bourse de mobilité de l'école doctorale MITT.

Cette thèse a été financée par la société Methodomics (`http://www.methodomics.com/`) et la région Languedoc-Roussillon Midi-Pyrénées.

Table des matières

Table des figures

Liste des tableaux

Introduction

Chapitre 1

Contexte biologique

1.1 Données d'expression de gènes

1.1.1 Introduction

Quelques notions de biologie

Les protéines sont des macromolécules composées d'acides aminés remplissant les fonctions vitales essentielles de la cellule. En effet, certaines protéines vont jouer un rôle d'enzyme au sein des cellules, certaines vont avoir un rôle de transmetteur d'information et d'autres permettent de réguler l'activité de certains gènes. Elles sont responsables de la structure cellulaire. Ce sont elles également qui produisent l'énergie et les biomolécules importantes qui participent aux constituants de la cellule. Elles sont aussi responsables de l'ensemble du métabolisme de l'acide désoxyribonucléique (ADN) : synthèse, réplication et réparation en cas de lésion.

Le mécanisme à l'origine de la production des protéines a pour point de départ l'ADN. L'ADN se trouve dans le noyau [1] de chaque cellule et porte l'information génétique. La molécule d'ADN, illustrée par la figure 1.1, est constituée de deux brins complémentaires. Ces brins sont composés de constituants de base, les nucléotides, qui se lient entre eux par une liaison covalente. Les nucléotides sont porteurs d'une des quatre bases azotées : l'adénine (A), la cytosine (C), la guanine (G) et la thymine (T). Les deux brins encodent pour la même information génétique en utilisant l'alphabet complémentaire : $A \leftrightarrow T$, $C \leftrightarrow G$. Chaque brin d'ADN est orienté. Il possède en effet deux extrémités appelées $5'$ et $3'$.

Le génome désigne l'ensemble du matériel génétique d'un organisme. Le génome correspond donc à l'ADN présent dans les cellules. Il est formé de :

— séquences non codantes : séquences non transcrites ou transcrites en acide ribonucléique (ARN) mais non traduites ;

— séquences codantes : séquences transcrites en ARN messagers (ARNm), puis traduites en protéines.

Un gène constitue une unité d'information génétique, codée sous forme d'une séquence de nucléotides, et correspond de ce fait à une petite partie du génome. Les gènes sont constitués d'une alternance d'exons (régions codantes) et d'introns (régions non-codantes).

Le nombre de gènes varie suivant l'organisme, indépendamment de la taille du génome et allant de quelques centaines à plusieurs dizaines de milliers de bases.

La première étape de la synthèse des protéines est la transcription des séquences codantes en ARN. Comme pour l'ADN, l'ARN est un support moléculaire de l'information

1. Note : une petite partie de l'ADN est en réalité de l'ADN mitochondriale provenant des mitochondries

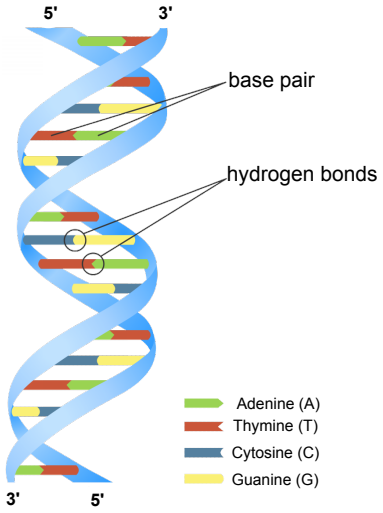


Figure 1.1 **Schéma de l'ADN.**
Source : Genome Research Limited.

génétique. L'ARN n'est constitué que d'un seul brin. L'information est également encodée à l'aide de quatre bases nucléiques. La thymine (T) est ici remplacée par l'Uracile (U). Il existe différents types d'ARN classés selon leur fonction. Parmi les ARN, certains portent en eux l'information génétique codant pour des protéines. Ces ARN, dits codants, sont appelés ARN messagers (ARNm). L'étape de construction des protéines à partir des ARNm s'appelle la traduction. Les autres ARN jouent néanmoins un rôle important dans le fonctionnement cellulaire et participent notamment à la régulation de l'information et à l'activité cellulaire.

Les étapes de transcription et de traduction, illustrées par la figure 1.2, vont être régulées par d'autres protéines, selon l'état de la cellule. Ainsi, pour une cellule, la population d'ARNm ou de protéines est caractéristique de son état à un moment donné. Par conséquent, il est possible de chercher à mesurer l'abondance de tous les ARNm ou de toutes les protéines pour comprendre ce qui se passe dans la cellule.

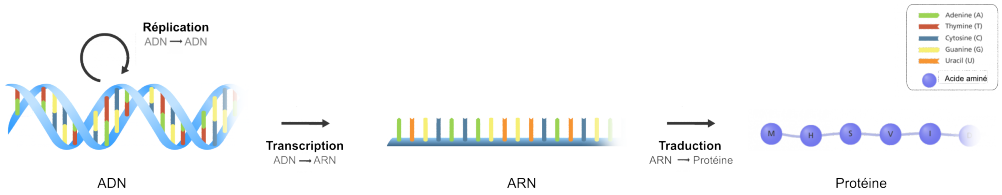


Figure 1.2 **Étapes de la synthèse des protéines.**
Source : figure (retravaillée) provenant de `https://www.yourgenome.org`.

Le transcriptome est l'ensemble des ARNm d'un tissu ou de cellules à un instant donné et dans des conditions données. Le transcriptome peut-être considéré comme un reflet de l'ensemble des protéines produites par la cellule. La caractérisation et la quantification du transcriptome, dans un tissu donné et dans des conditions données, permettent d'identifier des gènes qui sont transcrits, de déterminer les mécanismes de régulation d'expression de ces gènes et d'identifier les réseaux de régulation de l'expression des gènes.

En laboratoire

Pour mesurer la quantité de transcrits, plusieurs méthodes existent. Tout d'abord, les conditions expérimentales et le nombre d'échantillons par condition doivent être déterminés en fonction du type d'analyse. L'ARN est extrait du tissu ou des cellules et retranscrit en ADN complémentaire (ADNc). Le brin complémentaire de l'ADNc est également produit. Une étape commune aux méthodes décrites dans cette section consiste à dupliquer en grand nombre les séquences d'ADN afin que la quantité soit suffisante pour quantifier l'expression des gènes. Pour cela, on utilise des réactions en chaîne par polymerase [2] (PCR).

La PCR [156] permet d'amplifier in vitro une région spécifique d'une séquence d'acides nucléiques donnée afin d'en obtenir une quantité suffisante pour la détecter et l'étudier. La PCR exploite le processus de la réplication de l'ADN. Pour cela, elle s'appuie sur la capacité de l'ADN polymérase à synthétiser le brin complémentaire d'un brin simple d'ADN (servant de matrice). À partir d'une copie d'une séquence d'acides nucléiques, la séquence peut être amplifiée et détectée.

Un cycle de PCR peut-être décomposé en trois étapes :

1. **la dénaturation** : les brins d'ADN sont portés à une certaine température pour séparer les deux brins qui le composent ;
2. **l'hybridation** : des amorces (*primers*, en anglais) vont s'hybrider aux extrémités de la séquence d'ADN recherchée sur chaque simple brin d'ADN ;
3. **l'élongation** : des enzymes, les polymérases, parcourent le brin d'ADN matrice et synthétisent le brin complémentaire.

À la fin de chaque cycle PCR, les produits obtenus se présentent sous la forme d'ADN double brin. Ces trois étapes sont effectuées à des températures différentes, ce qui permet de contrôler l'activité enzymatique. Cet enchaînement de trois étapes est reproduit successivement, en utilisant les produits obtenus à la fin de chaque cycle. La figure 1.3 illustre les deux premiers cycles d'une PCR.

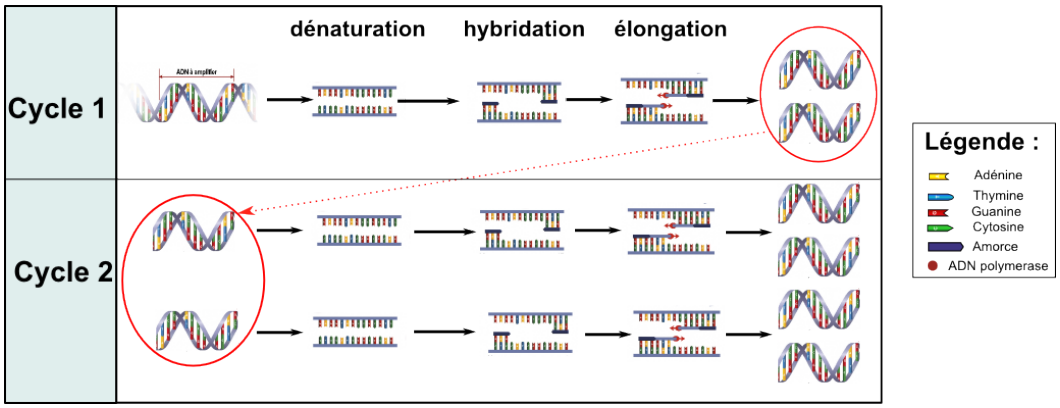


Figure 1.3 Schéma représentant les différentes étapes des deux premiers cycles d'une PCR.

Il est alors possible de quantifier la quantité d'ADNc. Pour cela, diverses méthodes existent, permettant de mesurer de façon exhaustive et simultanée l'expression de l'ensemble des gènes dans un type donné de cellule. Les méthodes peuvent être divisées en deux familles : celles basées sur le principe d'hybridation permettant d'obtenir des données

2. *Polymerase Chain Reaction*, en anglais

d'expression continues et celles basées sur des techniques de séquençage donnant des données d'expression discrètes.

1.1.2 Techniques basées sur l'hybridation

PCR quantitative, en temps réel (RT-qPCR)

La PCR quantitative [206] est une méthode particulière de réaction PCR permettant de mesurer la quantité initiale d'ADN. Cette technologie est basée sur la détection d'un signal (généralement un signal fluorescent) au cours des cycles de PCR. À chaque cycle, le taux de ce signal est mesuré et est proportionnel à la quantité d'amplicons (portions d'ADN définie par un couple d'amorce) générés.

Le principe de la PCR en temps réel (RT-qPCR) repose donc sur la possibilité de suivre la quantité d'ADN présente dans la réaction à tout instant. Elle s'oppose de ce fait à la PCR où les amplicons ne sont détectés qu'à la fin du processus.

Le profil d'une réaction PCR peut être décomposé en trois phases, illustrées par la figure 1.4 :

1. **une phase d'initiation** : elle s'achève lorsque le nombre de produits PCR néoformés dépasse la valeur seuil prédéfinie pour l'expérimentation ;
2. **une phase exponentielle** : durant cette phase, le nombre de produits PCR double à chaque cycle ;
3. **une phase de plateau** : elle commence lorsque les constituants nécessaires pour la PCR deviennent limitant.

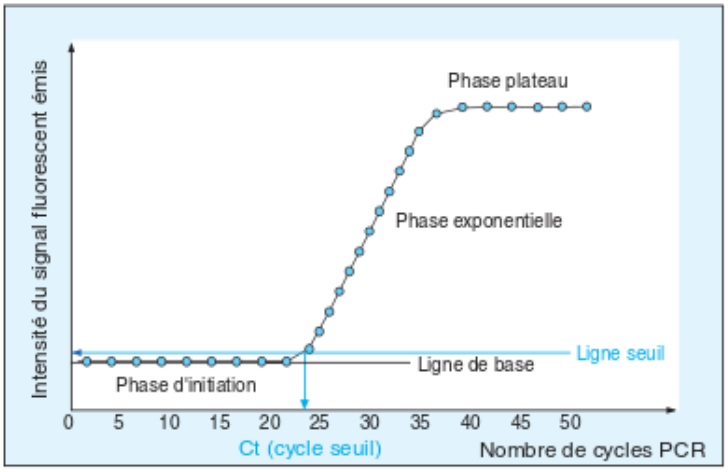


Figure 1.4 **Modèle en temps réel d'une PCR en temps réel.** L'intensité du signal (émission fluorescente) est exprimée en fonction du nombre de cycles PCR.
Source : figure provenant de [216]

L'évolution de l'intensité du signal en fonction du nombre de cycles PCR est appelée courbe d'amplification. Un seuil pour le niveau de fluorescence, suffisamment élevé pour être au-dessus du bruit de fond qui correspond au signal détecté pendant la phase d'initiation, est défini. Le point d'intersection entre ce seuil et la courbe d'amplification s'appelle le cycle seuil (*cycle threshold*, en anglais) et est noté C_t. Il correspond au nombre de cycles PCR minimal

pour lequel l'ADN amplifié est détectable. Il est atteint au début de la phase exponentielle. Plus la quantité initiale d'ARNm cible est faible, plus le C_t est élevé.

La quantité d'ARNm dans l'échantillon est déduite de sa valeur C_t. De plus, elle est exprimée par rapport à l'expression d'un gène de référence. Ce gène de référence (*housekeeping*, en anglais) est un gène qui reste stable (pas de changement d'expression) tout au long de l'expérimentation. Il permet de normaliser les données afin de supprimer certains biais techniques de la PCR (variations dans la quantité et qualité des échantillons, rendement d'extraction différent entre les échantillons, etc.).

La méthode RT-qPCR ne permet pas de découvrir de nouvelles régions codantes de l'ADN. Les données obtenues sont des données quantitatives, relatives (normalisation par un gène de référence) et continues. Pour cette approche, le nombre de gènes peut aller de quelques dizaines à quelques centaines. Il est possible de quantifier l'expression de gènes sur un nombre assez important d'individus (jusqu'à quelques centaines).

Puces à ADN

Cette technique de mesure (*microarray*, en anglais) permet de quantifier le niveau relatif de l'expression de plusieurs milliers de gènes à la fois, représentée par l'abondance des transcrits, dans un tissu donné, à un instant donné et/ou dans un état donné.

Comme pour la RT-qPCR, la technologie des puces à ADN est basée sur le principe d'hybridation développé par [197]. Ce principe repose sur le fait que deux fragments d'acides nucléiques complémentaires peuvent s'associer et se dissocier de façon réversible sous l'action de la chaleur et de la concentration saline du milieu. La puce à ADN est un support rigide (verre ou nylon) sur lequel a été fixé de façon ordonnée un ensemble de fragments d'ADN dont la séquence de nucléotides est connue, représentative d'un gène et complémentaire de l'ADNc. Ces fragments, appelés sondes, sont des oligonucléotides de synthèse ou des produits de PCR. Ce microdispositif est mis au contact des ADNc synthétisés à partir des ARN extraits des échantillons à analyser, appelés cibles. Ces cibles sont marquées par incorporation de radioéléments ou de fluorochromes. Chaque cible va s'apparier par complémentarité des bases avec la sonde lui correspondant sur la puce. Il est alors possible de quantifier les abondances relatives de chaque ARNm présent dans les échantillons grâce à la lecture des signaux fluorescents en mesurant l'intensité des signaux d'hybridation. La figure 1.5 schématise le principe d'une puce à ADN.

Après acquisition des images d'hybridation, la quantification des signaux d'hybridation reflète le niveau d'expression, dans l'échantillon initial, de chacun des gènes représentés sur la puce. Le niveau de précision de la méthode dépend du nombre de fragments d'ADN attachés sur la puce à ADN. Cette méthode nécessite de connaître par avance la séquence d'ADN que l'on souhaite hybrider. Elle ne permet donc pas de découvrir de nouvelles régions codantes de l'ADN.

Les données sont des mesures d'intensité de fluorescence. De par ce fait, elles sont bornées. En effet, la quantification du taux de fluorescence est limitée par les capacités technologiques. Les données obtenues sont donc quantitatives, relatives et continues. Le nombre de variables, soit de transcrits (ARNm) quantifiables, est de l'ordre de plusieurs milliers et le nombre d'individus est de l'ordre de la dizaine généralement.

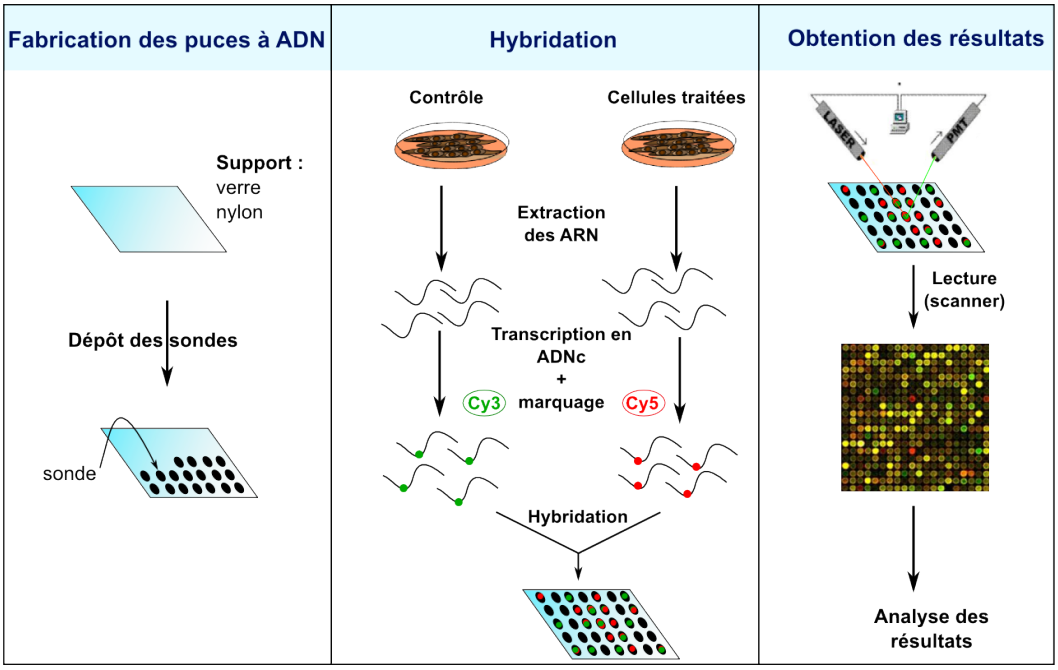


Figure 1.5 Schéma du principe d'une puce à ADN.

1.1.3 Techniques basées sur le séquençage

RNA-Seq

Le RNA-Seq [53] est une technique de séquençage, à haut débit, qui mesure l'abondance de séquences d'ARN dans une cellule ou un tissu donné pour des milliers de gènes simultanément. Le séquençage d'un fragment d'ARN consiste à déterminer l'ordre d'enchaînement des nucléotides qui le constituent.

Après avoir extrait et isolé l'ARNm, il est transformé en ADNc. Le brin complémentaire de l'ADNc est également produit. L'ADNc est alors amplifié par PCR afin d'en disposer une quantité suffisante pour le quantifier.

Nous décrivons ici les différentes étapes du séquençage de l'ADNc par la technologie RNA-Seq :

1. **lecture des brins d'ADNc** : l'ADNc est coupé en petits fragments de longueurs de 200 à 300 paires de bases. Ces fragments sont lus par un séquenceur. Les lectures données par ce dernier sont appelés « *reads* » ;

2. **contrôle de qualité du séquençage** : les *reads* de mauvaises qualités sont supprimés ou rétrécis (*trimming*). En effet, les premières bases d'un *read* sont généralement séquencées avec beaucoup de fiabilité, mais plus on avance dans la séquence, plus des erreurs sont probables. Le *trimming* consiste donc à amputer les extrémités des séquences de façon à améliorer la qualité des *reads* et par conséquent l'alignement ;

3. **alignement des *reads*** : l'alignement (*mapping*, en anglais) consiste à rechercher dans le génome la position d'une sous-séquence similaire à celle du *read* obtenu par séquençage. Si le génome de référence est disponible, les *reads* sont alignés sur celui-ci. Idéalement, le génome de référence est la séquence complète d'ADN à partir de laquelle

les ARN ont été produits. Généralement, il s'agit de la séquence d'ADN type de l'espèce étudiée. Celle-ci est disponible, par exemple, en téléchargement depuis le site NCBI [3] ;

4. **comptage des *reads*** : les *reads* alignés sur chaque région génomique d'intérêt sont comptés. Le nombre de *reads* lus et alignés sur une région d'intérêt est a priori considéré comme proportionnel au niveau d'expression de la région d'intérêt et à la taille de cette région. Le nombre moyen de *reads* alignés par position du génome s'appelle la couverture de séquençage. Plus elle est grande, plus le séquençage est complet. Le nombre total de *reads* alignés pour un échantillon est appelé profondeur de séquençage ou taille de librairie. Le résultat final est représenté sous la forme d'un tableau de comptage, comportant en ligne les gènes et en colonnes les échantillons et dont les entrées sont le nombre de *reads* trouvés dans un échantillon donné qui ont été alignés sur la séquence d'un gène donné.

La figure 1.6 permet de résumer les différentes étapes de la technique RNA-Seq sous la forme d'un schéma.

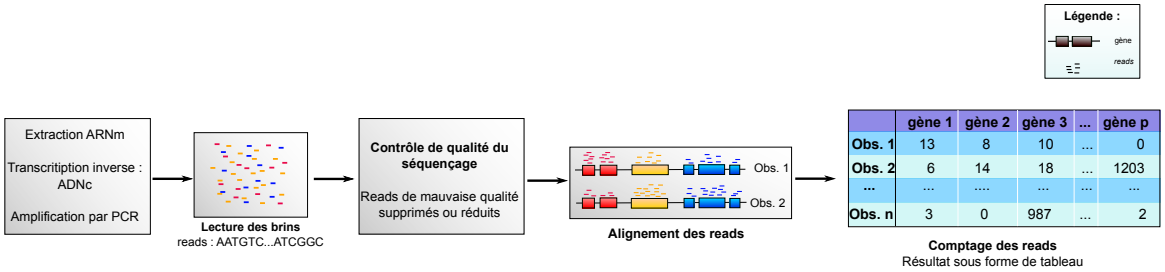


Figure 1.6 Les différentes étapes du RNA-Seq.

Les données obtenues sont quantitatives et discrètes. Tout le génome étant séquencé, le nombre de variables (soit de transcrits) est de l'ordre de plusieurs milliers. Cependant, la technologie RNA-Seq étant onéreuse, le nombre d'individus est généralement relativement faible. Il est de l'ordre de quelques dizaines pour des études standards. Pour des projets importants, ayant plus de moyens financiers, il est possible d'avoir des données avec quelques centaines d'individus.

La technologie RNA-Seq mesure l'expression sans limite d'amplitude (c'est-à-dire que les données sont non bornées) contrairement à la technologie des puces à ADN qui subissent une saturation de l'intensité de fluorescence pour les gènes très exprimés. De plus, l'alignement des séquences peut être réalisé sur un génome de référence incomplet (alignement *de novo*) permettant ainsi la découverte de nouveaux gènes.

Avec le développement des technologies de séquençage à haut débit, le RNA-Seq est devenu un nouveau standard dans l'analyse du transcriptome. Néanmoins, le coût d'acquisition reste assez élevé. Un nouveau protocole a donc été mis en place pour réduire les dépenses des différentes étapes nécessaires pour obtenir les données d'expression des gènes.

QuantSeq

Le QuantSeq [150] est une nouvelle méthode de séquençage à haut débit. Il permet de réduire le temps d'analyse des données et d'analyser plus d'échantillons simultanément. Le principe de cette technique est de générer un seul fragment pour chaque transcrit et de

3. `https://www.ncbi.nlm.nih.gov/`

l'amplifier. Les fragments séquencés sont situés près de l'extrémité 3' du transcript (voir figure 1.7). Cette façon de procéder permet d'obtenir des valeurs d'expression très précises.

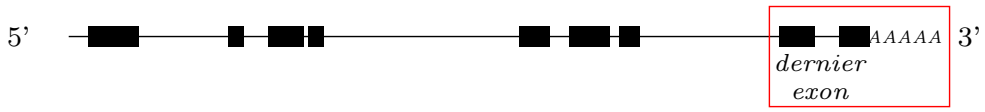


Figure 1.7 Zone du gène séquencé par le protocole QuantSeq.

Le protocole QuantSeq se déroule en plusieurs étapes, illustrées par la figure 1.8 :

1. **préparation des librairies** : cette étape est initiée grâce à une amorce en oligodT qui se fixe sur le brin d'ARN. Une transcription inverse est alors effectuée. L'ARNm initial est supprimé et une amorce aléatoire (*random priming*) permet de synthétiser le second brin. La librairie est purifiée pour supprimer diverses composants de réactions inutiles pour les étapes suivantes :
2. **amplification** : les librairies sont amplifiées par PCR. Cette étape a notamment pour objectif de générer une quantité suffisante de matériel pour le contrôle de la qualité et pour le séquençage ;
3. **séquençage** : les fragments sont lus par un séquenceur ;
4. **traitement des données** : les *reads* sont alignés le long du génome et comptés. Le niveau d'expression d'un transcript est proportionnel au nombre de *reads* alignés sur la séquence de ce transcript.

Un des principaux avantages de la technologie QuantSeq par rapport à celle du RNA-Seq est sa tolérance à la mauvaise qualité de l'ARN. Autrement dit, il est possible, avec la technique QuantSeq, de quantifier des transcrits provenant d'ARN de qualité médiocre. En outre, cette technologie est moins coûteuse et plus rapide. En revanche, avec cette technique, la recherche d'isoformes (variantes des transcrits pour un gène donné) n'est plus possible.

Comme pour le RNA-Seq, les données sont des données de comptage. Elles sont donc quantitatives et discrètes. Le nombre de gènes est également de l'ordre de plusieurs milliers. Comme cette technologie est moins chère que le RNA-Seq, il est possible de séquencer plus d'échantillons pour un prix équivalent, de l'ordre de cinq à dix fois.

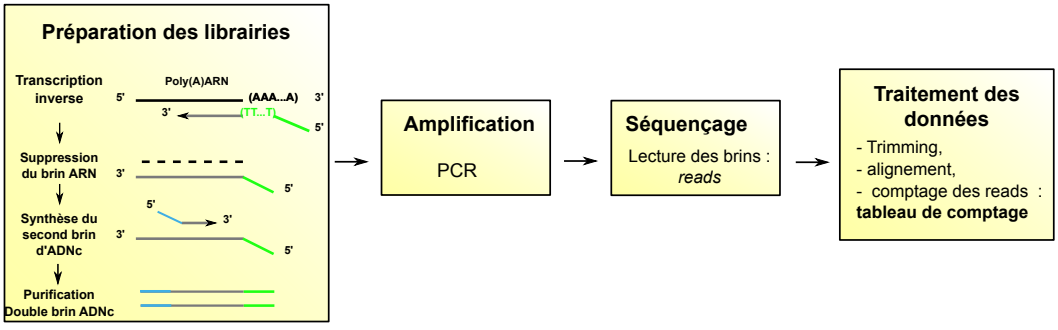


Figure 1.8 Schéma du protocole QuantSeq.

Le tableau 1.1 récapitule les différentes techniques, présentées dans ce chapitre, permettant de quantifier l'expression des gènes, ainsi que quelques unes de leurs caractéristiques.

Techniques	Type de données	Nombre de gènes	Nombre d'individus	Gènes
RT-qPCR	données continues	plusieurs dizaines à quelques centaines	plusieurs dizaines à quelques centaines	connus
Puces à ADN	données continues	plusieurs milliers	une dizaine	connus
RNA-Seq	données discrètes (comptages)	plusieurs dizaines de milliers	une dizaine	connus et inconnus
QuantSeq	données discrètes (comptages)	plusieurs dizaines de milliers	plusieurs dizaines	connus et inconnus

Tableau 1.1 Récapitulatif des méthodes pour quantifier l'expression des gènes.

Dans cette thèse, nous nous intéressons en particulier aux données issues des technologies de séquençage (RNA-Seq et QuantSeq). Des données RT-qPCR seront également utilisées pour apporter de l'information supplémentaire.

1.2 DiOGenes, une étude sur l'obésité

Cette thèse s'appuie sur un projet de recherche clinique sur l'obésité, DiOGenes (Diet, Obesity and Genes). À partir des diverses données mesurées au cours de cette étude, des questions biologiques mais aussi des problèmes rencontrés, nous avons proposé diverses méthodologies pour analyser ces données et intégrer l'information provenant des différents jeux de données disponibles.

1.2.1 L'obésité, une maladie chronique

L'obésité concerne aujourd'hui la quasi-totalité de la planète. Selon les estimations de l'Organisation Mondiale de la Santé (OMS [4]), en 2016, 39% de la population mondiale adulte était en surpoids et 13% était obèse. Entre 1975 et 2016, la prévalence de l'obésité a presque triplé à l'échelle mondiale. Maladie de l'adaptation aux récentes évolutions des modes de vie, l'obésité est à l'origine de nombreux troubles de santé comme le diabète de type II, l'hypertension artérielle, la dyslipidémie (excès de lipides dans le sang), les maladies cardiovasculaires, le syndrome d'apnée du sommeil et des maladies dermatologiques. De plus, l'obésité est associée à un risque accru pour certains cancers.

Définition de l'obésité

L'obésité est définie par l'OMS comme un excès de masse grasse ayant des conséquences néfastes pour la santé. La masse grasse corporelle est essentiellement constituée de tissu adipeux.

En pratique clinique, on définit l'obésité par l'Indice de Masse Corporelle (IMC [5]).

4. page de l'OMS consacrée à ce sujet : `http://www.who.int/topics/obesity/fr/`
5. en anglais : Body Mass Index (BMI)

Définition de l'Indice de Masse Corporelle (IMC)

L'IMC est la manière la plus simple pour évaluer le surpoids et l'obésité. Il s'agit d'une mesure du poids par rapport à la taille. Son unité est en kg/m^2 et il est calculé de la façon suivante :

$$IMC = \frac{\text{poids}}{\text{taille}^2}$$

L'IMC étant corrélé à la quantité de masse adipeuse, ce critère permet donc d'évaluer facilement et rapidement le surpoids et l'obésité pour un individu. La classification des individus par rapport à l'IMC est donnée dans le tableau 1.2. Cependant, il ne faut pas oublier qu'il ne s'agit que d'une indication approximative puisqu'il ne permet ni de faire la distinction entre masse grasse et masse maigre, ni de tenir compte de la répartition du tissu adipeux dans l'organisme. En effet, pour un même IMC, la composition corporelle peut varier d'un individu à l'autre. Prenons comme exemple un sportif de haut niveau qui aura un IMC élevé, du fait de sa masse musculaire, sans pour autant présenter d'excès de masse grasse.

IMC	Interprétation (selon l'OMS)
Moins de 16.5	dénutrition
16.5-18.5	maigreur
18.5 - 25	poids « idéal » (valeurs de référence)
25 - 30	surpoids
30-35	obésité modérée
35-40	obésité sévère
Au-delà de 40	obésité massive

Tableau 1.2 Classification des individus selon leur corpulence.

D'autres mesures permettent d'affiner les estimations obtenues par le calcul de l'IMC. Par exemple, le tour de taille est un autre critère qui permet d'estimer si un individu est atteint d'obésité. L'excès de masse grasse, localisé autour du ventre, est associé à un risque accru de maladies, telles le diabète ou encore des maladies cardiovasculaires, indépendamment de l'IMC. Lorsque le tour de taille est supérieur à 94 cm chez l'homme et à 80 cm chez la femme (en dehors de la grossesse), on parle alors d'obésité abdominale [5]. Les seuils, présentés ici, sont définis pour une population caucasienne. D'autres seuils spécifiques à chaque population (asiatique, africaine, etc.) existent afin d'ajuster au mieux les critères diagnostiques de l'obésité.

Si l'IMC n'est pas la mesure la plus fiable du surpoids, elle permet néanmoins de définir facilement et rapidement un problème d'obésité chez l'individu. Des techniques permettent de mesurer de manière plus précise la masse grasse. Elles sont cependant plus coûteuses et les plus précises demandent une logistique lourde. Elles peuvent donc être difficile à mettre en place pour de grands échantillons d'individus.

Les causes de l'obésité

L'obésité résulte d'une dérégulation, sur le long terme, de la balance énergétique et donc d'un bilan positif entre apports alimentaires et dépense énergétique. Ce déséquilibre aboutit à une inflation des réserves stockées dans le tissu adipeux, ce qui entraîne de nombreuses complications.

Les origines de l'obésité sont multiples. Son développement repose à la fois sur des facteurs génétiques, biologiques, comportementaux et environnementaux.

En effet, les modifications de l'alimentation et la réduction de l'activité physique jouent un rôle certain dans l'émergence récente de l'obésité. L'augmentation de la taille des portions, des aliments plus riches en lipides, sucre et sel, un accès plus simple des aliments ont favorisé les consommations caloriques excessives. En parallèle, l'urbanisation croissante, l'utilisation de transport (voiture ou transport en commun) dans les déplacements quotidien, etc., induisent quant à eux une diminution des dépenses énergétiques.

Cependant, ces facteurs influençant le bilan énergétique ne suffisent pas pour expliquer l'augmentation de la fréquence de l'obésité, ni l'inégalité des individus face à la prise de poids.

Une prédisposition génétique à la prise de poids peut rendre compte de ces différences individuelles. En effet, des gènes impliqués dans la prise de poids, l'obésité sévère et/ou les complications de l'obésité ont été identifiés [70].

Le rôle de l'environnement semble également jouer un rôle important comme le stress ou la privation de sommeil [198].

Il est donc important de connaître au mieux tous les facteurs et de voir comment ils interagissent entre eux. Accéder à une meilleur compréhension des causes et des mécanismes biologiques conduisant à l'obésité est encore un enjeu de la recherche médicale.

1.2.2 DiOGenes

Le projet DiOGenes [125] est une étude d'intervention diététique contrôlée sur des personnes obèses, réalisée dans huit pays européens. Son principal objectif est d'identifier l'efficacité de régimes particuliers dans ou contre la reprise de poids, après un régime amaigrissant, pour des personnes en surpoids ou obèses.

La partie clinique s'est déroulée en deux phases, illustrées par la figure 1.9. La première phase consiste en un régime hypocalorique, c'est-à-dire faible en calories (soit environ 800-1000 kcal par jour), de huit semaines avec pour objectif de perdre au moins 8 % du poids initial. La deuxième phase est une phase de suivi pondéral de six mois. Les sujets retenus pour cette phase, soit ceux qui ont perdu plus de 8% de leur poids initial pendant le régime hypocalorique, sont répartis aléatoirement en cinq groupes de régime, répertoriés dans le tableau 1.3 : quatre régimes avec différentes teneurs en protéines et indice glycémique et un régime témoin :

1. 25-30% de l'énergie provient des lipides, 10-15% des protéines et 57-62% des glucides à faible indice glycémique ;
2. 25-30% de l'énergie provient des lipides, 10-15% des protéines et 57-62% des glucides avec un indice glycémique élevé ;
3. 25-30% de l'énergie provient des lipides, 23-28% des protéines et 45-50% des glucides à faible indice glycémique ;
4. 25-30% de l'énergie provient des lipides, 23-28% des protéines et 45-50% des glucides avec un indice glycémique élevé ;
5. régime témoin correspondant à un régime équilibré du pays.

Avant et après chaque phase, des mesures cliniques et phénotypiques ont été obtenues et des prélèvements de sang et des biopsies de tissu adipeux ont été réalisées. Les ARN sont ensuite extraits de ces biopsies et des analyses de transcriptome du tissu adipeux réalisées.

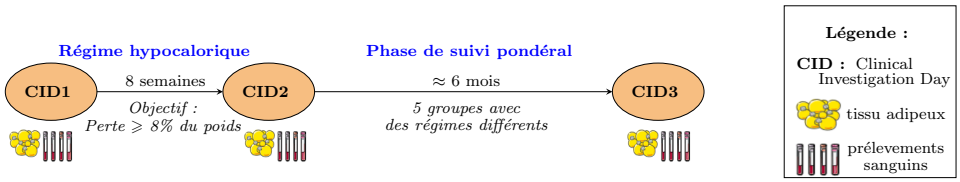


Figure 1.9 Schéma du protocole du projet DiOGenes.

	Indice glycémique faible	**Indice glycémique élevé**
Taux normal de protéine	1	2
Taux élevé de protéine	3	4
Régime témoin : 5		

Tableau 1.3 Groupes diététiques durant la phase de suivi pondéral.

Les données disponibles sont donc :

— des données phénotypiques et cliniques ;

— la quantification des transcrits provenant du tissu adipeux mesurés via trois techniques différentes : la RT-qPCR, le RNA-Seq et le QuantSeq.

Le nombre d'échantillons et de variables pour chacun de ces jeux de données et pour chaque pas de temps (CID : *Clinical Investigation Day*) est répertorié dans le tableau 1.4.

	Données cliniques	**RT-qPCR**	**RNA-Seq**	**QuantSeq**
Nombre de variables	> 80	284 gènes	54 043 gènes	32 041 gènes
Nombre d'échantillons avec données				
Point de départ de l'étude (CID1)	632	495	451	416
Après le régime hypocalorique (CID2)	622	544	389	291
Après la phase de suivi pondéral (CID3)	473	371	164	211

Tableau 1.4 Nombre de variables et d'échantillons disponibles pour les différents jeux de données.

D'autres types de données ont été mesurées comme les acides gras présents dans le tissu adipeux (analyse lipidomique), l'expression des microARN du tissu adipeux ou encore l'expression des microARN plasmatiques. Ces données n'ont pas été utilisées dans le cadre de cette thèse.

Chapitre 2

Cadre statistique

Les notations utilisées dans la suite de ce chapitre sont définies ici. Par abus de langage, toute région génomique d'intérêt est assimilée à un gène. Le terme échantillon est employé pour désigner les réplicats biologiques. Le tableau de comptage d'expression de gènes est représenté sous la forme d'une matrice, notée X. Cette matrice est de taille $n \times p$, où n est le nombre d'échantillons séquencés et p le nombre de gènes. La notation x_{ij} correspond à l'expression du gène j pour l'échantillon i. Cette expression est fortement dépendante de la *profondeur de séquençage*, c'est-à-dire du nombre de *reads* alignés pour un échantillon i. Nous notons, $N_i = \sum_{j=1}^{p} x_{ij}$, la profondeur de séquençage (ou taille de librairie), de l'échantillon i.

2.1 Modélisation statistique des données de comptage

2.1.1 Modélisation des données RNA-Seq

Dans les expérimentations de séquençage à haut débit, les données brutes correspondent à des millions de *reads* qui sont alignés sur une région spécifique du génome. Soit x_j le nombre de *reads* séquencés pouvant être assignés à une région spécifique (c'est-à-dire à un gène) j. La probabilité p_j pour un *read* d'être aligné sur une région j est estimée par la proportion de fragments d'ADN provenant d'une région génomique j.

Le comptage du gène j, X_{ij}, peut alors être modélisé par une distribution binomiale de paramètre N_i et p_j :

$$P(X_{ij} = k) = \binom{N_i}{k} p_j^k (1 - p_j)^{N_i - k}.$$

Le nombre de *reads* séquencés N_i étant très grand et la probabilité p_j très faible, la loi binomiale $Bin(N_i, p_j)$ peut être approchée par une loi de Poisson de paramètre λ_{ij} avec $\lambda_{ij} = N_i p_j$.

$$\begin{aligned} X_{ij} &\sim \mathcal{P}(\lambda_{ij}), \\ P(X_{ij} = x_{ij}) &= \frac{\lambda_{ij}^{x_{ij}}}{x_{ij}!} e^{-\lambda_{ij}}, \\ E(X_{ij}) &= \text{Var}(X_{ij}) = \lambda_{ij}. \end{aligned}$$

Cependant, pour les données d'expression RNA-Seq, la variance d'un gène est généralement plus élevée que sa moyenne. On parle de surdispersion, celle-ci provenant, en

particulier, d'une variabilité biologique. Le modèle de Poisson ne s'avère donc pas approprié pour modéliser correctement les données d'expression RNA-Seq.

Pour prendre en compte cette variabilité, des solutions alternatives ont été proposées :

— **des modèles de Poisson surdispersés** [16] : l'idée est d'imposer une relation linéaire entre la variance et la moyenne : $\text{Var}(X_i) = \phi E(X_i)$ où ϕ correspond au paramètre de surdispersion. Des approches basées sur la quasi-vraisemblance [230] sont utilisées pour estimer les paramètres ;

— **des modèles basés sur des distributions binomiales négatives** [10, 175] : la loi binomiale négative possède en effet un paramètre supplémentaire qui permet de modéliser la variance indépendamment de la moyenne. Elle peut être vue comme une loi de Poisson dont le paramètre serait une variable aléatoire suivant une distribution Gamma. Elle s'écrit de la façon suivante :

$$X_{ij} \sim \mathcal{NB}(\mu_{ij}, \phi_j),$$

$$P(X_{ij} = x_{ij}) = \left(\frac{\phi_j}{\phi_j + \mu_{ij}}\right)^{\phi_j} \frac{\Gamma(\phi_j + x_{ij})}{x_{ij}!\Gamma(\phi_j)} \left(\frac{\mu_{ij}}{\phi_j + \mu_{ij}}\right)^{x_{ij}},$$

$$E(X_{ij}) = \mu_{ij},$$

$$\text{Var}(X_{ij}) = \mu_{ij} + \mu_{ij}^2 \phi_j$$

où $\phi_j \geq 0$ correspond au paramètre de dispersion du gène j.

2.1.2 Transformation des données

Les données de comptage sont des données discrètes, très hétérogènes. Leur distribution est asymétrique et des valeurs extrêmes sont généralement présentes dans ce type de données [246]. Des modèles basés sur des lois discrètes comme la loi de Poisson ou la loi binomiale négative ont donc été proposés afin de les modéliser le plus correctement possible (section 2.1.1). Néanmoins, une autre approche est possible. En effet, des transformations plus ou moins complexes peuvent être appliquées sur ces données afin de se ramener à des modèles plus connus, fondés sur des lois gaussiennes.

Les transformations de variables sont souvent utilisées pour induire des propriétés « désirables » (par exemple, la normalité, l'homoscédasticité, la linéarité) afin de visualiser les données, d'utiliser des tests paramétriques ou des procédures d'estimation de paramètres. Le principe de la transformation est de générer une nouvelle variable X' à partir de la variable X, cette dernière ne respectant pas les propriétés souhaitées (distribution normale par exemple). La variable X' est définie comme une fonction de X :

$$X' = f(X)$$

où f est une fonction à choisir de telle sorte que les données transformées aient les propriétés recherchées. Dans le cas où l'objectif est d'obtenir des données suivant une distribution normale, le choix de cette fonction f va dépendre de l'allure de la distribution des fréquences des données brutes X.

Transformations classiques

Dans le cadre de l'analyse de données de comptage (épidémiologie, écologie, etc.), trois types de transformation sont couramment utilisés : la transformation logarithmique, la transformation racine-carrée et la transformation arcsin. Pour des données de comptage suivant une distribution de Poisson, la littérature [196, 141, 199] recommande d'utiliser la fonction racine-carrée. Lorsque le nombre de comptages nuls est important, [196] conseillent une fonction logarithmique ou la transformation suivante :

$$f(x) = \sqrt{(x + c)}$$

où c vaut 0,5 ou encore $3/8$. En outre, lorsque la variance est corrélée positivement avec la moyenne, il est conseillé d'utiliser la transformation logarithmique.

En ce qui concerne les données d'expression RNA-Seq, la transformation généralement utilisée est la transformation logarithmique. Les données RNA-Seq pouvant être nulles, la transformation logarithmique utilisée est la suivante :

$$f(x) = \log(x + c)$$

où c est une constante (et vaut généralement 1). Les données transformées ont une distribution qui se rapproche d'une distribution plus symétrique, proche de celle d'une distribution normale. La variabilité peut néanmoins encore être très importante après cette transformation [78]. La transformation Box-Cox [28], appartenant à la famille des transformations puissances, permet de généraliser la transformation logarithmique. Elle se définit comme suit :

$$f(x) = \begin{cases} \dfrac{x^\delta - 1}{\delta} \text{ si } \delta \neq 0 \\ \log(x) \text{ si } \delta = 0 \end{cases}$$

où la valeur de δ est choisie de façon à maximiser la log-vraisemblance des données transformées. Pour la même raison que la transformation logarithmique, la méthode est légèrement modifiée pour prendre en compte les valeurs nulles. Ainsi, à la place de $\log(x)$, nous utilisons $\log(x + 1)$.

Transformation normale inverse basée sur le rang [1]

Les transformations présentées ci-dessus sont des transformations paramétriques. Il est également possible d'utiliser des transformations non paramétriques telles que les transformations basées sur le rang. Cette approche consiste à rendre des distributions comparables en les transformant en rang permettant ainsi d'éliminer les unités de mesures, les ordres de grandeur et les différences de dispersion. Ces transformations peuvent donc être vues comme des méthodes de normalisation puisqu'elles permettent d'aligner les densités des divers échantillons et par conséquent de ramener les échantillons à des niveaux comparables.

La première étape consiste donc à convertir une variable en rangs :

$$r_{ij} = \text{rang}_{i=1,\ldots,n}(x_{ij})$$

1. Rank-based inverse normal transformation

Différentes méthodes existent mais elles sont toutes basées sur le modèle suivant :

$$f(x_{ij}) = \Phi^{-1}\left(\frac{r_{ij} - c}{n - 2c + 1}\right)$$

où Φ^{-1} correspond à la fonction quantile (ou probit dans certain cas). La différence réside dans le choix de la valeur de la constante c [23]. Parmi ces transformations, la transformation Blom [27] est généralement la plus utilisée. La valeur de la constante c pour la transformation Blom vaut $3/8$.

[246, 158] ont comparé différentes transformations des données RNA-seq respectivement pour améliorer la performance de prédiction et la classification utilisant des modèles gaussiens. [158] montrent que la transformation Blom est celle qui permet d'obtenir des données transformées se rapprochant le plus de données suivant une distribution normale. En ce qui concerne les performances de la classification, les résultats montrent qu'il est préférable d'utiliser une transformation logarithmique, voire la transformation VST plutôt que la transformation Blom ou de travailler avec des données non transformées. [246] montrent que le choix de la transformation appropriée est essentielle et a une influence importante sur les gènes étant sélectionnés comme différentiellement exprimés, sur le nombre de vrais positifs et sur les performances de la prédiction du modèle. Dans les simulations, la transformation la plus appropriée semble être la transformation basée sur les rangs. Elle est suivie par la transformation logarithmique et celle de Box-Cox sur données réduites. Ils montrent également l'importance de réduire la variance des covariables (quelle que soit la transformation utilisée).

Selon l'analyse que l'on souhaite effectuer, il est possible que ces transformations ne soit pas adaptées ou ne permettent pas d'obtenir exactement les propriétés désirées. Diverses transformations, utilisées dans un cadre spécifique, ont alors été proposées pour les données de comptage RNA-Seq.

Transformations pour stabiliser la variance

L'objectif de ces approches est de stabiliser la variance, notamment en limitant, voire supprimant la relation existant entre la moyenne et la variance.

Deux approches sont disponibles dans le package **DESEq2**. Dans ce package, les comptages X_{ij} pour le gène j dans l'échantillon i sont décrits avec un modèle linéaire généralisé en utilisant une famille binomiale négative avec un lien logarithmique :

$$X_{ij} \sim \mathcal{NB}(\mu_{ij}, \phi_j),$$

$$\text{avec } \mu_{ij} = s_i \lambda_{ij} \qquad \text{et } \log_2(\lambda_{ij}) = D_{i.}\beta_{j.}$$

où s_i est le facteur de normalisation (voir section 2.1.3), λ_{ij} est un paramètre proportionnel à l'expression du gène j dans l'échantillon i. Le vecteur β_j modélise les variations de l'expression du gène j en fonction des conditions expérimentales de chaque échantillon. Ces dernières sont résumées dans la matrice D (matrice de plan d'expérience, à n lignes). Les deux transformations sont :

— **la transformation VST** (*Variance stabilizing transformation*, en anglais) proposée par [10] : une transformation VST est une fonction dont l'objectif est d'obtenir des données

transformées, $\tilde{x} = f(x)$, telle que la variance des valeurs $\tilde{x}$ ne soit pas liée à leur moyenne. Il s'agit d'une transformation basée sur une réduction qui est définie par :

$$f(x) = \int_0^x \frac{1}{\sqrt{w(\lambda)}} d\lambda$$

où $w(\lambda)$ est la dépendance entre la variance et la moyenne, estimée par une approche paramétrique (modèle linéaire généralisé) ou non paramétrique (régressions locales) implémentées dans **DESeq2**. Cette transformation est appliquée sur les données de comptage normalisées (voir section 2.1.3 pour les méthodes de normalisation) ;

— **la transformation rlog** (*regularized logarithmic transformation*, en anglais) proposée par [139] : l'idée de cette méthode est de réduire les différences entre échantillons lorsque les comptages sont petits et de préserver les différences lorsque les comptages sont élevés. C'est une transformation basée sur une approche logarithmique qui donne des résultats similaires à une transformation $\log_2$ pour les comptages élevés et réduit les valeurs vers l'expression moyenne entre échantillons pour les gènes dont l'expression est faible.

Ces deux transformations utilisent la tendance expérimentale de la variance sur la moyenne afin de transformer les variables pour supprimer cette tendance.

La transformation VST est plus rapide que la transformation rlog. Cependant si les tailles de librairies des échantillons (et par conséquent les facteurs d'échelle) sont très hétérogènes, il est conseillé d'utiliser la transformation rlog. Ces transformations sont utiles pour visualiser les données afin de vérifier l'absence d'individus aberrants ou lorsque l'objectif est d'analyser les données à l'aide de méthodes de classification ou d'analyse linéaire discriminante.

Dans le contexte de l'analyse différentielle, [128] ont proposé une transformation, appelée voom. L'objectif est d'obtenir des données plus susceptibles d'être analysées par des méthodes basées sur des distributions gaussiennes (qui ont notamment été développées dans le cadre des puces à ADN). Comme les transformations VST et rlog, elle permet de stabiliser la variance et de supprimer le lien de dépendance entre la variance et la moyenne.

La transformation voom estime la relation variance/moyenne et génère des poids de précision pour chaque observation. La relation variance/moyenne est modélisée par une régression LOWESS et permet de donner un poids à chaque gène. Les poids obtenus sont alors incorporés dans la suite de l'analyse en utilisant les modèles linéaires créés pour l'analyse des puces à ADN.

Transformation pour la classification

L'objectif de la classification est de détecter des modules de gènes co-exprimés. Comme expliqué dans l'introduction de cette section, deux approches sont possibles :

— appliquer une transformation pour utiliser des modèles de mélange de lois gaussiennes ;

— utiliser des modèles de mélange de Poisson [169].

[88] ont proposé une transformation simple pour les données RNA-Seq permettant d'utiliser les modèles de mélange gaussien (qui sont des méthodes de classification bien établies dans le cas des données issues de puces à ADN [239]). Les données sont supposées être des réalisations d'un mélange de variables aléatoires suivant des lois gaussiennes (après

transformation) ou des lois de Poisson. Les travaux incluent une approche permettant de faire de la comparaison de modèles entre ces deux choix.

La transformation proposée est définie comme suit :

$$f(x_{ji}) = \log\left(\frac{x_{ji}/N_i + 1}{m_j + 1}\right) \qquad \text{avec } m_j = \frac{1}{n}\sum_{i'}^{n}\frac{x_{ji'}}{N_{i'}}$$

où N_i est la taille de librairie pour l'échantillon i et m_j correspond à l'expression moyenne du gène j à travers les n échantillons.

[168] proposent d'utiliser des méthodes de classification (k-means et modèle de mélange gaussien) sur des profils d'expression normalisés de données RNA-Seq. Les profils normalisés d'expression sont définis par :

$$p_{ij} = \frac{\frac{x_{ij}}{s_i} + 1}{\sum_l \frac{x_{lj}}{s_l} + 1}$$

où s_i correspond au facteur d'échelle permettant de normaliser les données (voir section 2.1.3). Les données p_{ij}, sont des données compositionnelles (i.e. dépendance linéaire des p_j) et il est nécessaire de les transformer avant d'utiliser des méthodes de classification basées sur des distributions gaussiennes. Selon la méthode utilisée (k-means ou modèle de mélange gaussien), [168] proposent diverses transformations. Par exemple, pour les modèles de mélange gaussien, des transformations classiques de type arcsinus ou logarithmique sont utilisées. Ces transformations sont disponibles dans le package **coseq**.

Transformation pour approcher une distribution de Poisson

Les données RNA-Seq sont des données de comptage. Il est donc naturel de vouloir utiliser des modèles basés sur des distributions de Poisson. Cependant, les données RNA-Seq sont surdispersées et ne respectent pas la propriété de données suivant une distribution de Poisson. La transformation puissance peut être utilisée afin de transformer légèrement les données pour que leur distribution s'approche de celle d'une distribution de Poisson et a l'avantage de la simplicité :

$$f(x) = x^{\alpha}$$

avec $\alpha \in]0, 1]$. Ce coefficient est choisi de façon à maximiser le critère d'adéquation entre la distribution des données transformées, x^{α} et une distribution de Poisson. Les exemples d'application de cette transformation pour les données RNA-Seq sont diverses : classification [233] ou encore l'inférence de réseau [6].

Le tableau 2.1 récapitule les différentes transformations et donne les packages R associés.

2.1.3 Normalisation

La normalisation est un processus destiné à identifier et supprimer des différences, dues à des biais techniques, entre les échantillons. Il est donc important de commencer par cette étape avant de chercher à analyser les données de séquençage. Le nombre de *reads* alignés pour un gène j et un échantillon donné est une mesure relative (non absolue) de l'expression du gène. Ce nombre dépend de la taille de la librairie N_i.

Transformation	Packages R
Transformations classiques	
Logarithmique, racine carrée et arcsin	disponible dans la version de base
Box-Cox	**MASS** [218], **bestNormalize** [161]
Blom	**RNOmni**
Stabilisation de la variance	
voom	**limma** [171]
VST	**DESeq2** [139]
rlog	**DESEq2**
Classification	
Transformation proposée par [88]	pas de package
Transformation pour profil d'expression	**coseq** [169]
Approcher une distribution de Poisson	
Transformation puissance	**PoiClaClu**

Tableau 2.1 Récapitulatif des packages R possibles pour les transformations.

Pour pouvoir comparer des expressions de gènes entre plusieurs réplicats, il convient de prendre en compte le nombre total de *reads* alignés pour chaque échantillon N_i. Un exemple de ce biais est illustré par le tableau 2.2 et la figure 2.1.

Comptages bruts

	gène 1	gène 2	gène 3	...	nombre total de *reads*
échantillon 1	77	61	120	...	1000
échantillon 2	157	122	244	...	2000

Comptages divisés par le nombre total de *reads* de l'échantillon

	gène 1	gène 2	gène 3	...	nombre total de *reads*
échantillon 1	0.077	0.061	0.12	...	1000
échantillon 2	0.0785	0.061	0.122	...	2000

Tableau 2.2 Influence de la taille de librairie N_i sur le nombre de *reads*.

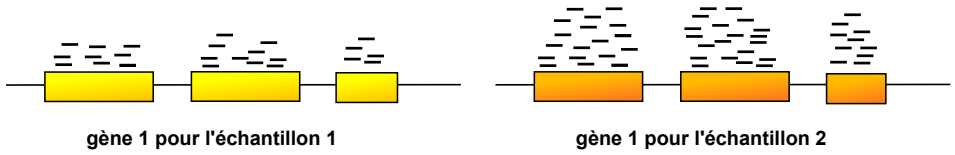


Figure 2.1 **Influence de la taille de librairie N_i sur le nombre de *reads*.** Ici, la profondeur de séquençage est plus élevée pour l'échantillon 2. Le gène 1 est donc exprimé de manière similaire dans ces deux échantillons.

Il existe un certain nombre de méthodes de normalisation pouvant être réparties en différentes familles : l'ajustement de la distribution, la prise en compte de la longueur du gène et le concept du nombre total de *reads* effectifs.

Le principe général des méthodes de normalisation est le même. L'objectif est de calculer un facteur de correction, C_i, pour chaque échantillon. Chaque comptage est alors multiplié par le facteur correctif correspondant à son échantillon. Les tailles de librairies pour les comptages normalisés sont alors approximativement égales. Généralement, les méthodes cherchent à définir un facteur d'échelle, s_i avec $\prod_{i=1}^{n} s_i = 1$, qui permet de définir une taille

de librairie corrigée, $\tilde{N}_i = s_i \times N_i$ à partir de laquelle les comptages normalisés (comptage par millions) sont obtenus avec :

$$\tilde{x}_{ij} = \frac{x_{ij}}{\tilde{N}_i} \times 10^6.$$

Ajustement des distributions

La méthode de normalisation naïve consiste à diviser chaque comptage d'un échantillon i par la taille de librairie de cet échantillon i, ce qui revient à choisir $s_i = 1$ pour tout i. Cette méthode s'appelle le comptage total [142].

Cependant, les tailles de librairies sont influencées par un nombre restreint de gènes fortement exprimés. Cela a pour conséquence des distributions très asymétriques. Pour remédier à ce problème, des variantes de cette méthode existent, comme la normalisation par la médiane ou par les quantiles. Par exemple, pour la méthode des quantiles, le facteur d'échelle est défini par :

$$s_i = \frac{Q_i^{(k)}}{\sqrt[n]{\prod_{l=1}^{n} Q_l^{(k)}}}$$

où $Q_i^{(k)}$ est le quantile choisi de la distribution des comptages dans l'échantillon i. Généralement le quantile choisi est le troisième quartile et cette normalisation est connue sous le nom « *upper-quartile* » [33].

Prise en compte de la longueur du gène

Pour comparer l'expression de différents gènes au sein d'un échantillon donné, il est important de corriger le biais induit par la longueur du gène sur laquelle les fragments sont alignés car, pour un même niveau d'expression, un long transcrit aura plus de chances d'être séquencé et donc plus de *reads* associés qu'un transcrit plus court. Un exemple de ce biais est illustré avec la figure 2.2.

Figure 2.2 **Influence de la longueur L_j des gènes sur le nombre de *reads*.** Le gène 2 est deux fois plus long que le gène 1 et produit par conséquent deux fois plus de fragments. Néanmoins, en terme de niveau d'expression, les deux gènes sont exprimés de la même façon.

Pour un gène j, la longueur, supposée connue, est notée L_j. Elle s'exprime en nombre de paires de base. La méthode de normalisation la plus simple consiste alors à diviser chaque comptage par la longueur du gène correspondant. Cette mesure ne prend cependant pas en compte la profondeur de séquençage. Pour cela, il faut opter pour la normalisation RPKM [2] [155] :

$$\text{RPKM}(x_{ij}) = \frac{x_{ij}}{(\frac{L_j}{10^3})(\frac{N_i}{10^6})}$$

Cette méthode de normalisation ne doit pas être utilisée lorsque l'objectif de l'analyse statistique est de rechercher des gènes différentiellement exprimés. Cette méthode permet certes d'obtenir des niveaux d'expression comparables entre gènes mais elle affecte la

2. *Reads Per Kilobase of exon per Million mapped Reads*

variabilité [159] et augmente le nombre de faux positifs dans l'analyse différentielle [67].

Concept du « nombre total de *reads* effectifs »

Le comptage total de *reads* dépend fortement d'un petit nombre de gènes fortement exprimés. De plus, la plupart des gènes ne sont pas différentiellement exprimés. Il est donc important de prendre en compte cette hypothèse dans la méthode de normalisation. C'est dans ce contexte que les méthodes de normalisations basées sur le concept du « nombre total de *reads* effectifs » ont été développées. Deux méthodes sont généralement utilisées pour prendre en compte ce concept : la méthode RLE [3] et la méthode TMM [4].

La méthode RLE [10] est disponible dans le package **DESeq2**. Les étapes pour obtenir le facteur d'échelle s_j sont les suivantes :

— un pseudo-échantillon, servant d'échantillon de référence, est créé pour lequel l'expression du gène j, R_j est définie comme la moyenne géométrique de l'expression de ce gène dans tous les échantillons :

$$R_j = (\Pi_{i=1}^{n} x_{ij})^{\frac{1}{n}} .$$

La moyenne géométrique est utilisée car elle est moins sensible aux valeurs extrêmes que la moyenne standard ;

— les comptages de tous les échantillons sont alors comparés à ceux de cet échantillon de référence : $\tilde{x}_{ij} = \frac{x_{ij}}{R_j}$;

— le facteur d'échelle est ensuite calculé comme :

$$s_i = \frac{\tilde{s}_i}{\exp(\frac{1}{n}\sum_{l=1}^{n} \log \tilde{s}_l)} \qquad \text{avec} \qquad \tilde{s}_i = \text{median}_j \left(\tilde{x}_{ij}\right).$$

La méthode TMM [175] est disponible dans le package **edgeR**. Cette méthode consiste à supprimer les valeurs extrêmes pour le log Fold-Change (M) et l'intensité moyenne en log (A) :

$$M_j(i,i') = \log_2\left(\frac{x_{ij}}{N_i}\right) - log_2\left(\frac{x_{i'j}}{N_{i'}}\right) \qquad \text{et} \qquad A_j(i,i') = \frac{1}{2}\left[\log_2\left(\frac{x_{ij}}{N_i}\right) + \log_2\left(\frac{x_{i'j}}{N_{i'}}\right)\right].$$

Pour la méthode RLE, un échantillon de référence est choisi parmi les échantillons disponibles $i = 1, \ldots, n$. Ce choix n'a pas de conséquence sur la suite de la méthode de normalisation mais généralement, l'échantillon i', dont le troisième quartile est le plus proche de la moyenne des troisièmes quartiles, est sélectionné comme échantillon de référence.

Après avoir filtré les données en supprimant les gènes avec des comptages nuls, les données sont filtrées en supprimant les valeurs les plus extrêmes. Ainsi, 30% des valeurs les plus extrêmes de M sont supprimées et 5% pour les valeurs de A. La moyenne pondérée des valeurs de M est alors calculée avec les échantillons restants :

$$\text{TMM}(i,i') = \frac{\sum_{j \in G*} w(i,i') M_j(i,i')}{\sum_{j \in G*} w_j(i,i')}$$

3. *Relative Log Expression*
4. *Trimmed Mean of M-values*

avec G^* l'ensemble des gènes qui n'ont pas été supprimés et

$$w_j(i,i') = \left(\frac{N_i - x_{ij}}{N_i x_{ij}} + \frac{N_{i'} - x_{i'j}}{N_{i'} x_{i'j}} \right).$$

Le facteur d'échelle est alors défini par :

$$s_i = \frac{\tilde{s}_i}{\exp\left(\frac{1}{n}\sum_{l=1}^{n} \log(\tilde{s}_l)\right)} \qquad \text{avec } \tilde{s}_i = 2^{\text{TMM}(i,i')}.$$

[67] ont comparé les différentes méthodes de normalisation dans le cadre d'une analyse différentielle. Ils mettent en évidence l'impact de la méthode de normalisation sur les résultats de l'analyse et formulent des recommandations pour le choix d'une méthode de normalisation appropriée en conseillant d'utiliser préférentiellement RLE ou TMM.

2.1.4 Analyse différentielle

L'objectif de l'analyse différentielle est de détecter les gènes différentiellement exprimés entre différentes conditions expérimentales.

La première étape consiste à définir les hypothèses à tester. Pour chaque gène j, l'analyse différentielle détermine si une différence d'expression est observée entre deux (ou plusieurs) conditions expérimentales. Pour cela, des tests d'hypothèses sont utilisés.

Soit x_{ij}^k l'expression du gène j pour l'échantillon i dans la condition k. La taille de librairie pour l'échantillon i dans la conditions k est notée s_i^k. Prenons un exemple avec deux conditions ($k = \{1, 2\}$). La question est de savoir si le gène j est plus exprimé dans une condition que dans l'autre. Le test d'hypothèse va donc chercher à déterminer s'il existe une différence entre λ_j^1, la moyenne d'expression du gène j dans la condition 1, et λ_j^2, la moyenne d'expression du gène j dans la condition 2 :

$$\mathcal{H}_{0j} = \{\lambda_j^1 = \lambda_j^2\} \qquad \text{contre} \qquad \mathcal{H}_{1j} = \{\lambda_j^1 \neq \lambda_j^2\}.$$

À partir des observations, une statistique de test est calculée pour chaque gène et est associée à une p-valeur. Si la p-valeur est inférieure au seuil fixé α (généralement 5%), l'hypothèse $\mathcal{H}_0$ est rejetée en faveur de l'hypothèse alternative $\mathcal{H}_1$.

Différentes approches existent pour réaliser ces tests, selon le nombre de conditions expérimentales et/ou la complexité du plan expérimental.

Modèles

La première idée naturelle est d'utiliser des tests de proportion standard ou le test de Fisher exact sur la table de contingence des données normalisées. Le test de Fisher n'estimant pas la variabilité des comptages, il a tendance à détecter un nombre important de faux positifs parmi les gènes fortement exprimés. Il n'est donc pas conseillé d'utiliser cette méthode pour l'analyse différentielle de données RNA-Seq.

Une approche alternative consiste à choisir une modélisation adéquate des données de comptage. Afin de prendre en compte la surdispersion, la distribution binomiale négative, décrite dans la section précédente, est souvent utilisée :

$$x_{ij}^k \sim NB(N_i^k \lambda_j^k, \phi_j)$$

avec N_i^k la taille de librairie (ou la taille de librairie corrigée) de l'échantillon i dans la condition k, λ_j^k la proportion de comptages pour le gène j dans la condition k et ϕ_j la dispersion du gène j. Cette dernière est supposée identique pour tous les échantillons. Les paramètres λ_j^k et ϕ_j sont alors estimés par maximum de vraisemblance avant de déduire une p-valeur associée aux données observées.

Une première approche pour le calcul de la p-valeur est le test exact pour la loi binomiale négative [176]. Les données sont normalisées afin d'obtenir des tailles de librairies semblables $N = s_i N_i$, ce qui implique que, pour chaque condition, k, la somme des comptages pour un gène j sur tous les échantillons suit la loi :

$$x_{1j}^k + \cdots + x_{n_k j}^k \sim NB(N\lambda_j^k, \phi_j / n_k).$$

Les paramètres λ_j^k et ϕ_j sont estimés et le test effectué ensuite est similaire au test de Fisher.

Pour l'estimation de ϕ_j, diverses approches, prenant en compte la faible taille d'échantillon, sont disponibles. Deux méthodes sont couramment utilisées pour le calcul des dispersions : celle proposée par [175], disponible dans le package **edgeR** et celle de [139] disponible dans le package **DESeq2**. Le package **DESeq2** modélise la tendance moyenne-variance afin d'estimer le paramètre de dispersion. Le package **edgeR** utilise un compromis entre une dispersion commune à tous les gènes et une dispersion spécifique à chaque gène.

Lorsque le nombre de conditions expérimentales est supérieur à deux ou lorsque le plan expérimental est plus complexe, une approche basée sur des modèles linéaires généralisés (GLM) est employée pour rechercher les gènes différentiellement exprimés. Des covariables, décrivant le plan expérimental, sont utilisées. Les données de comptage sont alors modélisées par un modèle GLM :

$$x_{ij} \sim NB(\mu_{ij}, \phi_j) \qquad \text{avec } \log(\mu_{ij}) = \log(\lambda_{ij}) + \log(N_i).$$

$\log(\lambda_{ij})$ est estimé par :

$$\log(\lambda_{ij}) = \lambda_0 + y_i^T \beta_j$$

où y est le vecteur des covariables utilisées pour décrire le plan expérimental. Les modèles GLM permettent de décomposer les effets au moyen de différents facteurs mais aussi de leurs interactions. Ces approches sont implémentées dans les packages **edgeR** et **DESeq2**.

Une dernière approche consiste à transformer les données avec la transformation voom [128] et à appliquer des méthodes d'analyse différentielle développées pour les données continues (par exemple les puces à ADN), c'est-à-dire, à utiliser des modèles linéaires gaussiens. Cette approche est disponible dans le package **limma** [171].

[103, 57] ont écrit de courtes revues résumant les différentes approches pour l'analyse différentielle des données d'expression mesurées par RNA-Seq et les packages permettant d'appliquer ces méthodes.

Correction pour tests multiples

Lors de l'analyse différentielle, de nombreux tests sont réalisés simultanément (un pour chaque gène). Cependant, le fait de multiplier les tests augmente le nombre de faux positifs, c'est-à-dire le nombre de cas dans lesquels l'hypothèse $\mathcal{H}_0$ est rejetée alors qu'elle est vraie. Par exemple, si 20 hypothèses indépendantes, toutes vraies, sont testées avec un risque de 5%, la probabilité de rejeter (à tord) au moins une de ces hypothèses est :

$$P_{\{\mathcal{H}_{0l}, l=1,\ldots,20\}}(\exists l \in \{1,\ldots,20\} : \mathcal{H}_{0l} \text{ est rejetée}) = 1-(1-\alpha)^{20} \approx 0.64.$$

Il est donc nécessaire de contrôler le risque de faux positifs (erreur de type I). Cette mesure est définie généralement soit par V le nombre de faux positifs, soit par Q la proportion de fausses découvertes définie par :

$$Q = \begin{cases} V/R \text{ si } R > 0 \\ 0 \text{ sinon} \end{cases}$$

où R correspond au nombre total d'hypothèses rejetées. Q correspond donc à la proportion de faux positifs parmi les hypothèses rejetées. Le tableau 2.3 résume le nombre des différentes erreurs possibles lors de la procédure de tests multiples.

	Hypothèse vraie	**Hypothèse fausse**	**Total**
Hypothèse rejetée	V	U	R
Hypothèse non rejetée	$m_0 - V$	$m_1 - V$	$m - R$
Total	m_0	m_1	m

Tableau 2.3 **Table de contingence pour les tests d'hypothèse multiples**. m correspond aux nombres d'hypothèses, m_0 le nombre d'hypothèses vraies, R le nombre d'hypothèses rejetées et V correspond au nombre d'erreur de type I (nombre de faux positifs).

Deux grandes familles de méthodes existent pour cela : le contrôle du FWER[5] et le contrôle du FDR[6].

La première méthode (FWER) consiste à calculer la probabilité d'avoir au moins un faux positif sur l'ensemble des comparaisons :

$$\text{FWER} = P(V > 0).$$

Ce type de méthode permet de majorer le risque de première espèce. Plus le nombre de gènes à tester est important, moins il y a de gènes déclarés différentiellement exprimés. Parmi cette famille de méthodes, la correction de Bonferroni [101] est la plus couramment utilisée.

Le second type de méthode cherche à contrôler la proportion attendue de faux positifs parmi les différences déclarées comme significatives :

$$\text{FDR} = E(Q).$$

5. *Family-Wise type I Error Rate*
6. *False Discovery Rate*

La méthode classiquement utilisée est celle proposée par [24]. Ces méthodes sont moins stringentes que celles de la famille FWER.

2.2 Inférence de réseaux de gènes

Un réseau $\mathcal{G}$ (ou graphe) est un objet mathématique utilisé pour modéliser les relations entre des entités. Dans sa forme la plus simple, il est composé de deux ensembles $\mathcal{G} = (V, E)$:

— l'ensemble $V = \{v_1, \ldots, v_p\}$ qui est un ensemble de p sommets (ou nœuds) représentant les entités étudiées ;

— l'ensemble E qui est un sous-ensemble de l'ensemble des paires de sommets, $E \subset \{(v_i, v_j, i, j = 1, \ldots, p, i \neq j\}$. Les paires de sommets dans E sont appelées arêtes du graphe. Elles modélisent un type donné de relations entre les entités.

La figure 2.3 permet d'illustrer un exemple simple de réseau. [126, 71, 231] sont les principaux ouvrages de référence sur les modèles graphiques.

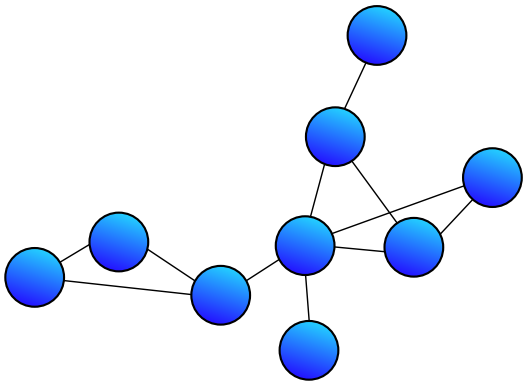


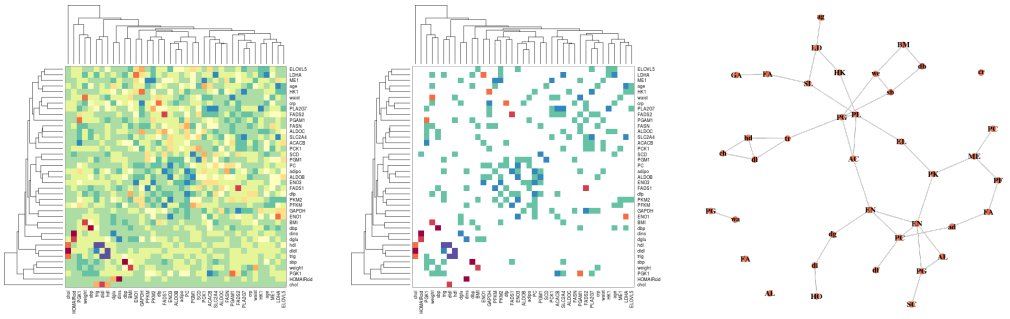
Figure 2.3 **Illustration d'un réseau.** Exemple d'un réseau avec 9 nœuds (cercles bleus) et 11 arêtes (lignes connectant deux nœuds).

Dans le cas d'inférence de réseaux de gènes, l'objectif est d'inférer un réseau à partir des données d'expression où les p gènes vont être représentés par les sommets du graphe. Les arêtes vont alors représenter un lien direct et fort (lien de régulation ou de co-expression) entre deux gènes.

2.2.1 Réseaux construits à partir des corrélations (*relevance networks*)

Une approche naïve [36, 35] est d'utiliser les corrélations entre les gènes pour déterminer les arêtes du réseau de gènes. L'approche, pour construire un tel réseau, se décompose en trois étapes, illustrées par la figure 2.4. Tout d'abord, les corrélations de Pearson (ou une autre similarité) sont calculées deux à deux entre les gènes. La seconde étape consiste à fixer un seuil à partir duquel l'expression est considérée comme négligeable.

Le réseau peut alors être construit : deux gènes sont liés par une arête si et seulement si la corrélation entre ces deux gènes est supérieure au seuil fixé.



matrice de corrélations entre les expressions | seuillage | réseau

Figure 2.4 **Étapes principales pour construire le réseau de corrélation.**
Source : wikistat `http://wikistat.fr/` « An introduction to network inference and mining ».

Bien que l'interprétation des arêtes soit facile, cette approche peut mener à une mauvaise interprétation biologique des relations existant entre les gènes, les liens directs et indirects étant confondus.

Illustrons ce problème par un exemple (figure 2.5). Soit trois gènes, notés x, y et z. L'expression du gène y et celle du gène z sont fortement régulées par un gène commun x. Les corrélations entre les gènes x et y et entre les gènes x et z sont donc élevées. Cependant, ces corrélations fortes ont pour conséquence une importante corrélation entre y et z. Bien qu'il n'y ait pas lien direct entre y et z, le modèle graphique avec des corrélations va pourtant construire une arête entre ces deux gènes. D'un point de vue biologique, ce type de modélisation n'est donc pas le plus pertinent.

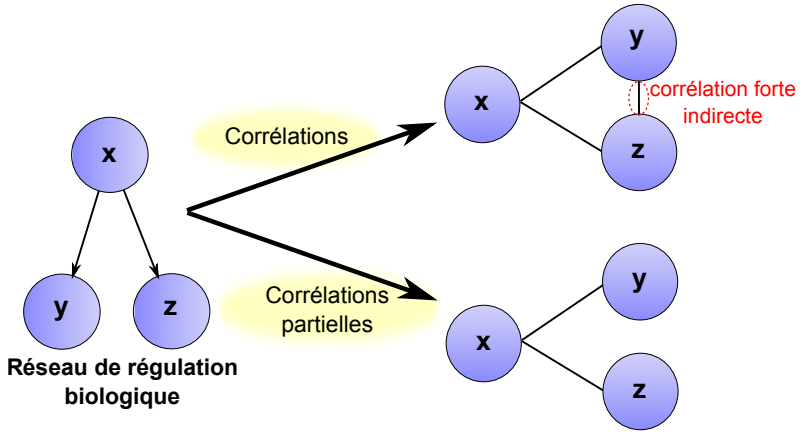


Figure 2.5 Limite de l'utilisation des corrélations pour l'inférence de réseau.

Pour éviter ce problème, on utilise comme mesure d'interaction entre variables les corrélations partielles : les corrélations entre deux gènes sont calculées sachant l'expression de tous les autres gènes. Cette approche permet de ne s'intéresser qu'aux interactions directes entre les variables. Deux nœuds sont reliés par une arête si et seulement si les deux variables associées aux nœuds sont dépendantes conditionnellement aux autres variables, c'est-à-dire si leur corrélation partielle sachant toutes les autres variables est non nulle. De tels réseaux peuvent être construits avec les modèles graphiques gaussiens.

2.2.2 Modèle graphique gaussien

L'objectif des modèles graphiques gaussiens (GGM) est de chercher à estimer le graphe des dépendances conditionnelles entre p variables (dans notre cas, il s'agit de gènes) à partir de n observations i.i.d., $(X_{ij})_{i=1,\dots,n}$, pour tout $j \in \{1,\dots,p\}$. Nous supposons que X suit une loi normale multivariée $\mathcal{N}(0,\Sigma)$ où Σ est la matrice de covariance de X, de taille $p \times p$, définie positive.

Les modèles GGM sont basés sur le résultat suivant, mis en avant par [64] : quel que soit le couple $(X_j, X_{j'})_{j \neq j'}$, les variables X_j et $X_{j'}$ sont indépendantes conditionnellement aux autres variables si et seulement si $\Sigma^{-1}_{jj'}$ est nulle. En effet, l'inverse de la matrice de covariance, $K = \Sigma^{-1}$, connue sous le nom de matrice de concentration (ou de précision), permet de décrire la structure de dépendances conditionnelles entre les variables [126] puisque chaque élément, $K_{jj'}, j \neq j'$, est directement lié aux coefficients de corrélations partielles $\pi_{j,j'} = \text{Cor}(X_j, X_{j'}|(X_k)_{k\neq j})$ via la relation suivante [231] :

$$\pi_{j,j'} = -\frac{K_{jj'}}{\sqrt{K_{jj}K_{j'j'}}}. \tag{2.1}$$

Une première approche pour inférer le réseau consiste donc à calculer S la matrice de variance covariance empirique :

$$S = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})^T$$

où $\bar{x}$ est le vecteur de moyenne empirique. Il suffit alors d'inverser cette matrice afin d'obtenir une estimation de K. Néanmoins, lorsque le nombre de variables est équivalent ou supérieur au nombre d'observations ($p \sim n$ ou $p > n$), cette approche n'est plus possible puisque la matrice S est alors soit mal conditionnée, soit non inversible.

Une première proposition pour aborder ce problème a consisté à stabiliser l'inversion par régularisation ridge. [186, 185] ont par exemple proposé de rendre la matrice K plus stable par inversion en l'estimant par :

$$\hat{K} = \left(\hat{\Sigma} + \rho I\right)^{-1}$$

(avec $\rho \in R^+$) et optionnellement, de combiner cette approche avec du bootstrap. Cette méthode produit une matrice K avec seulement des valeurs non nulles. Le problème consiste alors à décider quelles valeurs sont « significativement » non nulles. [186] proposent de sélectionner les arêtes en utilisant un test statistique basé sur un modèle bayésien avec une correction pour tests multiples. Ce modèle est implémenté dans le package R **GeneNet**.

La méthode précédente est une méthode en deux étapes qui estime, dans un premier temps, les corrélations partielles et dans un second temps, sélectionne les arêtes considérées comme les plus significatives. Une autre approche basée sur des modèles linéaires existe. En effet, les corrélations partielles sont liées à l'estimation des modèles linéaires suivants :

$$X_j = \sum_{j' \neq j} \beta_{j'j} X_{j'} + \epsilon_j \tag{2.2}$$

par la relation :

$$\beta_{jj'} = -\frac{K_{jj'}}{K_{jj}}.$$

En combinant avec l'équation (2.1), les deux équations ci-dessus montrent que les coefficients non nuls des modèles linéaires (équation (2.2)), correspondent aux corrélations partielles non nulles.

[143] ont été les premiers à proposer d'ajouter une pénalisation de type lasso [211] dans les modèles 2.2, permettant à la fois de résoudre le problème de grande dimension et de sélectionner les arêtes. Leur approche, connue sous le nom de sélection de voisinage [7], consiste donc à résoudre p problèmes de régressions pénalisées, où p correspond aux nombres de variables :

$$\min_{\beta \in \mathbb{R}^{p-1}} \frac{1}{n} \|X_j - X_{-j}\beta\|_2^2 + \rho\|\beta\|_1 \tag{2.3}$$

Un inconvénient majeur de cette approche est la nécessité d'ajouter une étape de symétrisation pour la matrice estimée obtenue. En effet, il est possible que le coefficient de régression de $X_{j'}$ sur X_j soit nul tandis que celui de X_j sur $X_{j'}$ soit non nul (ou inversement). Pour résoudre ce problème, [143] ont proposé deux règles :

— « *AND* » : si les deux coefficients ($\beta_{jj'}$ et $\beta_{j'j}$) sont non nuls alors une arête est construite entre les deux sommets ;

— « *OR* » : si au moins un des deux coefficients ($\beta_{jj'}$ ou $\beta_{j'j}$) est non nul alors une arête est construite entre les deux sommets.

D'autres auteurs [240, 18, 85] ont abordé le problème en le visualisant comme un problème de maximum de vraisemblance pénalisé. Au lieu de considérer p problèmes de régressions différents, ils proposent de pénaliser directement les éléments de la matrice de concentration avec une pénalisation de type lasso. Le problème est alors de maximiser la log-vraisemblance :

$$\mathcal{L}(K|X) = \log(\det(K)) - \text{Trace}(SK) - \rho||K||_1. \tag{2.4}$$

Pour résoudre ce problème d'estimation complexe, [18] ont proposé une approche basée sur un algorithme de descente par bloc. [85] ont revisité l'approche de [18] en combinant leur modèle de descente par bloc avec une seconde méthode de descente par coordonnées. Cet algorithme, appelé lasso graphique (*graphical lasso*, en anglais), est implémenté dans divers packages R : **glasso**, **huge** [244].

Notons que résoudre les p problèmes de régression (2.3) décrits par [143] est vu comme une approximation du problème exact (2.4) [18, 9].

Le modèle GGM suppose que les données suivent une loi normale multivariée. Il est donc généralement utilisé pour inférer des réseaux à partir de données d'expression continues (par exemple celles obtenues avec des puces à ADN). En revanche, ce modèle ne peut pas être utilisé directement sur des données discrètes, comme les données RNA-Seq. Pour pouvoir inférer un réseau à partir de telles données, il faut transformer les données afin que leur distribution suive une loi normale multivariée (voir section 2.1.2).

Une autre stratégie consiste à prendre en compte le caractère discret des données RNA-Seq en utilisant des modèles basés sur des lois de Poisson.

7. *neighbourhood selection*, en anglais

2.2.3 Modèle graphique log-linéaire Poisson

[26] a introduit le modèle graphique de Poisson (PGM). Il a montré que pour garantir une distribution jointe cohérente, il est nécessaire d'imposer des contraintes sévères sur les dépendances modélisées. Le modèle PGM ne peut que capturer les dépendances négatives. [238] ont alors proposé des variantes du modèle PGM de [26] mais aucun n'était complètement satisfaisant. En effet, les modèles proposés manquaient d'avoir des distributions de Poisson conditionnelles ou marginales. [6] ont alors proposé un modèle local PGM : le modèle graphique log-linéaire de Poisson (LLGM).

Étant basé sur une distribution de Poisson, le modèle LLGM suppose que la moyenne des données soit égale à la variance. Cependant, une des particularités des données RNA-Seq est leur surdispersion ; c'est-à-dire que la variance des comptages pour un échantillon est généralement plus grande que la moyenne. Ainsi pour pouvoir appliquer ce modèle à des données RNA-Seq, il est nécessaire de corriger au préalable cette surdispersion. Pour cela, la première étape consiste à utiliser une transformation puissance (section 2.1.2) sur les données : $z_{ij} = x_{ij}^{\alpha}$ avec $\alpha \in]0, 1]$.

Soit $Z_j = (x_{1j}^{\alpha}, \ldots, x_{nj}^{\alpha})$ le vecteur des données transformées du gène j pour les n échantillons. On suppose que la distribution conditionnelle Z_{ij} sachant tous les autres gènes $z_{i(-j)} = (z_{i,1}, \ldots, z_{i,j-1}), z_{i,j+1}, \ldots, z_{i,n})$ est une distribution de Poisson de paramètre λ_j avec $\log(\lambda_j)$ modélisé par une régression linéaire sur tous les autres gènes :

$$p(Z_{ij}|z_{i(-j)}) \sim P(\lambda_j)$$

avec

$$\log(\lambda_j) = \sum_{j \neq j'} \beta_{jj'} \widetilde{z}_{ij'}$$

où $\widetilde{z}$ correspond aux données log-transformées et réduites.

Une arête est présente dans le réseau inféré si les deux paramètres $\beta_{jj'}$ et $\beta_{j'j}$ sont non nuls.

La log-vraisemblance de chaque gène j s'écrit alors :

$$L(\beta) = \sum_{i=1}^{n} \left[z_{ij} \exp\left(\sum_{j \neq j'} \beta_{jj'} \widetilde{z}_{ij'} \right) - \sum_{j \neq j'} \beta jj' \widetilde{z}_{ij'} \right]. \tag{2.5}$$

Le vecteur β_j est supposé parcimonieux. Une pénalité lasso est donc ajoutée à la log-vraisemblance 2.5. Ainsi, de nombreux coefficients dans le vecteur β_j estimé sont mis à 0. Les paramètres β_j sont estimés par un algorithme de gradient coordiné. Comme suggéré par [6], le paramètre de régularisation est choisi avec la méthode StARS (*Stability Approach to regularization Selection criterion*) [135].

Ce modèle est implémenté dans deux packages R : le package **XMRF** [228] et le package **RNAseqNet** [108].

[6] ont donc proposé d'utiliser un modèle local afin de supprimer les contraintes sur les dépendances. Bien que cette approche permette d'estimer une structure de réseau plus général que le modèle PGM, il ne permet pas d'avoir un modèle graphique joint cohérent global.

Les technologies de séquençage à haut débit étant relativement récentes, les modèles d'inférence de réseaux adaptés aux données de comptage RNA-Seq sont encore en développement. En effet, des développements théoriques et méthodologiques sont

nécessaires pour pouvoir travailler sur un modèle graphique de Poisson joint proprement défini et n'ayant pas les contraintes actuelles du modèle actuellement proposé. De nouveaux modèles ont alors été proposés au cours des dernières années.

2.2.4 Autres modèles graphiques pour données de comptage

[50] proposent un modèle graphique de Poisson pénalisé pour des données de comptage comprenant un nombre important de comptages nuls. Le réseau est construit à partir d'un modèle de Poisson spatial à inflation de zéros et pénalisé. Un algorithme EM construit sur une descente par coordonnées est utilisé pour estimer les paramètres. Des résultats sur données simulées montrent que ce modèle est plus performant que le modèle LLGM en présence de données comprenant un nombre important de comptages nuls.

De nombreux modèles se sont développés afin de ne pas transformer les données au préalable et de prendre en compte la surdispersion directement dans le modèle. Tous ces modèles sont basés sur un modèle log-normal Poisson. Ils différent dans la formulation du problème et/ou dans sa résolution.

[87] ont été les premiers à proposer un modèle hiérarchique log-normal Poisson ne nécessitant aucune transformation au préalable. À partir des travaux de [143] et [6], ils ont proposé de remplacer la régression généralisée de Poisson par une régression généralisée mixte, basée sur une loi de Poisson hiérarchique. L'expression du gène j pour l'échantillon i, $i \in \{1, \dots, n\}$ est modélisée par $X_{ij} \sim P(\lambda_{ij})$ avec :

$$\log(\lambda_{ij}) = \sum_{j' \neq j} \beta_{jj'} \widetilde{x}_{ij'} + \epsilon_{ij},$$

$$\epsilon_j = (\epsilon_{1j}, \dots, \epsilon nj) \sim N(0, \sigma_j^2 I_n)$$

où $\widetilde{x}$ correspond aux données log-transformées et réduites. Comme pour [143], la matrice obtenue est rendue symétrique soit avec la règle « *AND* », soit avec la règle « *OR* ». Une pénalité de type lasso est ajoutée à la vraisemblance. Les paramètres sont estimés en utilisant une approximation de Laplace de la vraisemblance pénalisée et un algorithme de descente par coordonnées. Le paramètre de régularisation ρ est ici sélectionné en maximisant le critère BIC (voir section 2.2.5).

[51] proposent un modèle log-normal de Poisson. Comme pour le modèle proposé par [87], les données suivent une distribution de Poisson $\mathcal{P}(\lambda_{ij})$ mais avec $\log(\lambda_i)j \sim \mathcal{N}(\mu, \Sigma), i = 1, \dots, n$. Une pénalité de type lasso est ajoutée à la log-vraisemblance. Ils utilisent la méthode de Laplace pour approcher la vraisemblance et un algorithme ADMM [8] [29] pour obtenir l'estimateur du maximum de vraisemblance. Le paramètre de régularisation ρ est choisi via le critère eBIC.

[48, 49] proposent d'inférer le réseau en utilisant un modèle log-normal Poisson défini par [2] pour modéliser les données de comptage :

$$Z_i \sim \mathcal{N}(0_p, \sigma)$$

$$Y_{ij}|Z_{ij} \sim \mathcal{P}\left(\exp(o_{ij} + x_i^T \beta_j + Z_{ij})\right)$$

8. *Alternating Directions Method of Multipliers*

où o est la matrice contenant les termes *offsets*, supposés connus et qui peuvent être utilisés, par exemple, pour corriger les différences de tailles de librairies, β correspond aux coefficients de régression et x_i aux covariables pour l'observation i. Les données de comptage sont donc modélisées à l'aide de distributions de Poisson qui sont conditionnelles à des variables latentes corrélées gaussiennes. La structure de dépendance est complètement capturée par les variables latentes gaussiennes Z_i. Contrairement au cadre des modèles graphiques gaussiens, l'optimisation de la log-vraisemblance pénalisée est généralement insoluble. Pour construire le réseau, [49] ont donc recours à une procédure d'inférence variationnelle. Le problème d'optimisation est résolu en alternant un algorithme de gradient (*gradient ascent*) pour estimer les paramètres variationnels et une étape de modèle graphique en utilisant l'algorithme graphique lasso défini par [85]. Le package R contenant le modèle est disponible sur un github : `https://github.com/jchiquet/PLNmodels`. Dans ce package, deux critères pour choisir le paramètre de régularisation sont disponibles : StARS et eBIC.

Enfin, [235] proposent un modèle log-normal multivarié de Poisson. Les données d'expression X suivent une distribution multivariée de Poisson. Ils proposent d'estimer les paramètres en utilisant un algorithme de Monte-Carlo EM (MCEM) qui permet d'estimer simultanément les coefficients de régression et l'inverse de la matrice de covariance. Comme pour les autres approches, une pénalité de type lasso est ajouté au modèle pour obtenir des résultats parcimonieux. Le critère eBIC est utilisé pour sélectionner le paramètre de régularisation.

2.2.5 Choix des paramètres

Le choix du paramètre ρ est important puisqu'il permet de contrôler le niveau de parcimonie du réseau. Des valeurs élevées de ρ ont tendance à inférer des réseaux pratiquement vides et de petites valeurs fournissent généralement des réseaux trop denses.

Diverses méthodes ont été proposées pour sélectionner une valeur optimale de ce paramètre de régularisation ρ. Les méthodes usuelles pour choisir le paramètre de régularisation sont le critère AIC [3], le critère BIC [191] et la validation croisée [203, 59]. Bien que ces méthodes aient de bonnes propriétés théoriques en « faible dimension », elles ne sont pas adaptées pour des problèmes en grande dimension. Par exemple, dans le cadre de problèmes de régression, [229] ont montré que la validation croisée sur-ajuste les données (sur-apprentissage). De même, les critères AIC et BIC ont tendance à avoir des résultats médiocres : ils ont tendance à sélectionner plus de variables que nécessaire [121] lorsque le nombre de variables est beaucoup plus important que le nombre d'échantillons.

Pour l'inférence de réseau, deux familles de méthodes sont utilisées. La première approche consiste à modifier les critères habituels afin de les adapter aux cas où $p > n$ ou $p \gg n$. La seconde approche se base sur un concept de stabilité [144, 135].

Sélection de modèle

BIC Certains modèles, comme celui proposé par [87], optent pour le critère BIC afin de choisir le paramètre de régularisation ρ. Dans son modèle, [87] proposent deux étapes pour choisir un ρ commun pour toutes les régressions gène par gène. Dans un premier temps, un paramètre ρ_j est choisi pour chaque gène j en utilisant le critère BIC (en maximisant la log-vraisemblance pénalisée pour le gène j avec le critère BIC).

Dans un second temps, un paramètre de régularisation unique est obtenu en calculant la moyenne des ρ_j : $\rho = \sum_{j=1}^{p} \rho_j / p$. Puisque BIC est un critère asymptotique, prendre la moyenne des paramètres de régularisation sur l'ensemble des régressions, aide à améliorer la performance de l'inférence de réseau.

eBIC, ≪ BIC étendu ≫ Pour pallier le problème de grande dimension, [44] proposent de modifier le critère BIC en proposant une nouvelle famille de BIC étendu. L'objectif est de pénaliser à la fois le nombre de paramètres inconnus et la complexité de l'espace du modèle. [83] ont alors adapté ce critère pour les modèles graphiques. Le critère eBIC correspond au critère BIC auquel un terme de pénalité a été ajouté :

$$\text{eBIC}_\gamma(E) = -2\mathcal{L}(\hat{\theta}(E)) + |E| \log(n) + 4|E|\gamma \log(p)$$

avec γ un hyperparamètre compris entre 0 et 1. Il est important de ne pas confondre l'hyperparamètre γ avec ρ le paramètre de régularisation du modèle graphique lasso. Cet hyperparamètre γ doit être choisi manuellement. Lorsqu'il vaut 0, le critère eBIC revient à calculer le critère BIC. Plus ce paramètre est élevé, plus la parcimonie du réseau est importante (c'est-à-dire des réseaux moins denses). [83] montre qu'une valeur de 0,5 pour γ reste un bon compromis entre le nombre de faux positifs et de faux négatifs.

RIC Dans le cadre des méthodes de régression, [140] a proposé un nouveau schéma de sélection qui permet de diminuer le biais de sélection avec une pénalité qui s'adapte à la dimension des données ainsi qu'à leur structure de corrélation. L'idée fondamentale derrière cette approche est la création de données de référence n'ayant aucune relation avec la variable réponse mais avec les mêmes caractéristiques que les données réelles. Ces données sont obtenues à partir de permutations (PIC [9]) ou de rotations (RIC [10]) des données réelles. Les données de référence sont combinées aux données réelles et l'algorithme de sélection est appliqué sur cette nouvelle matrice de données : il s'arrête lorsque la première variable du jeu de référence (considérée comme une variable de bruit) est sélectionnée.

Le critère RIC permet de choisir directement le meilleur paramètre de régularisation ρ en se basant sur des rotations aléatoires plutôt que de trouver le meilleur ρ sur tout le chemin de régularisation en utilisant des méthodes coûteuses en temps de calcul (validation croisée ou ré-échantillonnage).

Approches basées sur la stabilité

L'estimation de structure discrète comme la sélection de variables ou la modélisation graphique est difficile, notamment dans le cadre de la grande dimension. Une nouvelle approche pour la sélection de modèle, basée sur la notion de stabilité, a généré un intérêt croissant dans la littérature récente. Cette nouvelle approche se base sur des techniques comme le ré-échantillonnage ou le bootstrap pour augmenter la stabilité des algorithmes de sélection (et pour quantifier leur incertitude). Cette notion de stabilité a été introduite par [30] dans le contexte de la prédiction.

Le principe de ces approches, cherchant à identifier la structure stable, se base sur l'idée suivante : le même algorithme de sélection doit fournir des résultats similaires sur des jeux de données semblables.

Selection stability [144] généralisent le concept de stabilité et proposent une version adaptée à l'inférence de réseau. L'objectif de leur approche est de fournir un réseau parcimo-

9. *Permutated Inclusion Criterion*
10. *Rotation Information Criterion*

nieux et stable tout en contrôlant le nombre de faux positifs parmi les arêtes. Leur méthode diffère des modèles de sélection puisqu'elle cherche à estimer la probabilité de sélection des variables et non directement le paramètre de régularisation optimal.

Cette probabilité de sélection est estimée à l'aide de B exécutions de la méthode de régression pénalisée sur différents sous-échantillons des données. Les arêtes considérées comme stables seront souvent sélectionnées par les modèles parmi les divers sous-échantillons. Une arête peu stable sera plus sensible au ré-échantillonnage et ne sera sélectionnée que par peu de modèles.

Le résultat est obtenu sous la forme d'un chemin de stabilité (*stability path*, en anglais) représentant la probabilité de sélection en fonction de la valeur de ρ. Une caractéristique attractive de cette méthode est le contrôle de l'erreur qui est fourni en posant une borne supérieure sur le nombre attendu de faux positifs parmi les variables sélectionnées.

StARS [11] S'inspirant des travaux de [144], [135] proposent une nouvelle approche pour choisir le paramètre de régularisation ρ : le critère StARS. Contrairement à la méthode *stability selection* dont l'objectif est de limiter le nombre de faux positifs, le critère StARS cherche à inférer un réseau de telle sorte que le vrai réseau (inconnu) soit inclus dans le réseau obtenu, autrement dit à limiter le nombre de faux négatif.

À partir des données initiales, B sous-échantillons de taille $m < n$ sont créés ainsi qu'un vecteur Λ contenant les valeurs des paramètres de régularisation ρ. Un réseau $\Omega^{(b,\rho)}$ est inféré pour chaque sous-échantillon b et chaque valeur ρ de Λ. La fréquence d'inclusion de l'arête e, arête présente entre les gènes j et j', est calculée comme suit :

$$p_e^\rho = \sharp\{b : \Omega_{jj'}^{(b,\rho)} \neq 0\}/B$$

et sa variance vaut :

$$v_e^\rho = p_e^\rho(1 - p_e^\rho).$$

La stabilité $stab(\rho)$ du réseau se définit par :

$$stab(\rho) = 1 - 2\bar{v}^\rho$$

où $\bar{v}^\rho$ correspond à la moyenne de v_e^ρ. Le critère StARS sélectionne le plus petit ρ (réseau plus dense) tel que $stab(\rho) \geq 1 - 2\tau$. En se basant sur des résultats théoriques, [135] suggèrent d'utiliser $2\tau = 0.05$ et des sous-échantillons de taille $m = \lfloor 10\sqrt{n} \rfloor$.

11. *Stability Approach to regularization Selection criterion*

Chapitre 3

Contributions

Les travaux des cette thèse s'appuient sur les données du projet DiOGenes. L'objectif de ce projet est d'étudier les effets à long terme d'un régime hypocalorique chez des personnes obèses. Cette étude cherche à approfondir les connaissances sur les mécanismes biologiques du contrôle pondéral et des pathologies associées à l'obésité.

Pour les études cliniques cherchant à comprendre l'ensemble des mécanismes liés à une maladie (l'obésité dans notre cas), divers types de données peuvent être mesurées : des mesures phénotypiques, cliniques mais aussi des analyses transcriptomiques effectuées à partir de prélèvements sanguins et/ou de tissus humains (le tissu adipeux pour DiOGenes). Les données peuvent également être mesurées à différents temps clés de l'étude. Dans DiOGenes, des données cliniques, phénotypiques et transcriptomiques du tissu adipeux sont disponibles avant et après chacune des deux phases de l'étude. Obtenues à différents pas de temps, les données acquises sont donc volumineuses et complexes.

Un obstacle important survenant face à cette masse de données hétérogènes est de définir une façon appropriée pour les exploiter et les modéliser, tout en tenant compte de leur grande dimensionnalité, de leur hétérogénéité au niveau biologique (données acquises à différents niveaux de l'échelle du vivant et à divers moments d'une expérience) mais aussi au niveau de leur nature (données numériques, discrètes, continues, etc.). Un autre problème rencontré est la présence d'observations avec des valeurs incomplètes (dans un même ensemble de données) ou manquantes totalement (par exemple des individus présents seulement dans certains jeux de données).

L'objectif de cette thèse est donc de proposer des approches permettant d'intégrer des données hétérogènes complexes à partir de divers tableaux de tailles déséquilibrées afin de répondre à diverses questions biologiques telles que :

— quelles sont les interactions entre les différents ensembles de données, situées à divers niveaux de l'organisme ?

— comment ces interactions évoluent-elles au cours du protocole ?

— quels sont les mécanismes biologiques clés qui expliquent le succès (maintien de l'amélioration des paramètres métaboliques) ou l'échec de l'intervention nutritionnelle ?

Les approches proposées dans le cadre de cette thèse cherchent à prendre en compte les spécificités des données d'expression de gènes mesurées avec des techniques de séquençage à haut débit (RNA-Seq et QuantSeq). Pour cela, les travaux présentés utilisent des approches permettant de prendre en compte le caractère discret des données ainsi que leur surdispersion.

La thèse s'articule en trois grandes parties : la gestion des données manquantes, l'inférence de réseau en présence d'individus manquants et l'intégration de différents types de données (cliniques et transcriptomiques) en utilisant une approche basée sur de l'inférence de réseau.

3.1 Données manquantes

Les données manquantes sont fréquentes en recherche clinique. L'analyse des données manquantes constitue donc un point essentiel dans l'analyse des données. L'origine des données manquantes peut varier, ce qui conditionne à la fois l'impact et la manière de prendre en compte les données manquantes ainsi que la nature des biais qui peuvent en découler. La méthode standard pour gérer les données manquantes a été, pendant longtemps, de supprimer les individus présentant au moins une valeur non observée, ce qui conduisait à une perte d'information considérable et une perte de puissance. Des méthodes ont alors été développées pour gérer les données manquantes afin d'extraire le plus d'information possible des données et de limiter les biais dues aux données manquantes (et à l'imputation de ces données par des valeurs de remplacement).

Dans le chapitre 4, nous faisons une revue sur la gestion des données manquantes. Ce chapitre permet de présenter plus en détail comment appréhender les données manquantes, les visualiser et les différentes approches pouvant les prendre en compte.

Après avoir défini les notations et la typologie des données manquantes, nous présentons les différentes approches : les approches utilisant seulement les données observées, les approches de modélisation jointe, les méthodes d'imputation simple, les méthodes permettant d'évaluer l'incertitude liée à l'imputation et des approches plus spécifiques utilisées dans le cas où les données non observées sont manquantes de manière non aléatoire. Nous présentons également, quand cela est possible, des packages R dans lesquels sont implémentées les méthodes décrites dans cette revue.

Ce chapitre fait l'objet d'une publication dans le journal de la société française de statistique.

3.2 Inférence de réseau de gènes en présence d'individus manquants

3.2.1 Illustration du problème

Dans les études longitudinales, comme c'est le cas avec le projet DiOGenes, il est fréquent que certains individus ne soient pas observés pour certains pas de temps de l'étude. Or, dans l'optique d'évaluer l'évolution des réseaux de gènes inférés à partir de données d'expression de ces gènes au cours du temps, il est préférable d'inférer ces réseaux à tous les pas de temps à partir des mêmes individus. Deux configurations, illustrées par la figure 3.1, sont alors possibles :

— utiliser le plus d'information possible en inférant le réseau à un pas de temps donné avec tous les échantillons disponibles à ce pas de temps ;

— de manière plus comparable : ne garder que les échantillons communs entre deux pas de temps pour inférer les réseaux.

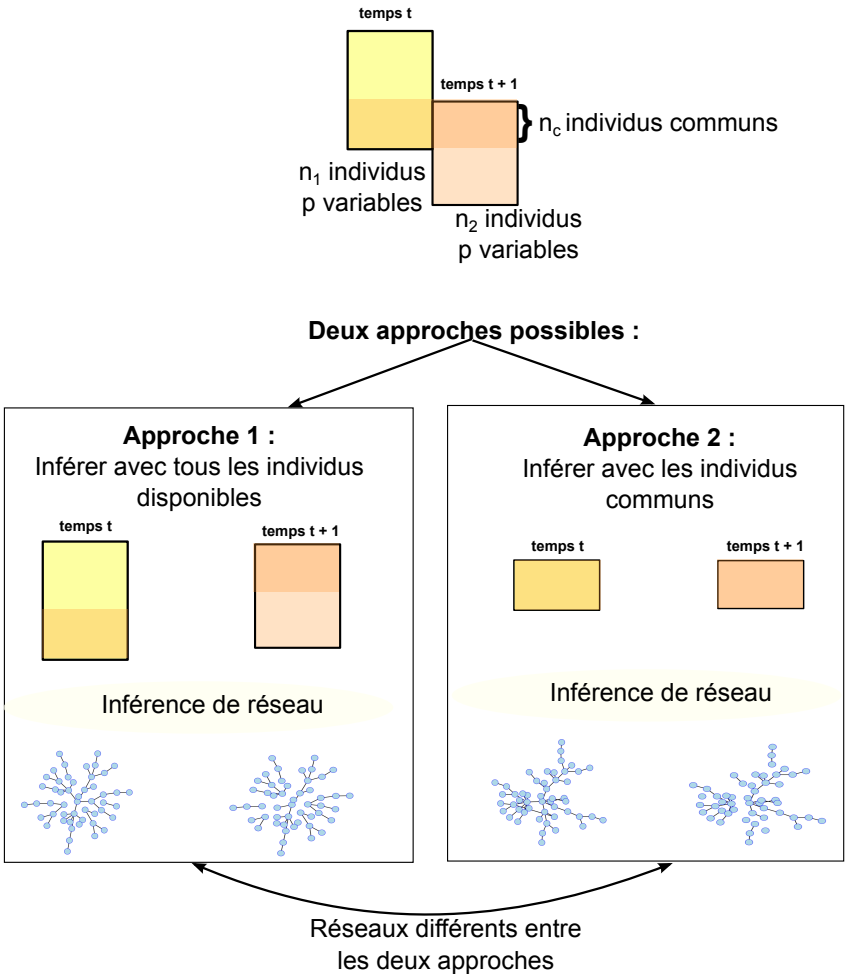


Figure 3.1 Choix des individus pour l'inférence de réseau entre deux pas de temps.

L'objectif est donc de trouver une solution pour limiter la perte d'information tout en ayant un nombre d'individus communs le plus important possible entre deux pas de temps.

Dans DiOGenes, l'expression des gènes dans le tissu adipeux a été mesurée par RNA-Seq. Les données obtenues sont des données de comptage. Elles sont discrètes et ont la particularité d'être également surdispersées. En outre, le coût associé à la collecte de ce type de données fait, qu'en général, le nombre d'observations collectées est souvent très faible devant le nombre de variables. Il est donc important d'utiliser un modèle graphique adapté aux caractéristiques de ces données. Nous avons choisi d'utiliser le modèle graphique log-linéaire de Poisson [6].

Dans cette partie, nous nous intéressons donc au problème d'individus manquants à certains pas de temps et à l'inférence de réseaux de gènes à partir de données d'expression de gènes mesurées par RNA-Seq.

3.2.2 Une méthode d'imputation pour le cas d'individus manquants dans le cadre de l'analyse de réseau

L'inférence de réseau est souvent sensible à la présence ou l'absence de certaines observations. Des travaux se sont intéressés à la question de la recherche de structures stables

(au sens de structures communes à la majorité des observations disponibles) dans l'inférence : [135] ont proposé une approche par ré-échantillonnage permettant de sélectionner un paramètre de régularisation maximum assurant la stabilité de l'inférence. [19] ont montré que l'inférence de réseau peut être sensible à quelques observations dites « influentes ». Ils ont alors proposé des mesures d'influence afin de filtrer les données pour supprimer les observations influentes et par conséquent stabiliser le réseau inféré.

Nous proposons ici une approche différente. En effet, il est possible de mesurer, simultanément aux données RNA-Seq, d'autres types de données d'expression de gènes. D'un coût moins élevé, ces données sont généralement disponibles pour un nombre plus important d'individus. Nous présentons ici une approche utilisant cette information supplémentaire au travers d'une nouvelle méthode d'imputation : l'imputation multiple hot-deck (hd-MI) Cette méthode d'imputation pour l'inférence de réseau à partir de données RNA-Seq est une méthode se basant sur deux méthodes d'imputation :

— l'imputation hot-deck qui permet à la fois d'imputer des individus en entier afin de ne pas détruire la structure de corrélation existant entre les variables et d'imposer des contraintes sur les données (par exemple la positivité) ;

— l'imputation multiple qui permet de mesurer l'incertitude liée à l'imputation.

L'objectif est d'imputer des individus manquants à certains pas de temps pour améliorer la qualité de l'inférence de réseau et pour pouvoir comparer les réseaux entre eux.

Les détails sur la méthode hd-MI ainsi que la comparaison avec d'autres méthodes d'imputation sur des données réelles sont présentés dans le chapitre 5. Ce chapitre a fait l'objet d'une publication dans le journal Bioinformatics [107] et de diverses communications orales dans des conférences nationales et internationales.

3.3 Intégration de données cliniques et transcriptomiques via une approche basée sur de l'inférence de réseau

3.3.1 Problématique

Dans les études cliniques, les données disponibles sont à la fois de grande dimension, de sources multiples et de nature hétérogène.

Dans un tel contexte, les besoins en méthodes intégratives sont de plus en plus pressants afin de considérer le système biologique dans son ensemble. Selon la question biologique et l'objectif, plusieurs types d'analyses peuvent être associés avec l'intégration de données hétérogènes. Nous nous intéressons ici à une approche basée sur de l'inférence de réseau à partir de données d'expression de gènes et de données cliniques.

Cependant, trouver les relations existant entre gènes et variables cliniques nécessite des approches spécifiques. Inférer directement un réseau du système biologique à partir de l'ensemble des données ne permet pas de trouver les liens entre gènes et variables cliniques.

En effet, les relations entre gènes et données cliniques sont masquées par les relations plus fortes existant entre les gènes.

3.3.2 Une analyse intégrative basée sur une approche réseau

Afin d'étudier les relations entre deux ensembles de données (variables cliniques et expression de gènes) pour chaque contraste, nous proposons ici de travailler avec les logarithmes des niveaux de changement d'expression (log Fold-Change, logFC) permettant d'introduire de l'appariement dans le modèle et d'utiliser une analyse de réseau basée sur un modèle graphique gaussien. Celle-ci sépare la construction du réseau d'intéractions global gènes/données cliniques en deux étapes :

— la première étape consiste à inférer un réseau de gènes en utilisant un modèle graphique gaussien sur les logFC des gènes ;

— la deuxième étape consiste à étudier les relations gènes/données cliniques par des modèles linéaires mixtes.

Le réseau du système global est alors obtenu en fusionnant les divers types de relations.

Cette approche intégrative est détaillée dans le chapitre 6. Elle est appliquée sur les données réelles de DiOGenes. Les deux ensembles de données utilisés sont des données transcriptomiques mesurées par QuantSeq [150], une nouvelle technique de séquençage à haut débit, et un ensemble de variables cliniques. Ce chapitre est une pré-publication qui sera prochainement soumise pour publication à un journal.

Données manquantes

Chapitre 4

Décrire, prendre en compte, imputer et évaluer les valeurs manquantes dans les études statistiques : une revue des approches existantes

Ce chapitre est un article publié dans le journal de la Société Française de Statistique.

4.1 Introduction

L'apparition de données manquantes est intimement liée à l'analyse statistique, au fait de collecter et préparer les données pour l'analyse statistique et elle a des origines multiples. Les données manquantes peuvent être la conséquence de non réponses (en sondages), de problèmes expérimentaux divers (en biologie), d'une mauvaise saisie de l'information ou de données aberrantes que l'on supprime après la première analyse exploratoire, ... La donnée manquante est parfois partielle [1] (pour un individu donné, seules quelques valeurs sont manquantes) ou bien totale [2] (toutes les variables d'un individu donné sont non observées).

L'objectif des méthodes permettant de traiter les données manquantes est multiple : il peut s'agir d'estimer les valeurs manquantes elles-mêmes, pour reconstituer une vision réaliste des données. Toutefois, dans de nombreux cas, les données contenant des valeurs manquantes sont utilisées pour des analyses statistiques de natures diverses : estimation d'un paramètre de la population dont sont tirées les données, analyses exploratoires (types ACP), modèles prédictifs... Dans ces divers cas, la manière d'aborder les données manquantes, en utilisant uniquement l'information disponible ou bien en tentant de reconstituer les données manquantes (imputation), doit tenir compte de l'objectif lui-même, afin de limiter la perte de précision dans les méthodes de prédiction ou bien les biais d'estimation dans les méthodes d'inférence.

[187], [7], [133], [189], [89], [20], [37] et [42] constituent les principaux ouvrages de référence sur les données manquantes. L'objectif de cet article est de proposer au lecteur une vision générale des divers problèmes liés aux données manquantes et des principales stratégies qui peuvent être mises en œuvre pour tenir compte de leur présence dans les analyses statistiques.

L'article est organisé comme suit : la section d'introduction présente les notations et la typologie usuelle des données manquantes. La section 4.2 présente les approches utilisant uniquement les données observées (c'est-à-dire, les méthodes qui ne recourent pas à l'imputation des données manquantes). La section 4.3 présente les approches de

1. *item non-response* en anglais.
2. *unit non-response* en anglais.

modélisation jointe principalement utilisées dans les problèmes d'inférence statistique. La section 4.4 présente les méthodes d'imputation simple qui permettent d'obtenir un tableau de données complet. La section 4.5, quant à elle, décrit les diverses approches permettant d'évaluer la qualité de l'imputation ou l'incertitude liée à l'imputation ou à la présence de valeurs manquantes dans les résultats de l'analyse statistique. Enfin, la section 4.6 décrit les approches plus spécifiquement dédiées au cas le plus complexe, celui dans lequel les données sont manquantes MNAR (c'est-à-dire, manquantes de manière non aléatoire). En complément, compte tenu de l'impact croissant de l'utilisation du logiciel R dans l'analyse statistique, nous nous attacherons, quand cela est possible, à présenter des packages dans lesquels les diverses méthodes décrites dans cette revue sont implémentées.

4.1.1 Notations

Soit un vecteur $Y = (Y_1, ..., Y_p)$ de p variables aléatoires numériques ou catégorielles. On notera y_{ij} l'observation de la variable Y_j pour un individu $i \in \{1, \ldots, n\}$, $\mathbf{y}_i = (y_{i1}, \ldots, y_{ip})$ le vecteur des observations des p variables de Y et $\mathbf{Y}$ la matrice des observations $(y_{ij})_{i=1,\ldots,n,\, j=1,\ldots,p}$ dont les lignes sont des observations i.i.d. de Y. Pour simplifier, on confondra la notation de la variable aléatoire Y_j et de son observation $Y_j = \begin{pmatrix} y_{1j} \\ \vdots \\ y_{nj} \end{pmatrix}$ sur les n individus.

On définit aussi la *matrice indicatrice des valeurs manquantes*, $\mathbf{R}$, dont les valeurs, $(r_{ij})_{i=1,\ldots,n,\, j=1,\ldots,p}$, sont :

$$r_{ij} = \begin{cases} 1 & \text{si } y_{ij} \text{ est observée} \\ 0 & \text{sinon} \end{cases}$$

et on note R la variable aléatoire associée. De manière similaire, Y_{obs} et Y_{miss} correspondent (respectivement) aux parties observées et manquantes de Y de telle sorte que $Y = RY_{\text{obs}} + (1 - R)Y_{\text{miss}}$.

Le *mécanisme de génération des données manquantes* est défini comme étant la distribution conditionnelle de R sachant Y, $f(R \mid Y)$ [133]. Ce mécanisme peut éventuellement dépendre de paramètres, notés ψ. Également, dans certains cas, des covariables $(X_j)_{j=1,\ldots,q}$ sont complètement observées sur tous les individus (on note alors x_{ij} l'observation de la covariable j pour l'individu i et X les variables aléatoires correspondantes). Dans ces cas plus complexes, le mécanisme de génération des données manquantes est alors noté $f(R \mid Y, X; \psi)$ ou $f(R|Y; \psi)$.

Enfin, quelques-unes des notions de cette revue seront illustrées sur des données de questionnaire, présentes dans le package R **naniar** et qui concernent une enquête annuelle produite en 2009 par le Behavioral Risk Factor Surveillance System (BRFSS) [3] destinée à évaluer les comportements à risque dans la population adulte aux États-Unis. Le jeu de données contient la mesure de 34 variables (État de résidence, sexe, âge, statut marital, grossesse, tabagisme...) pour 245 adultes de 18 ans et plus. Ces données contiennent un total de 1186 valeurs manquantes.

3. `https://www.cdc.gov/brfss/annual_data/annual_2009.htm`

4.1.2 Répartition des données manquantes

Pour décider de l'approche la plus judicieuse pour prendre en compte les valeurs manquantes dans l'analyse (suppression d'individus ou de variables, correction manuelle, imputation par prédiction, ...), il est recommandé de réaliser une analyse exploratoire permettant de comprendre la distribution des valeurs manquantes dans le jeu de données. [133] définissent trois types de répartition des données manquantes, illustrés par la figure 4.1 :

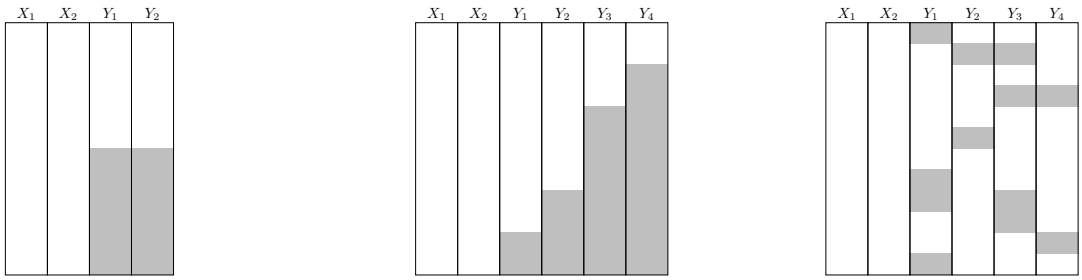


Figure 4.1 **Répartition des données manquantes, (a) univariée, (b) monotone et (c) sans structure.** Les zones grisées indiquent la position des données manquantes.

— la structure des valeurs manquantes est *univariée* (figure 4.1 (a)) si les mêmes individus ont des valeurs manquantes pour les mêmes $d < p$ variables ;

— les valeurs manquantes sont *monotones* (figure 4.1 (b)) si les variables peuvent être ordonnées de telle sorte que, lorsque l'observation y_{ij} est manquante pour la variable Y_j, alors toutes les variables suivantes pour ce même individu, $\{y_{ik}\}_{k>j}$, sont aussi manquantes. Ce cas est fréquemment rencontré dans les études longitudinales, particulièrement en épidémiologie (il peut correspondre, par exemple, à la sortie de l'étude d'un individu : on parle alors de données censurées) ;

— les valeurs manquantes sont *sans structure* (voir figure 4.1 (c)), si elles sont réparties sans structure particulière dans le jeu de données.

En outre, la quantité de données manquantes peut être définie de manière variée selon que l'on considère une proportion de manquants par rapport aux individus (lignes), aux variables (colonnes) ou bien aux valeurs elles-mêmes (entrées du tableau).

Comme souligné par [208] et [212], comprendre la répartition des valeurs manquantes dans le jeu de données permet d'adapter la stratégie de traitement de celles-ci, qu'il s'agisse d'exclure des variables ou individus (qui contiennent une fréquence de manquants trop importante), de collecter de nouvelles données, d'estimer ou de remplacer les valeurs manquantes (imputation). Pour aborder cette question, le package R **mi** [205] identifie les motifs identiques de valeurs manquantes entre paires de variables à la création du tableau de données avec la fonction `missing_data.frame` (voir figure 4.2).

Une autre manière standard d'explorer la répartition et la structure des valeurs manquantes est d'avoir recours à des graphiques diagnostiques, qui peuvent s'avérer particulièrement efficaces en raison de la capacité de l'œil humain à détecter facilement des motifs [212]. Le package R **VIM** ([208] et [123]) permet ce type d'analyse exploratoire et peut aider à identifier le mécanisme de génération des données manquantes (voir section suivante) ainsi qu'à déceler des anomalies ou des erreurs dans les données imputées

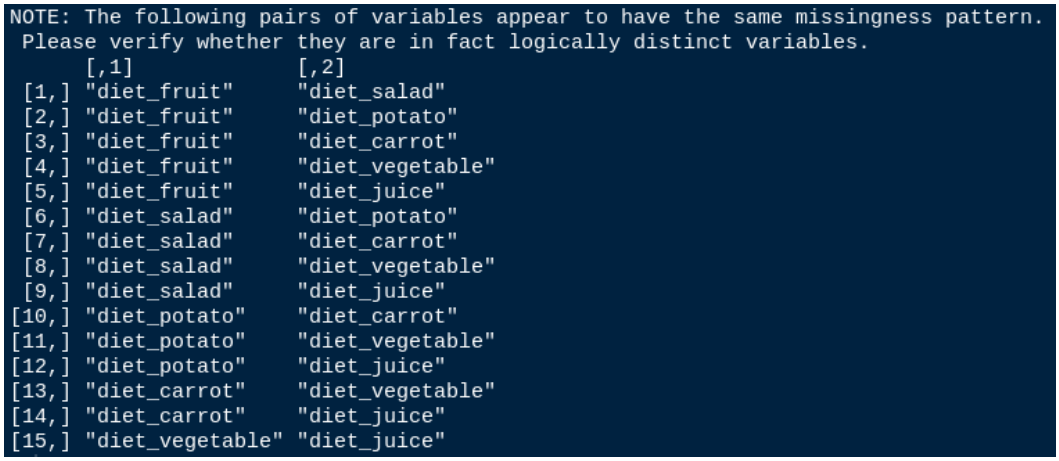


Figure 4.2 Message concernant les motifs de valeurs manquantes identiques entre diverses variables tel que fourni par le package **mi**.

(voir section 4.5.1). **VIM** contient, en outre, quelques méthodes d'imputation des données que nous décrirons dans les sections suivantes. Enfin, **VIM** peut être facilement utilisé au travers de l'interface graphique **VIMGUI**. Sur l'exemple décrit brièvement en section 4.1.1, la figure 4.3 montre le type de graphiques disponibles dans ce package : la répartition du nombre de valeurs manquantes par variable est visualisée par un diagramme en barres, les motifs et fréquences de ces motifs sont visualisés par un diagramme en grille et la relation entre les niveaux de valeurs des variables et les valeurs manquantes est disponible sous la forme d'un graphique en matrice (ordonné, dans cet exemple, selon la variable « age », en troisième colonne).

De manière similaire, le package **naniar** est dédié à la manipulation et la visualisation des données manquantes selon les principes développés dans la collection de packages « tidyverse »[4]. Parmi les graphiques disponibles dans ce package, on trouve un graphique en matrice permettant de visualiser la répartition des manquants et très similaire à celui du package **visdat** de visualisation de données. On trouve également un graphique en bâtons permettant de visualiser le nombre de valeurs manquantes par variable.

Dans l'exemple des figures 4.3 et 4.4, on peut, par exemple, identifier de manière immédiate que, si la plupart des variables sont renseignées pour presque tous les individus, quelques variables ont une forte proportion de valeurs manquantes (parmi lesquelles la variable indiquant si la personne est enceinte, « pregnant » ou la variable précisant la fréquence à laquelle la personne fume, « smoke_day »). Ces variables sont souvent manquantes simultanément. On observe également un groupe de variables qui sont manquantes de manière simultanée sur la droite des graphiques et qui correspondent aux variables décrivant les habitudes alimentaires, « diet_... », comme déjà identifié par le message de la figure 4.2 (ce sous-groupe présente donc une structure univariée). De même, le sous-groupe relatif aux habitudes de consommation d'alcool, « drink_... » a une structure monotone. Enfin, les valeurs manquantes de la variable « pregnant » sont clairement liées à la variable « age » (les personnes les plus âgées de l'échantillon ayant systématiquement un statut manquant pour la variable « pregnant »). Comme nous le verrons dans la section suivante, ces observations simples donnent des indices sur la nature du mécanisme des données manquantes et orientent l'utilisateur vers des manières de prendre en charge l'information manquante.

4. https://www.tidyverse.org/

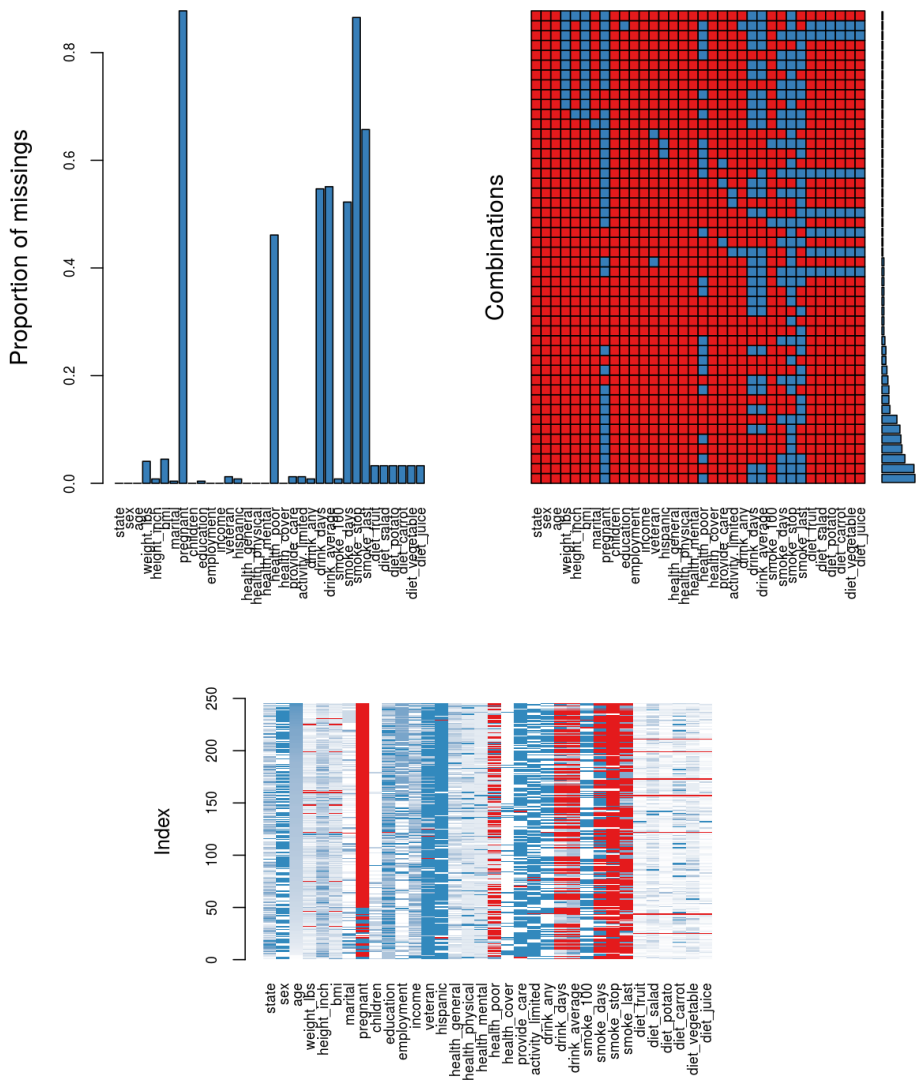


Figure 4.3 **Graphiques de visualisation de la distribution des valeurs manquantes disponibles dans VIM.** En haut à gauche : diagramme en barres du nombre de valeurs manquantes par variable. En haut à droite : diagramme en grille des motifs et fréquences de ces motifs. En bas : diagramme de la répartition des valeurs manquantes (en rouge) dans la distribution des valeurs de chaque variable (en niveaux de bleu) dans lequel les individus sont ordonnés selon la valeur de la variable « age ».

Enfin, notons que, si les deux packages précédents proposent des visualisations statiques de la répartition des données manquantes, [208] soulignent le très grand intérêt pratique, pour déceler des problèmes de collectes de données ou des motifs dans la distributions des valeurs manquantes, des représentations interactives. Le logiciel GGobi[5] [56], accessible dans R via le package **rggobi**, permet une telle visualisation. Des exemples d'utilisation des fonctionnalités d'interactivité, sous forme de vidéos, sont disponibles sur le site web associé au livre `http://www.ggobi.org/book/`. Elles illustrent, par exemple, comment le fait de pouvoir lier des graphiques différents à la souris permet d'explorer la distribution des valeurs manquantes ou bien comment visualiser les effets de l'imputation sur la distribution des variables.

5. `http://www.ggobi.org/`

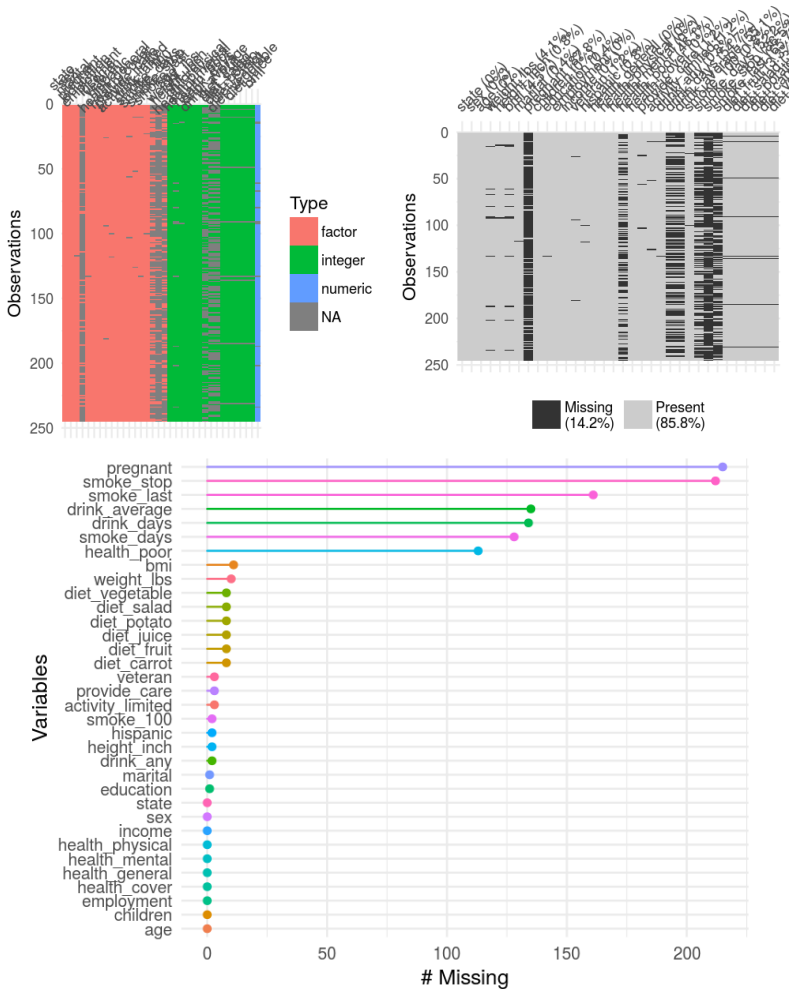


Figure 4.4 **Graphiques de visualisation de la distribution des valeurs manquantes disponibles dans visdat (en haut à droite) et dans naniar.** En haut : diagrammes de la répartition des valeurs manquantes. En bas : diagramme en bâtons du nombre de manquants par variable.

4.1.3 Mécanisme de génération des données manquantes

Au-delà du simple aspect descriptif de la répartition des données manquantes, il est souvent nécessaire d'appréhender la loi de probabilité à l'origine des données manquantes (càd le mécanisme de génération des données manquantes). La connaissance de ce mécanisme (ou plutôt de son type) est, en effet, une hypothèse standard des garanties théoriques qui existent pour certaines méthodes qui prennent en compte les valeurs manquantes, comme nous le verrons dans les sections suivantes.

Typologie générale

[133] définissent une typologie générale des données manquantes en trois catégories qui dépendent de la relation statistique entre les données et le mécanisme de génération des données manquantes. Les définitions suivantes sont données dans le cas où il n'y a pas de covariables complètement observées, X, pour alléger les notations, mais s'étendent de manière triviale au cas où elles sont présentes.

— **Données manquantes complètement aléatoirement ou MCAR** [6]
Les données sont *manquantes complètement aléatoirement* si la probabilité d'absence est la même pour toutes les observations. Cette probabilité ne dépend que des paramètres extérieurs indépendants de cette variable. De manière formelle, ce cas est défini par :

$$f(R|Y, X; \psi) = f(R; \psi).$$

Dans ce cas-ci, les données manquantes sont nécessairement sans structure. Un exemple typique de données MCAR est le cas où une personne oublie par accident de répondre à une question lors d'une enquête. Les données manquantes des variables présentes au centre du tableau de la figure 4.4 (en haut à droite) pourraient être de ce type (par exemple, les variables niveau d'éducation, « education » et statut vis-à-vis du service militaire « veteran ») : elles présentent peu de manquants, pour lesquels on ne décèle, de manière visible, aucune relation avec les valeurs ou le statut des autres variables.

— **Données manquantes aléatoirement ou MAR** [7]
Le cas des données manquantes complètement aléatoirement est rare : si la probabilité d'absence est liée à une ou plusieurs variables observées, les données manquantes sont dites *données manquantes aléatoirement*. De manière formelle, ce cas est défini par :

$$f(R|Y, X; \psi) = f(R|Y_{\text{obs}}, X; \psi).$$

Dans l'exemple introduit dans la figure 4.3 (bas), le couple $(\text{age}, \text{pregnant})$ pourrait constituer un exemple de données MAR : les valeurs manquantes de la variable « pregnant » sont liées de manière visible à la variable « age » de l'individu, qui est complètement observée.

— **Données manquantes non aléatoirement ou MNAR** [8]
Enfin, le dernier cas est de données *manquantes de façon non aléatoire* se présente lorsque la probabilité d'absence d'une variable dépend de la variable elle-même ou d'autres variables non observées. De manière formelle, ce cas est défini par :

$$f(R|Y, X; \psi) = f(R|Y_{\text{obs}}, Y_{\text{miss}}, X; \psi).$$

Ce type de données manquantes est plus complexe à traiter. Il peut être abordé par analyse de sensibilité (voir section 4.6 pour des détails sur le traitement spécifique de ce type de données manquantes). Un exemple typique de ce type de données manquantes est le cas de questions sensibles dans un questionnaire où le niveau de non-réponse dépend de la réponse elle-même. Dans les données de l'exemple précédent, on peut suspecter, par exemple, une plus grande propension des gros fumeurs ou des gros consommateurs d'alcool à ne pas répondre (variables « smoke_... » et « drink_... »).

Notons que les exemples donnés ne sont fondés que sur des hypothèses liées à l'observation de la distribution des variables. Dans le cas des variables $(\text{age}, \text{pregnant})$, on peut aussi imaginer que les données sont MNAR si le statut de la variable « pregnant » est lui-même lié à la présence de manquants sur cette variable (les valeurs négatives de « pregnant » étant, par exemple, plus fréquemment non collectées) et que l'observation d'un lien entre âge et statut manquant de « pregnant » est lié à une dépendance (qui existe de manière évidente)

6. *Missing Completely At Random* en anglais.
7. *Missing At Random* en anglais.
8. *Missing Not At Random*

entre ces deux variables. Même dans le cas de la variable « education », il est impossible de distinguer une potentielle absence MCAR du cas où toutes les valeurs manquantes de cette variable correspondent, par exemple, à une même modalité de la variable (« *n'est jamais allé à l'école ou seulement à l'école maternelle* », par exemple), qui correspondrait à un cas MNAR.

Pourquoi s'intéresser aux valeurs manquantes ?

Une approche naïve, en présence de données manquantes, est d'analyser les données en utilisant uniquement les observations disponibles. Prenons, par exemple, le cas simple de l'inférence statistique, dans lequel on chercherait à estimer l'espérance de Y_1, $\mu_1 = \mathbb{E}(Y_1)$. Dans ce cas, l'estimateur habituel de μ_1 est $\hat{\mu}_1 = \frac{1}{n}\sum_{i=1}^{n} y_{i1}$ qui est sans biais ($\mathbb{E}(\hat{\mu}_1) = \mu_1$) mais n'est pas nécessairement observé (si certaines valeurs de la variable Y_1 sont manquantes). Remplacer cet estimateur par $\widetilde{\mu_1} = \frac{1}{n_1}\sum_{i=1}^{n} r_{i1}y_{i1}$ avec $n_1 = \sum_{i=1}^{n} r_{i1}$ le nombre de valeurs observées pour Y_1 a des conséquences variées selon le type de mécanisme des données manquantes :

— si les données manquantes sont MCAR, R et Y sont indépendantes et $\widetilde{\mu}_1$ est donc aussi un estimateur sans biais de μ_1. Toutefois, cet estimateur est obtenu avec $n_1 < n$ observations et il en résulte une perte de précision de l'intervalle de confiance autour de μ_1 ou (dans le cas de tests statistiques) une perte de puissance ;

— si les données manquantes sont MAR ou MNAR, R et Y ne sont plus indépendantes. Cela peut être le cas, par exemple, si l'observation de Y_1 est liée à la variable Y_2 comme suit :

$$R_1 = \begin{cases} 0 & \text{si } Y_2 \leq a \\ 1 & \text{sinon.} \end{cases}$$

pour un $a \in \mathbb{R}$, fixé. Dans ce cas,

$$\mathbb{E}(\widetilde{\mu}_1) = \mathbb{E}\left(Y_1 \mathbf{1}_{\{Y_2 > a\}}\right)$$

ce qui résulte en un biais de $\mathbb{E}\left(Y_1 \mathbf{1}_{\{Y_2 > a\}}\right)$ dans l'estimation de μ_1. La différence entre le cas MAR et le cas MNAR réside dans la dépendance de R aux données non observées. Dans l'exemple précédent, si Y_2 est complètement observée, le mécanisme de génération des données est MAR.

Identification et utilisation de la typologie des valeurs manquantes

Il est donc important de connaître le type de données manquantes pour éviter les erreurs conduisant à des biais d'analyse dans leur prise en compte. Un test statistique, permettant de tester l'hypothèse selon laquelle les données manquantes sont MCAR contre MAR, est décrit dans [130]. Il est fondé sur une statistique de test qui suit une loi du χ^2. Le test fait l'hypothèse d'une distribution gaussienne, $\mathcal{N}(\mu, \Sigma)$ de Y et son principe est de grouper les individus en K sous-groupes de profils de valeurs manquantes distincts, $\mathcal{C}_k$ ($k = 1, \ldots, K$). Si les données manquantes sont MCAR, la statistique de test proposée et fondée sur le calcul des moyennes et variances conditionnelles aux K groupes de profils, a une distribution asymptotique suivant une loi du χ^2.

Ce test est implémenté dans la fonction `LittleMCAR` du package R **BaylorEdPsych**. S'il permet de tester l'hypothèse MCAR, il n'indique pas, en revanche, quelles variables ne sont pas MCAR. Comme le test est fondé sur une distribution asymptotique, son efficacité

est fortement conditionnée à la taille de l'échantillon. Lorsque le nombre d'individus est trop faible ou que l'hypothèse de distribution gaussienne n'est pas réaliste, [109] ont proposé un test non paramétrique. Ce test est disponible dans le package R **missMech** [110]. En revanche, comme souligné dans [37], il n'existe pas de test de l'hypothèse MAR contre l'hypothèse MNAR car l'information qui serait nécessaire pour réaliser un tel test est, justement, l'information manquante.

Par ailleurs, lorsque les données sont manquantes MAR, [181] décrit les conditions minimales requises qui permettent d'ignorer le processus de génération des données manquantes dans l'inférence statistique (le processus de génération des données manquantes est alors dit « ignorable »). Pour cela, les données doivent être manquantes aléatoirement (cas MAR et MCAR) et les paramètres régissant le mécanisme de génération des données manquantes et des données doivent être « distinguables » : cela signifie que les paramètres du modèle de génération des données, ϕ, peuvent s'écrire $\phi = (\psi, \theta)$ où ψ désigne les paramètres qui régissent la distribution de R et où θ sont les paramètres qui régissent celle de Y. Ces paramètres sont distinguables lorsqu'ils vivent dans des espaces en produits cartésiens. Dans ce cas, lorsque les données manquantes sont MAR, il est possible de factoriser la densité des données observées de la façon suivante :

$$f(Y_{\text{obs}}, R; \theta, \psi) = f(R|Y_{\text{obs}}; \psi) \times \int f(Y; \theta) dY_{\text{miss}} = f(R|Y_{\text{obs}}; \psi) f(Y_{\text{obs}}; \theta), \tag{4.1}$$

et la vraisemblance des données observées est donc proportionnelle à la vraisemblance ignorant le mécanisme à l'origine des données manquantes $\mathcal{L}(\theta|Y_{\text{obs}})$:

$$\mathcal{L}(\theta, \psi|Y_{\text{obs}}, R) \propto \mathcal{L}(\theta|Y_{\text{obs}}).$$

En présence d'un mécanisme ignorable, [181] montre qu'il n'est donc plus nécessaire de modéliser la distribution du mécanisme à l'origine des données manquantes pour estimer θ. Ce type d'approche est à la base des approches fondées sur la maximisation de la vraisemblance qui sont décrites dans la section 4.3.

Enfin, pour utiliser au mieux les informations sur la répartition des données manquantes et leur mécanisme de génération, un autre type d'approche est décrit dans [212]. Les auteurs proposent l'utilisation d'arbres de décision pour déterminer quelles sont les variables permettant d'expliquer la présence de manquants. Ces approches peuvent permettre d'utiliser l'information obtenue sur la présence de valeurs manquantes pour mettre en œuvre des stratégies plus efficaces d'analyse des données manquantes (pondération des cas complets, comme décrit dans la section 4.2.1, modèles à effets aléatoires ou modèles de mélange de profil, comme décrits dans la section 4.6, par exemple). Ils montrent également que cette approche est performante, y compris dans le cas MCAR, sur un cas pratique de données médicales.

4.2 Méthodes fondées uniquement sur les données observées

Une première approche pour pouvoir utiliser et analyser des données contenant des valeurs manquantes consiste à utiliser uniquement les observations disponibles. Ces approches présentent l'avantage de ne pas avoir recours à la spécification un modèle d'imputation (c'est-à-dire de remplacement des données) dont la qualité conditionne fortement les résultats de l'analyse. En revanche, elles sont souvent relativement inefficaces, biaisées ou induisent une perte de puissance importante.

Nous présentons, dans cette section, les approches possibles fondées sur ce paradigme, en décrivant les avantages et limites de celles-ci.

4.2.1 Analyse des cas complets et pondération

Une des premières possibilités pour traiter un jeu de données présentant des données manquantes est l'*analyse des cas complets*[9]. Cette méthode est la plus simple et la plus courante et c'est la méthode souvent implémentée par défaut dans les logiciels. Elle consiste à ne considérer que les individus pour lesquels toutes les données sont disponibles et donc à supprimer tout individu ayant au moins une valeur manquante.

Comme déjà souligné dans la section 4.1.3, l'analyse des cas complets est principalement valable dans le cas où les données manquantes sont MCAR et, même dans ce cas-ci, elle peut conduire à la suppression d'un nombre important d'individus (et donc à une perte de puissance dans les problèmes d'inférence). [94] déconseille l'utilisation de cette méthode lorsque les individus présentant des valeurs manquantes représentent plus de 5% de la population. En outre, dans le cas de la régression linéaire de Y_1 sur les autres variables $Y^{-1} = (Y_2, ..., Y_p)$, [192] montrent que l'analyse des cas complet produit des estimations non biaisées du modèle linéaire uniquement dans le cas où R est indépendante de Y_1 sachant Y^{-1} : $\mathbb{P}(R = 1|Y) = \mathbb{P}(R = 1|Y^{-1})$.

Une approche pour réduire les biais d'estimation dans l'analyse des cas complets consiste à repondérer les cas complets disponibles : c'est la *pondération par probabilité inverse* (IPW) [10] (voir [192] pour une revue de ce type d'approches). Généralement, la pondération est choisie comme l'inverse de la probabilité d'un individu d'être observé complètement, $\frac{1}{\eta_i}$. Les probabilités $(\eta_i)_{i=1,...,n}$ étant inconnues, elles sont estimées par un modèle de régression dont la variable à prédire est la variable R. Des équivalences asymptotiques ont été montrées dans [172] et [170] entre IPW et l'imputation multiple (voir section 4.5.2), dans le cas où Y est MAR et où les modèles d'imputation (pour l'imputation multiple) et de génération des données manquantes (IPW) sont correctement spécifiés. En pratique, [192] notent que les études empiriques donnent, en général, un avantage d'efficacité à l'imputation multiple mais soulignent aussi quelques avantages de IPW : sa simplicité conceptuelle et de mise en œuvre, sa meilleure efficacité lorsque la distribution de $Y_{1,\text{obs}}$ est très différente de celle de $Y_{1,\text{miss}}$ ou lorsque les cas non complets tendent à avoir des valeurs manquantes pour beaucoup de (et non pour quelques) variables. Enfin, [192] soulignent que IPW peut produire des poids très instables, lorsque l'estimation de η_i est faible. Les auteurs proposent quelques

9. *listwise deletion* en anglais.
10. *inverse probability weighting* en anglais.

solutions pour aborder ce problème, comme la stabilisation des poids et l'augmentation de IPW (AIPW). Enfin, au niveau de l'implémentation, le package **ipw** [227] permet de déterminer les probabilités inverses à utiliser pour l'imputation.

Conclusion et recommandations :

— *Avantages* : faciles à mettre en œuvre ; ne requièrent pas de spécifier un modèle d'imputation correct ;

— *Désavantages* : principalement valables dans les cas MCAR (analyse des cas complets) et MAR (IPW) ; requièrent que le nombre de cas complets corresponde à une proportion importante des données de départ ; en pratique souvent moins efficaces que l'imputation multiple.

4.2.2 Analyse des cas disponibles

Afin d'éviter la diminution trop importante du nombre d'individus dans l'analyse statistique, une alternative à l'analyse des cas complets est l'*analyse des cas disponibles* [11] ([7] et [164]). Cette approche consiste à estimer différents aspects du problème avec différents sous-échantillons en utilisant le maximum d'information disponible dans chacun des sous-problèmes. On inclut aussi dans l'analyse des cas disponibles, le cas où une variable entière est retirée du jeu de données parce que son taux de valeurs observées est trop faible ou inférieur à 1 (dans ce dernier cas, la méthode prend le nom d'*analyse des variables complètes*).

De manière plus précise, deux exemples typiques d'utilisation de cette approche sont présentés ci-dessous :

— si l'analyse statistique requiert l'estimation d'une matrice de covariance des variables Y_j, on peut estimer la covariance entre chaque paire de variables à partir de

$$\text{Cov}(Y_j, Y_{j'}) = \frac{1}{n_{jj'}} \sum_{i=1}^{n} y_{ij} y_{ij'} r_{ij} r_{ij'} - \overline{\mathbf{y}}_j^{jj'} \overline{\mathbf{y}}_{j'}^{jj'}$$

où $n_{jj'} = \sum_{i=1}^{n} r_{ij} r_{ij'}$ est le nombre de cas disponibles pour Y_j et $Y_{j'}$ et $\overline{\mathbf{y}}_j^{jj'} = \frac{1}{n_{jj'}} \sum_{i=1}^{n} y_{ij} r_{ij} r_{ij'}$ est la moyenne empirique de Y_j sur ces cas disponibles. Parfois, pour utiliser l'information maximale disponible, la moyenne est estimée par $\overline{\mathbf{y}}_j^{jj'} = \frac{1}{n_{jj'}} \sum_{i=1}^{n} y_{ij} r_{ij}$, moyenne empirique sur les cas disponibles pour Y_j. Cet estimateur peut être utilisé, par exemple, dans le cas d'un modèle linéaire (avec Y la variable à expliquer ou les variables explicatives) dans lequel l'estimation des paramètres ne fait intervenir que des estimateurs des moments du premier et du second ordre (càd, de la moyenne et des variances/covariances) mais [7] indique qu'alors, en dehors du cas MCAR, les estimations sont biaisées, comme pour l'analyse des cas complets ;

— pour l'apprentissage d'arbres de classification ou de régression ([84] et [32]), les données sont partitionnées récursivement de manière binaire en recherchant, pour dans chaque nœud t déjà construit, une variable Y_j et un seuil s_j^* qui maximisent un critère d'homogénéité des ensembles $\{i \in t : \ y_{ij} < s_j^*\}$ et $\{i \in t : \ y_{ij} \geq s_j^*\}$. Les données manquantes lors de l'apprentissage sont prises en compte en définissant, pour chaque variable, le seuil de partition optimal, s_j^*, à partir des observations non

11. *pairwise deletion ou available-case analysis* en anglais.

manquantes, $\{i \in t :\ r_{ij} \neq 0\}$, uniquement. Le critère d'homogénéité est également construit sur ces observations uniquement.

Une fois le meilleur ensemble (Y_j, s_j^*) défini par minimisation du critère d'homogénéité, les données sont ensuite partitionnées en deux sous-ensembles (non disjoints) $\{i \in t :\ y_{ij} < s_j^*$ et $r_{ij} = 1\} \cup \{i \in t :\ r_{ij} = 0\}$ et $\{i \in t :\ y_{ij} \geq s_j^*$ et $r_{ij} = 1\} \cup \{i \in t : r_{ij} = 0\}$, ce qui correspond à la propagation des observations manquantes dans les deux branches de l'arbre. Cette approche est appelée *partitionnement probabiliste* et une alternative à celle-ci est la définition de *variables de substitution* (voir section 4.2.4 pour une discussion et des éléments de comparaison).

Les approches d'analyse des cas disponibles posent en général des problèmes de deux types différents, qui viennent du fait que les différents composants des modèles (covariances ou bien partition dans un arbre) sont calculés sur des sous-échantillons différents :

— d'une part, cette approche peut favoriser (ou défavoriser) de manière artificielle certaines variables selon leur taux de valeurs manquantes dans l'analyse ou la prédiction. Par exemple, en présence de données manquantes MAR, [32] montrent que cette approche ne dégrade que peu les performances en apprentissage de la méthode sauf si les données sont manquantes de manière plus importantes pour les variables susceptibles d'être les plus pertinentes pour partitionner l'échantillon. Dans ce dernier cas, l'utilisation de la stratégie d'analyse des cas disponibles a des effets sur les performances en apprentissage : les erreurs en apprentissages sont majorés par rapport à d'autres approches comme l'utilisation de variables de substitution ;

— d'autre part, dans le cas du calcul d'une matrice de covariance ou de corrélation, l'analyse des cas disponibles produit une matrice avec des corrélations calculées sur des individus différents et/ou sur un nombre différent d'individus. Les résultats de cette méthode résultent d'une série d'analyses sur divers sous-échantillons qui peuvent être représentatifs de populations différentes. Ce problème complique les interprétations des corrélations et limite la généralisation à une population spécifique : comme les corrélations sont calculées sur des sous-échantillons de tailles différentes, les erreurs standards des estimateurs habituels sont difficiles à obtenir. Par exemple, la stratégie consistant à les calculer en utilisant la taille moyenne des échantillons sous-estime les erreurs standards [134]. Enfin, si les moyennes sont calculées sur les cas disponibles pour chacune des deux variables indépendamment, il est possible d'obtenir des corrélations incohérentes (non comprises entre -1 et 1), en particulier pour des variables fortement corrélées [37].

Le package **regtools** propose des implémentations de type « analyse des cas disponibles » de plusieurs méthodes statistiques en étendant, par exemple, les fonctions `lm` (régression linéaire), `prcomp` (ACP) et `loglin` (modèles log-linéaires).

Conclusion et recommandations :

— *Avantages* : facile à mettre en œuvre ; ne requiert pas de spécifier un modèle d'imputation correct ; permet de prendre en compte plus d'individus par rapport à l'analyse des cas disponibles ;

— *Désavantages* : principalement valable dans le cas MCAR ; favorise artificiellement certaines variables ; produit des statistiques sur des sous-populations différentes, difficilement comparables.

4.2.3 Ajustement par variable binaire

L'*ajustement par variable binaire* [12] s'utilise dans des modèles de régression lorsque l'analyse des cas complets n'est pas possible en raison d'un trop faible nombre de cas complets [54]. Elle consiste à associer à chaque variable explicative incomplète, Y_j, la variable Y_j^* définie par :

$$Y_j^* = \begin{cases} Y_j & \text{si } Y_j \text{ est observée,} \\ A & \text{sinon.} \end{cases}$$

où $A \in \mathbb{R}$ est une constante arbitraire (souvent 0 ou la moyenne de Y_j, mais sa valeur n'est pas importante). Il suffit alors de remplacer chaque variable incomplète Y_j par le couple (Y_j^*, R_j).

Par rapport à l'analyse des cas complets, cette méthode permet d'améliorer la précision de certains estimateurs en utilisant l'intégralité des individus disponibles dans le jeu de données initial. Néanmoins, cette méthode produit des estimateurs qui sont biaisés dans tous les cas.

Conclusion et recommandations :

— *Avantages* : facile à mettre en œuvre ; alternative à l'analyse des cas complets lorsque le nombre de cas complets est trop faible ;

— *Désavantages* : produit presque systématiquement des estimateurs biaisés dans le cadre de problèmes d'inférence ; pas recommandée en pratique.

4.2.4 Approche par substitution de variables

Dans le cas particulier d'un modèle de prédiction (régression ou classification supervisée) dans lequel Y sont les variables explicatives, on peut aussi obtenir des prédictions à partir d'observations incomplètes de Y en utilisant des approches par substitution de variables. Ces approches sont particulièrement utilisées dans le cas d'arbres de régression ou de classification [32], qui utilisent la notion de « partition de substitution » : la partition de substitution d'une partition du nœud t par la variable Y_j et le seuil s_j^* est définie comme la partition par la variable $Y_{j'}$ (pour un $j' \neq j$) et le seuil $s_{j'}^*$ qui minimise une mesure d'association entre les deux partitions sur les individus observés.

[32] montrent que l'utilisation des partitions de substitution pour la prédiction d'une observation avec des données manquantes donne des performances de qualité dès lors que les observations sont manquantes aléatoirement et que plusieurs des variables explicatives Y_j sont corrélées (ce qui induit des mesures d'association élevées entre une partition et sa ou ses partitions de substitution). [68] vont plus loin et proposent une étude exhaustive, théorique et empirique, des diverses méthodes classiques de prise en charge des valeurs manquantes : analyse des cas disponibles (section 4.2.1) et analyse des variables complètes (section 4.2.2), imputation par la moyenne (section 4.4.1), création d'une modalité particulière « *manquant* » utilisé comme une modalité supplémentaire (qui est à rapprocher de la méthode décrite en section 4.2.3), utilisation de partitions probabilistes (section 4.2.2) et, enfin, variables de substitution. Les résultats théoriques et empiriques montrent que la qualité de l'approche dépend de deux critères :

12. *dummy variable adjustment* en anglais.

— si les données à prédire (et pas seulement les données d'apprentissage) contiennent elles-aussi des données manquantes et que la variable à prédire est liée au processus de génération des données manquantes (ce cas contient des situations MAR et MNAR) alors l'approche par création d'une modalité supplémentaire est la plus efficace en terme d'erreur de prédiction ;

— dans tous les autres cas, les approches par substitution de variables, utilisation des variables complètes et partitions probabilistes sont, de manière à peu près équivalentes, les meilleures, avec un désavantage pour l'approche par variables complètes dans les cas de taux de manquants faibles et un désavantage pour l'approche par imputation dans les cas de taux de manquants importants.

Conclusion et recommandations :

— *Avantages* : a montré son efficacité empirique dans le cadre des arbres de régression et de classification ;

— *Désavantages* : principalement valable dans le cas MAR et lorsque les covariables sont fortement corrélées ; gourmande en temps de calcul.

4.3 Inférence statistique en présence de valeurs manquantes

Lorsque l'objet de l'analyse statistique est l'inférence, les approches fondées sur la modélisation paramétrique de la distribution multivariée des données, $f(Y;\theta)$ permettent d'obtenir des estimations de θ sans avoir à imputer les données et en garantissant une estimation non biaisée de ce paramètre, à condition que l'hypothèse d'ignorabilité du mécanisme de génération des données manquantes soit vérifiée. Les premiers travaux de ce type ont été proposés par [187] et se fondent sur des approches de maximisation de la vraisemblance dans le cadre d'un modèle gaussien. On les retrouve fréquemment résumés sous le nom générique de « modélisation jointe [13] », qui regroupe des approches fréquentistes et bayésiennes.

4.3.1 Approches fréquentistes

Lorsque la densité $f(Y;\theta)$ est spécifiée et dans le cas d'un mécanisme ignorable, l'équation (4.1) indique que la vraisemblance de θ pour les données observées est de la forme

$$\mathcal{L}(\theta|Y_{\text{obs}}) \propto \log \int f(Y;\theta) dY_{\text{miss}}.$$

Les estimateurs du maximum de vraisemblance offrent des estimations non biaisées de θ mais, à cause de l'intégration, le calcul direct de la vraisemblance précédente n'est possible que dans de très rares cas en présence de données incomplètes. Les approches fréquentistes pour l'estimation de θ dans ce cadre-ci peuvent être regroupées en deux grands types de méthodes : la première utilise une approche EM [65] et la seconde se fonde une approche

13. *Joint Modelling* en anglais

par maximum de vraisemblance à information incomplète [14], originellement proposée par [81].

— **Algorithme EM.** L'idée de l'utilisation de l'algorithme EM consiste à alterner deux étapes :

une étape E (Expectation) dans laquelle les statistiques suffisantes du modèle sont « complétées » en tenant compte des valeurs observées et de la valeur courante du paramètre, $\theta^{(t)}$. La forme de ces statistiques dépend du modèle considéré ;

une étape M (Maximization) dans laquelle la valeur du paramètre courant est mise à jour pour obtenir $\theta^{(t+1)}$ par maximisation de la vraisemblance complétée à l'étape E.

L'approche EM présente l'avantage d'être convergente [65]. Toutefois, si dans le cas d'une distribution gaussienne, les formules explicites des étapes E et M sont données dans ([133] et [72]), la mise en œuvre de cette approche peut s'avérer plus complexe pour d'autres distributions, comme discuté par [146]. Enfin, [72] liste un certain nombre de désavantages à cette approche, en particulier, le fait qu'elle ne fournit pas d'estimation de la variabilité des estimations de θ : une étape supplémentaire (utilisant par exemple une approche par bootstrap ; voir [94] et section 4.5.3) est nécessaire pour obtenir des estimations des erreurs types.

— **FIML.** L'approche par maximum de vraisemblance à information incomplète, quant à elle, ne remplit pas les valeurs manquantes mais détermine une vraisemblance partielle pour chaque observation i. Celle-ci, notée $\mathcal{L}_i$, est obtenue par calcul de la vraisemblance ordinaire sur les variables observées pour i (les paramètres non estimables car fondés sur des variables manquantes pour i sont remplacés par 0). Dans le cas gaussien, si on note $\theta = (\mu, \Sigma)$ les paramètres (moyenne et variance) de la loi jointe, on obtient

$$\mathcal{L}_i = K_i - \frac{1}{2}\log|\Sigma_i| - \frac{1}{2}(\mathbf{y}_i^* - \mu_i)^\top \Sigma_i^{-1}(\mathbf{y}_i^* - \mu_i)$$

où $\mathbf{y}_i^*$ est le vecteur des variables observées pour l'individu i, μ_i et Σ_i correspondent respectivement au vecteur moyenne et à la matrice de covariance restreints aux variables observées pour i. K_i est une constante qui dépend du nombre de valeurs observées pour i.

Ces n quantités sont alors sommées pour obtenir la fonction de log-vraisemblance sur l'ensemble de l'échantillon :

$$\widetilde{\mathcal{L}}(\theta|Y) = \sum_i^n \mathcal{L}_i.$$

θ est enfin obtenu comme le maximum de cette vraisemblance $\widetilde{\mathcal{L}}$. Outre l'estimation du paramètre de la loi jointe des données, cette approche permet d'obtenir des erreurs types sur le paramètre, ce qui est un avantage sur l'approche précédente. Comme noté par [72], elle peut aussi être plus simple à mettre en œuvre que l'approche EM car elle ne nécessite pas de dériver une étape E spécifique à chaque modèle.

Notons que les deux approches décrites ci-dessus dépendent toutes les deux de l'hypothèse d'ignorabilité du mécanisme de génération des données manquantes. Elles sont donc restreintes au cas de données MAR et non applicables dans le cadre MNAR. Elles sont,

14. FIML : *Full Information Maximum Likelihood*, aussi connue sous les noms de *direct maximum likelihood* ou *raw maximum likelihood*.

en outre, fortement dépendantes de la véracité du modèle sous-jacent de génération des données, souvent supposé gaussien.

Enfin, ces approches sont fréquemment utilisées dans le cadre de l'imputation de données (voir section 4.4) : une fois θ estimé, l'imputation, c'est-à-dire, le remplacement de la valeur manquante par une valeur plausible, peut être réalisée en échantillonnant selon la loi $f(Y;\theta)$ pour compléter les valeurs manquantes. Notons toutefois que le cadre d'application de l'approche dépasse celui de l'imputation : l'approche par maximum de vraisemblance est initialement destinée à l'estimation du paramètre de la loi jointe de Y, θ, et peut donc être utilisée directement (sans avoir recours à l'imputation) si l'estimation de θ est la question d'intérêt pour le statisticien. Elle offre, en particulier, un cadre général pour l'inférence et rend possible l'utilisation de tests du rapport de vraisemblance.

Conclusion et recommandations :

— *Avantages* : bien adaptées au cadre de l'inférence statistique ; ne requièrent pas l'imputation de valeurs ; fournissent des estimations non biaisées dans le cadre d'un mécanisme ignorable ; peuvent être utilisées également pour l'imputation des valeurs manquantes ; fournissent des estimations des erreurs sur les paramètres estimés ;

— *Désavantages* : seulement valables dans le cas MAR ; requièrent des hypothèses fortes sur la loi jointe des données ; gourmande en temps de calcul ; garanties asymptotiques qui requièrent des échantillons de grande taille.

4.3.2 Approches bayésiennes

Une autre approche pour estimer le paramètre θ de la loi jointe $f(Y;\theta)$ est le recours à une approche bayésienne dans laquelle une loi a priori est définie sur θ, $p(\theta)$. Cette loi *a priori* est utilisée pour déterminer la loi *a posteriori* du paramètre connaissant les données observées :

$$p(\theta|Y_{\text{obs}}) \propto f(Y_{\text{obs}}|\theta)p(\theta).$$

L'inférence bayésienne consiste à déterminer cette loi *a posteriori*.

En présence de valeurs manquantes, comme dans le cadre fréquentiste, l'hypothèse d'un mécanisme ignorable permet d'écrire

$$p(\theta|Y_{\text{obs}}) = \int p(\theta|Y) f(Y_{\text{miss}}|Y_{\text{obs}},\theta) dY_{\text{miss}}. \tag{4.2}$$

[207] proposent un cadre général pour l'inférence bayésienne sous cette hypothèse, avec une approche par augmentation de données. Celle-ci consiste à itérer deux étapes :

une étape d'imputation (étape I) dans laquelle M tableaux de données complets sont générés selon la loi $f(Y_{\text{miss}}|Y_{\text{obs}},\theta)$ courante. Cette étape consiste à échantillonner θ M fois dans la distribution courante de $p(\theta|Y_{\text{obs}})$, $p_t(\theta|Y_{\text{obs}})$, et à utiliser les valeurs échantillonnées et la donnée de $f(Y|\theta)$ pour générer les données complètes $\mathbf{Y}^{(1)}, \ldots, \mathbf{Y}^{(M)}$. [207] notent la similarité d'approche entre cette étape et l'imputation multiple (décrite en section 4.5.2), d'où le nom « étape d'imputation » qu'ils lui ont donné ;

une étape postérieure (étape P) dans laquelle la valeur courante de la loi *a posteriori* est obtenue par

$$p_{t+1}(\theta|Y_{\text{obs}}) = \frac{1}{M}\sum_{m=1}^{M} p\left(\theta|Y = \mathbf{Y}^{(M)}\right)$$

L'étape P demande de pouvoir calculer analytiquement $p(\theta|Y)$. La mise en œuvre de cette approche peut donc être plus ou moins facile selon le choix de l'*a priori* effectué. Dans le cas gaussien de paramètre $\theta = (\mu, \Sigma)$, si on choisit pour *a priori* de θ l'*a priori* (non informatif) de Jeffrey [90], on sait que $\mathcal{L}(\Sigma|Y)$ est une loi de Wishart inverse à $n-1$ degré de liberté et de paramètre d'échelle S^{-1}, où S est la matrice de covariance empirique de θ, et $\mathcal{L}(\mu|\Sigma, Y) \sim \mathcal{N}(\overline{\mathbf{y}}, \frac{1}{n}\Sigma)$. Toutefois, la loi a posteriori n'a pas toujours une forme explicite simple à déterminer et on peut alors avoir recours à des algorithmes itératifs (comme l'algorithme de Gibbs) pour pouvoir échantillonner dans la loi *a posteriori*.

[207] montrent que l'approche proposée, sous des conditions relativement peu restrictives, converge bien vers la vraie loi a posteriori $p(\theta|Y_{\text{obs}})$. En outre, l'estimation bayésienne s'adapte assez bien à tout type de répartition de données manquantes et, contrairement aux approches fréquentistes, elle est bien adaptée aux échantillons de petites tailles puisqu'elle ne repose pas sur des résultats asymptotiques. Par ailleurs, elle fournit directement une estimation de la variance associée à l'estimation des paramètres via la loi *a posteriori* et permet également, comme les approches fréquentistes, de pratiquer une imputation des données manquantes en utilisant un échantillonnage similaire à l'étape I décrite plus haut.

Conclusion et recommandations :

— *Avantages* : bien adaptée au cadre de l'inférence statistique ; ne requiert pas l'imputation de valeurs ; garanties théoriques de convergence ; peut être utilisée également pour l'imputation des valeurs manquantes ; adaptée aux échantillons de petite taille ;

— *Désavantages* : seulement valable dans le cas MAR ; gourmande en temps de calcul ; requiert des hypothèses fortes sur la loi jointe des données.

4.3.3 Packages R

Divers packages proposent des implémentations de ces approches :

— **Amelia** [102] propose diverses méthodes d'imputation EM et IP et des graphiques diagnostiques. Le package propose des versions fondées sur des approches bootstrap ou bayésienne pour estimer les incertitudes et gère les imputations multiples (voir section 4.5). Le package possède une interface graphique (AmeliaView) permettant aux personnes non familières avec R de l'utiliser ;

— **lavaan** [178] propose une approche par maximum de vraisemblance à information incomplète pour prendre en compte les données manquantes dans les modèles à équations structurelles ;

— **norm** [190] est un package proposant l'analyse de données multivariées suivant une distribution normale. La fonction `em.norm` donne les estimations des paramètres obtenues par approche EM. Pour obtenir une imputation des données manquantes, la fonction `imp.norm` peut être utilisée avec les paramètres estimés par la fonction précédente. Enfin, la fonction `da.norm` implémente l'approche bayésienne décrite ci-dessus. En particulier, [190] conseillent l'utilisation des résultats de l'algorithme EM pour initialiser l'approche bayésienne et mieux calibrer le nombre d'itérations nécessaires pour celle-ci. **cat** et **mix** [190] sont l'équivalent du package **norm** pour l'imputation de variables catégorielles et mixtes. **cat** estime les paramètres d'une distribution multinomiale pour les variables catégorielles.

4.4 Imputation simple

Une alternative aux approches qui se fondent sur les données observées uniquement est probablement l'approche la plus courante de traitement des données manquantes : l'imputation de celles-ci par une valeur unique utilisée pour « remplacer » la valeur non observée. On appelle cette approche *imputation simple*. Comme souligné par [189], les approches par imputation présentent plusieurs avantages par rapport aux approches utilisant uniquement les données observées : d'une part, elles permettent de limiter la perte de puissance liée à la taille réduite de l'échantillon correspondant aux individus complètement observés. D'autre part, si les données observées contiennent suffisamment d'information pour permettre de prédire les valeurs non observées, l'inférence statistique conserve sa précision initiale. Enfin, une fois les données manquantes imputées, l'utilisateur obtient un tableau de données complet de n individus sur lequel n'importe quelle analyse statistique classique peut être pratiquée, sans nécessité d'avoir un traitement particulier personnalisé pour les valeurs non observées : ces approches ne sont donc pas restreintes au cadre de l'inférence statistique.

Selon que l'objectif de l'imputation est l'inférence statistique d'une quantité d'intérêt ou bien l'obtention d'un tableau complet permettant diverses analyses statistiques, l'impact des erreurs d'imputation est différent. Les différentes méthodes d'imputation s'intéressent donc à conserver au mieux certains aspects dans les variables observées (distribution univariée, corrélations entre variables, etc) en fonction de l'objectif de l'utilisateur. L'erreur commise par la méthode d'imputation est alors mesurée soit en terme d'erreur commise sur la valeur imputée elle-même (erreur d'imputation, voir section 4.5.1 sur les outils de diagnostic), soit sur le résultat de l'analyse.

Dans cette section, nous décrivons les méthodes les plus courantes d'imputation simple, que nous avons organisées en trois grandes familles (complétion stationnaire, imputation fondée sur des similarités entre individus, imputation fondée sur des méthodes de prédiction) auxquelles s'ajoutent les méthodes d'imputation adaptées à l'analyse factorielle des données. Dans toutes ces familles, des approches existent pour imputer des variables numériques ou catégorielles. Nous présentons les avantages et inconvénients de ces méthodes, qui sont toutes principalement adaptées au cadre MAR. En particulier, nous essayons de systématiquement mettre en avant le cadre approprié d'utilisation de celles-ci, qui est lié, à la fois, à l'usage que l'utilisateur souhaite avoir du tableau imputé, mais aussi, au type de répartition des données manquantes. Enfin, nous discutons, en conclusion de la section, d'une méthodologie appropriée pour l'analyse globale d'un tableau de données contenant des valeurs manquantes ainsi que d'ouvertures pour l'imputation dans le cadre de données ayant une structure particulière (séries temporelles, par exemple).

4.4.1 Complétion stationnaire

L'imputation par *complétion stationnaire* ([189] et [118]) consiste à remplacer les valeurs manquantes de la variable Y_j par une valeur identique, m_j, pour tous les individus. Différents types de complétion stationnaire existent :

— pour une variable catégorielle prenant ses valeurs dans un ensemble fini $\{1, \ldots, M\}$, le mode des valeurs $\{y_{ij} : r_{ij} \neq 0\}$ est utilisé pour l'imputation [15] : $m_j = \arg\max_{u=1,\ldots,M} \text{Card}\{i : r_{ij} \neq 0 \text{ et } y_{ij} = u\}$;

— pour une variable numérique, la valeur moyenne ou médiane des $\{y_{ij} : r_{ij} \neq 0\}$ est utilisée pour l'imputation : $m_j = \frac{1}{\sum_{i=1}^{n} r_{ij}} \sum_{i=1}^{n} y_{ij} r_{ij}$. L'imputation par la moyenne est simple à mettre en œuvre mais ses propriétés sont limitées : elle distord la distribution de la variable d'intérêt même dans le cas MCAR. Par conséquent, certaines caractéristiques de la distribution sont biaisées, en particulier la variabilité qui est réduite ;

— pour une variable numérique, une combinaison convexe des valeurs $\{y_{ij} : r_{ij} \neq 0\}$ peut également être utilisée pour l'imputation : $m_j = \frac{1}{\sum_{i=1}^{n} r_{ij}} \sum_{i=1}^{n} w_i y_{ij} r_{ij}$ où w_i sont les poids de la combinaison linéaire tels que $\frac{\sum_{i=1}^{n} w_i r_{ij}}{\sum_{i=1}^{n} r_{ij}} = 1$. L'imputation par la moyenne est un cas particulier d'imputation par combinaison linéaire (dans lequel $w_i = 1$).

[189] soulignent un des principaux problèmes de cette approche : dans le cas simple de l'estimation de la moyenne de la variable Y_j, contenant des valeurs manquantes, l'imputation par la moyenne diminue la taille attendue de l'intervalle de confiance d'une part en introduisant un biais qui diminue la valeur de l'écart type empirique de Y_j et d'autre part en sur-estimant, par n, le nombre de valeurs observées. Les auteurs montrent que pour 25% de valeurs manquantes, le taux d'erreur observé sur l'intervalle de confiance de la moyenne est près de trois fois ce qu'il devrait être. Enfin, outre une sous-estimation de la variabilité des variables, y compris dans le cas MCAR, cette approche modifie les corrélations entre variables. Pour limiter ces problèmes, des variantes de l'imputation stationnaire peuvent être mises en œuvre : en particulier, lorsque la population est naturellement stratifiée en sous-populations homogènes, l'imputation par complétion stationnaire peut être réalisée indépendamment dans chacune des sous-populations.

Enfin, un autre exemple de méthode d'imputation se rapprochant de la complétion stationnaire est celui de données longitudinales où la variable Y_j est mesurée pour les individus i à divers pas de temps $t = 1, \ldots, T$. Dans ce cas, l'imputation d'une valeur manquante y_{ijt} peut être faite par la dernière valeur connue de cette variable pour cet individu, y_{ijt^*}, pour $t^* = \arg\max_{u=1,\ldots,t-1}\{r_{iju} \neq 0\}$. Cette approche, souvent abrégée par LOCF [16] et aussi connue sous le nom de « analyse du point final [17] », fait l'hypothèse implicite qu'il n'y a pas eu de changement entre t^* et t. C'est une approche de gestion des données manquantes très largement pratiquée dans le cadre d'études cliniques longitudinales, plus particulièrement des études dites « en intention de traiter », dans lesquelles deux groupes de malades, un groupe traité et un groupe contrôle, sont suivis de manière longitudinale. [151] soulignent que, dans le cas où le taux de sortie de l'étude du groupe traité est lié au traitement, cette approche biaise les conclusions en faveur du traitement, avec des conséquences potentiellement très importantes pour la prise en charge médicale des malades. Ces conclusions sont confirmées par l'étude par simulations de [217] qui montre une violation du degré de significativité et une perte de puissance importante dans les tests de comparaison entre les deux groupes dans ce cas-ci.

15. *Concept Common Attribute Value Fitting* en anglais.
16. *Last Observation Carried Forward* en anglais.
17. *endpoint analysis* en anglais

Les approches par complétion stationnaire sont disponibles, par exemple, dans les packages **simputation** (imputation par la médiane), **Hmisc** (imputation aléatoire, par la moyenne, par la médiane, par le mode...) et **ForImp** (imputation par la moyenne, par la médiane, par le mode). De manière plus générique, la fonction `impute` du package **Hmisc** permet d'utiliser une fonction arbitraire des valeurs observées pour une imputation par complétion stationnaire.

Conclusion et recommandations :

— *Avantages* : facile à mettre en œuvre ; permet d'obtenir un jeu de données complet sur lequel n'importe quelle analyse statistique peut être pratiquée ;

— *Désavantages* : biaise (diminue) l'estimation des variabilités des variables ; modifie les corrélations entre variables ; sur-estime la taille de l'échantillon observé ; non recommandée en pratique, même dans les cas MCAR, sauf si le nombre de valeurs manquantes est très faible et que l'on ne sait pas mettre en œuvre une autre méthode décrite dans ce papier.

4.4.2 Méthodes fondées sur des similarités entre individus

Une autre approche pour l'imputation simple consiste à utiliser les valeurs observées des individus similaires à l'individu pour lequel une valeur est manquante. Ces méthodes sont liées à des imputations par k plus proches voisins (kNN) ou à des méthodes regroupées sous le nom générique d'approches « hot-deck » (les deux dénominations étant parfois confondues selon les publications).

Méthode des k plus proches voisins (kNN)

La méthode kNN est une méthode d'imputation multivariée fondée sur une notion de distance entre individus, $d(i, i')$, obtenue à partir de q covariables entièrement observées, X. Pour une valeur manquante y_{ij}, l'approche consiste, d'une part, à calculer l'ensemble des distances $d(i, i')$ pour les $i' \neq i$ tels que $r_{i'j} \neq 0$ et à retenir les k observations (pour un $k \in \mathbb{N}^*$), $y_{(1)j}$, ..., $y_{(k)j}$, correspondant aux k plus petites distances. Les k valeurs $(y_{(i)j})_{i=1,\ldots,k}$ des plus proches voisins sont alors agrégées pour imputer la valeur manquante y_{ij}. Généralement, si la variable Y_j est numérique, la valeur manquante est imputée par la moyenne (ou la médiane) des $(y_{(i)j})_{i=1,\ldots,k}$. L'approche se généralise facilement au cas où il n'y a pas de covariables complètement observées en calculant des distances, pour chaque individu, qui sont basées sur un sous-ensemble d'individus et/ou de variables complètement observées.

La méthode requiert le choix de deux hyper-paramètres : d, la distance choisie, et k, le nombre de voisins utilisés pour l'estimation. Des choix classiques pour d sont la distance euclidienne entre valeurs observées,

$$d(i, i') = \sum_{j'=1}^{q} (x_{ij'} - x_{i'j'})^2, \tag{4.3}$$

ou la distance de Mahalanobis. Lorsque le jeu de données contient des variables catégorielles, [243] propose l'utilisation d'une distance particulière prenant en compte l'existence de ces variables et la valeur imputée est alors le mode des $(y_{(i)j})_{i=1,\ldots,k}$. [148] proposent une approche alternative fondée sur l'analyse canonique des corrélations entre les covariables X

et les cas complets de Y : les plus proches voisins sont alors définis dans l'espace factoriel de projection de X. L'idée sous-jacente est de sélectionner les plus proches voisins dans un espace de corrélation optimale avec les variables à imputer.

Pour le choix de k, [112] soulignent que les recommandations pour le choix de cette valeur varient selon les auteurs : par exemple, [45] et [105] utilisent $k = 1$ ou 2, [21] recommandent d'utiliser une valeur faible de k alors que [215] recommandent une valeur de k comprise entre 10 et 20 pour des jeux de données de grande taille. Dans leurs expériences, [112] mettent en valeur une dépendance de k à la taille du jeu de données et suggèrent de choisir k égal à la racine carrée du nombre moyen de cas complets des variables utilisées pour l'imputation.

L'imputation par kNN est implémentée dans de nombreux packages R. Parmi ceux-ci, on peut citer :

— **DMwR** : ce package regroupe des fonctions utiles pour la fouille de données et est associé à l'ouvrage de [214]. La fonction `knnImputation` de ce package propose deux méthodes d'imputation des valeurs manquantes. La méthode par défaut est une moyenne pondérée, le poids de l'individu i' étant donné par $\exp(-d(i,i'))$ où d est la distance euclidienne entre l'individu imputé, i et i'. L'approche alternative consiste à remplacer chaque valeur manquante par la médiane des kNN (ou bien par le mode quand la variable à imputer est catégorielle) ;

— **impute** [215] : ce package Bioconductor est destiné à l'imputation de données d'expressions de gènes (puces à ADN) et requiert donc un tableau de variables numériques. La méthode proposée dans ce package calcule des voisins dans l'espace des gènes et non dans l'espace des individus. Pour accélérer le calcul des distances euclidiennes entre gènes, le package utilise un pré-traitement par classification non supervisée et réduit le calcul des distances à un sous-groupe de gènes. L'imputation par la moyenne des kNN est finalement réalisée ;

— **VIM** [123] : ce package autorise l'imputation par kNN pour des données mixtes. Pour ce faire, les k voisins sont choisis en utilisant une variation de la distance de Gower [92]. Cette distance peut s'appliquer à un ensemble de variables à la fois numériques, catégorielles et binaires. Elle est fondée sur une notion de *contribution* de la covariable X_j qui est définie par

$$\begin{array}{cc} X_j \text{ est numérique} & X_j \text{ est catégorielle} \\ S_{ii'j} = \frac{|x_{ij} - x_{i'j}|}{\max_l(x_{lj}) - \min_l(x_{lj})} & S_{ii'j} = \begin{cases} 1 & \text{si } x_{ij} = x_{i'j} \\ 0 & \text{sinon.} \end{cases} \end{array}$$

De cette notion, on peut déduire une distance entre individus i et i' comme suit :

$$d(i,i') = \frac{\sum_{j=1}^{p} S_{ii'j}}{n}.$$

Les variables numériques sont finalement imputées par la médiane des valeurs des voisins tandis que les variables catégorielles sont imputées par le mode des valeurs des voisins ;

— **yaImpute** [60] : ce package met à disposition une grande variété de méthodes d'imputation par kNN, dont l'approche d'imputation par analyse canonique des corrélations décrite plus haut et propose plusieurs outils diagnostiques pour l'évaluation et la comparaison des approches d'imputation.

Hot-deck

L'imputation hot-deck est une approche qui a été introduite en 1947 pour traiter les valeurs manquantes dans les réponses des sondages démographiques (*Current Population Survey*) par le bureau national américain des sondages (*US Census Bureau*). [11] font une revue des méthodes hot-deck et de leurs propriétés.

L'imputation hot-deck est fondée sur le concept de *donneur*, qui est proche du concept de plus proche voisin. De manière plus précise, pour un individu i ayant une valeur manquante y_{ij}, on définit un ensemble de donneurs $\mathcal{D}(i)$ qui sont des individus i' « similaires » à l'individu i et pour lesquels $r_{i'j} \neq 0$. Une des valeurs $y_{i'j}$ pour $i' \in \mathcal{D}(i)$ est alors imputée pour y_{ij}. Les variantes de la méthode hot-deck diffèrent à deux niveaux : dans la phase de définition de l'ensemble des donneurs et dans la phase d'imputation.

Généralement, l'ensemble des donneurs d'un individu i est défini par le biais d'une mesure de similarité ou de distance calculée sur des covariables complètement observées, X, mais d'autres approches sont parfois pratiquées. Les plus courantes sont les suivantes :

— **Hot-deck métrique ou plus proches voisins**
Dans cette variante, l'ensemble des donneurs est défini comme l'ensemble des kNN de l'individu i pour une distance donnée calculée sur un ensemble de covariables X, complètement observées. La distance euclidienne est généralement utilisée. Cette approche est similaire au cas de l'approche kNN (voir section 4.4.2) mais diffère dans la phase d'imputation (voir ci-dessous), sauf pour le cas $k = 1$.

— **Hot-deck métrique avec score d'affinité**
Une autre méthode pour calculer la similarité entre deux individus a été proposée par [58] : le score d'affinité. Le score d'affinité $s(i, i')$ mesure le degré de similarité qui existe entre l'individu receveur i et chaque donneur potentiel i', pour lequel les p variables du jeu de données ont été observées. Il a été établi, dans un premier temps, pour des données discrètes et se définit alors comme la proportion de valeurs communes entre i et i' parmi les variables observées pour le receveur i :

$$s(i, i') = \frac{\sharp\{j = 1, \ldots, p : \, r_{ij} = 1 \text{ et } y_{ij} = y_{i'j}\}}{\sum_{j=1}^{p} r_{ij}}.$$

Dans le cas de variables numériques continues, [58] proposent d'adapter le score d'affinité de la manière suivante :

$$s(i, i') = \frac{\sum_{j=1}^{p} r_{ij} \mathbf{1}_{\{|y_{ij} - y_{i'j}| < \sigma\}}}{\sum_{j=1}^{p} r_{ij}}$$

où σ est un seuil à fixer (qui peut éventuellement être adapté en fonction de l'échelle de la variable). Dans les deux cas, l'ensemble des donneurs, $\mathcal{D}(i)$, se définit alors par $\mathcal{D}(i) = \{i' : \; s(i, i') = \max_{l \neq i} s(i, l)\}$.

— **Hot-deck hiérarchisé**
L'approche hot-deck hiérarchisé est similaire au cas d'imputation de données longitudinales décrit dans la section 4.4.1. Elle est utilisée lorsqu'il existe un ordre naturel entre les variables ($j = 1, \ldots, p$) et consiste à remplacer la valeur manquante y_{ij} par la valeur d'un individu qui a les mêmes valeurs pour les variables $Y_1, Y_2, ..., Y_{j-1}$. S'il n'en existe pas, elle est remplacée par la valeur d'un individu ayant les mêmes valeurs pour les variables $Y_1, Y_2, ..., Y_{j-2}$. Ce processus est itéré jusqu'à obtention d'au moins un individu correspondant à un critère de correspondance. Cette méthode est donc fondée

sur une définition modifiée de l'ensemble des donneurs $\mathcal{D}(i)$ qui sont des individus identiques à l'individu i pour certaines variables et a une phase d'imputation spécifique bien définie.

Une fois l'ensemble des donneurs $\mathcal{D}(i)$ défini, l'imputation est pratiquée selon diverses méthodes :

— **Hot-deck aléatoire avec ou sans remise**
L'approche hot-deck aléatoire consiste à remplacer une valeur manquante y_{ij} par la valeur $y_{i'j}$ pour un i' choisi au hasard dans $\mathcal{D}(i)$. Cette approche peut être utilisée pour des variables numériques ou catégorielles mais nécessite que les individus du jeu de données aient un profil homogène pour que les valeurs imputées ne soient pas éloignées de la vraie valeur. Aussi, si la population s'avère trop hétérogène, il est préférable de constituer des classes d'imputation réputées plus homogènes. La méthode hot-deck aléatoire est alors appliquée à l'intérieur de ces sous-populations et on parle alors de « hot-deck par classes ». En pratique, les classes d'imputation sont souvent définies en stratifiant le jeu de données selon des covariables entièrement observées ou en appliquant des procédures usuelles de classification sur le jeu de données [111].

— **Hot-deck séquentiel**
L'approche hot-deck séquentielle [133] est utilisée lorsqu'il existe un ordre naturel au sein des individus $i = 1, \ldots, n$. Si une valeur y_{ij} est manquante, elle est alors imputée par la valeur non manquante la plus récente parmi l'ensemble des donneurs $\mathcal{D}(i)$, y_{i^*j} avec $i^* = \arg\max_{i'=1,\ldots,i-1}\{y_{i'j} : r_{i'j} = 1\}$. En pratique, les variables sont ordonnées par le choix d'une variable (ou de plusieurs variables) de tri parmi les covariables X_j observées pour tous les individus. Celle-ci doit expliquer au mieux la variable à imputer (à partir des observations correspondant aux individus répondants) et, si besoin, les covariables de tri suivantes sont utilisées pour ordonner les ex-aequos. Comme l'estimateur obtenu dépend de l'ordre dans lequel les données sont ordonnées, il est nécessaire que la covariable de tri choisie ne soit pas fortement corrélée avec la probabilité de non-réponse. La conséquence du non respect de cette règle est l'imputation de la même valeur pour un grand nombre d'individus et donc la distorsion de la distribution de la variable imputée [119], qui entraîne une distorsion de la distribution des données et diminue artificiellement la variance estimée. Une solution de type hot-deck hiérarchisé, comme décrite ci-dessus, permet de limiter ce type de problème.

L'imputation hot-deck est implémentée dans les packages R suivant :

— **hot.deck** [58] : outre l'imputation simple par hot-deck métrique avec score d'affinité, ce package propose une imputation multiple (voir section 4.5.2) ;

— **HotDeckImputation** : ce package propose différentes méthodes d'imputation hot-deck : hot-deck séquentiel, hot-deck aléatoire, hot-deck métrique par kNN ainsi qu'une méthode appelée « hot-deck séquentiel CPS ». Cette dernière permet d'appliquer l'approche hot-deck séquentiel parmi les classes d'imputation ;

— **VIM** : outre les fonctionnalités d'analyse exploratoire des données manquantes, ce package propose également plusieurs approches d'imputation hot-deck (dont le hot-deck aléatoire et le hot-deck séquentiel) dans la fonction `hotdeck`. Le package **simputation** possède également une fonction `impute_hotdeck`, qui utilise les fonctions de VIM et permet divers types d'imputation hot-deck ;

Cold-deck

Cette approche est proche de la méthode hot-deck présentée dans la section précédente mais, dans ce cas-ci, les donneurs ne sont pas des individus du jeu de données initial. De manière plus précise, les mêmes variables Y ont été observées sur un second ensemble d'individus $i = n+1, \ldots, n+m$ et les donneurs sont définis au sein de cet ensemble. Par exemple, l'imputation de la valeur manquante y_{ij} pour un $i \leq n$, requiert la définition de l'ensemble des donneurs $\mathcal{D}(i) \subset \{n+1, \ldots, n+m\}$, par exemple par calcul des distances euclidiennes :

$$\forall\, i' = n+1, \ldots, n+m, \qquad d(i,i') = \sum_{j' \neq j} r_{ij}(y_{ij'} - y_{i'j'})^2.$$

Les cas typiques d'utilisation sont les cas où les donneurs proviennent d'enquêtes antérieures, de données historiques ou de l'expertise d'un spécialiste [11].

Conclusion et recommandations

L'avantage principal des méthodes basées sur des mesures de similarité est qu'elles ne requièrent pas d'hypothèses sur la distribution des données : elles peuvent être utilisées de manière souple avec des données de types variés et peuvent même s'adapter à des métriques d'intérêt spécifiques aux données étudiées (comme les distances basées sur la phylogénie entre espèces utilisées en biologie, par exemple ; [58]).

Les approches hot-deck préservent la distribution univariée des données dans le cadre MCAR [73, chap. 2] et des modifications de l'approche permettent d'obtenir des estimateurs sans biais de la moyenne dans le cadre MAR [11]. Les valeurs imputées sont des valeurs observées donc réalistes et elles ne nécessitent pas d'hypothèses paramétriques fortes. Elles permettent, en outre, d'imputer à la fois des variables numériques et catégorielles. Toutefois, ces méthodes produisent des estimateurs biaisés de nombreux paramètres pour tout type de mécanisme de génération des données manquantes (y compris MCAR). En particulier, ces approches ne sont pas adaptées à l'estimation des mesures d'association entre les variables [189], même si quelques solutions ont été proposées pour résoudre ce problème dans le cas de données manquantes monotones [11]. [76] montre également, sur des simulations, que la variance de l'estimateur de la moyenne est sous-estimée lorsque calculée directement sur les données imputées par hot-deck. Ce dernier problème peut être limité par l'utilisation de méthodes de ré-échantillonnage ou par l'imputation multiple (voir section 4.5.2). Enfin, [11] soulignent que hot-deck est moins sensible à une mauvaise spécification des hypothèses qui sous-tendent l'imputation (imputation hiérarchique, par plus proches voisins, ...) que les méthodes paramétriques mais que cet avantage est principalement visible lorsque la taille de l'échantillon est suffisamment grande. L'imputation hot-deck est, en effet, très dépendante de la richesse de l'ensemble des donneurs potentiels et celle-ci se dégrade rapidement lorsque la taille de l'échantillon est faible.

Enfin, les approches kNN sont principalement étudiées et évaluées d'un point de vue empirique. En particulier, [21] montrent, sous divers types de mécanismes de génération des données manquantes, que prendre $k > 1$ permet d'améliorer la qualité de l'imputation par rapport à $k = 1$ en terme d'erreur sur la valeur imputée et d'erreur quadratique moyenne sur l'estimation de diverses statistiques (coefficient de corrélation et de régression) à partir des données imputées mais l'augmentation de k tend à déformer, de manière croissante, la distribution univariée des variables imputées et, notamment, à modifier leurs variances.

Conclusion et recommandations :

— *Avantages* : faciles à mettre en œuvre ; permettent d'obtenir un jeu de données complet sur lequel n'importe quelle analyse statistique peut être pratiquée ; non paramétriques et peuvent prendre en compte divers types de distance ; préserve la distribution univariée des données (HD) ; sans biais dans le cas MCAR pour l'estimation de la moyenne (HD) ;

— *Désavantages* : déforme la distribution univariée des données (kNN) ; déforment les relations multivariées ; pas recommandée si n est faible (HD).

4.4.3 Approches par prédiction

Une approche alternative pour imputer des valeurs manquantes est d'avoir recours à des approches par prédiction. Pour imputer la valeur manquante y_{ij}, ces méthodes estiment un modèle de régression de Y_j sur les autres variables, $(Y_{j'})_{j' \neq j}$, pour lesquelles $y_{ij'}$ est observée ou sur les covariables complètement observées, X. La prédiction obtenue pour l'individu i est alors utilisée pour imputer y_{ij}.

Parmi ces méthodes, on peut citer la régression locale (ou LOESS, [52]), fréquemment utilisée. Elle consiste à construire un polynôme de faible degré, ajusté autour de la donnée manquante, par kNN. De manière plus précise, si seule la valeur y_{ij} est manquante pour l'individu i, les kNN de i sont sélectionnés parmi l'ensemble des individus pour lesquels toutes les variables sont observées. Si ces observations sont notées $(1), \ldots, (k)$, le problème de régression linéaire par moindres carrés est estimé :

$$\hat{\beta} = \arg \min_{\beta \in \mathbb{R}^{p-1}} \sum_{i'=1}^{k} \left(\beta^\top \mathbf{y}_{(i')}^{-j} - y_{i'j} \right)^2$$

où $\mathbf{y}_{(i')}^{-j}$ est le vecteur des observations des $p-1$ variables autres que Y_j pour l'individu (i'). La valeur y_{ij} est alors imputée par

$$\hat{\beta}^T \mathbf{y}_i^{-j}.$$

Cette approche se généralise de manière évidente au cas où plusieurs variables sont à imputer pour un même individu i ou bien en remplaçant la sélection des kNN par un poids décroissant en la distance entre l'individu pour lequel la valeur est manquante, i, et les autres individus, utilisés pour estimer β.

Au-delà de la méthode LOESS décrite ci-dessous, de nombreuses autres approches de régression ou de classification, paramétriques et non paramétriques, sont utilisées de la même manière pour imputer des valeurs manquantes pour des variables numériques ou catégorielles. Parmi celles-ci, on peut citer les plus courantes comme

— la régression linéaire ou sa version robuste utilisant des M estimateurs [104] ;

— les régressions linéaires pénalisées de types Lasso [211], ridge [99], elasticnet [245], régression pas à pas [98] ;

— les méthodes de régression non paramétriques comme les arbres de régression CART [32] ou les forêts aléatoires [31]. [202] proposent également une approche d'imputation par prédiction qui est itérative et fondée sur les forêts aléatoires. Celle-ci est implémentée dans le package **missForest**.

En outre, le package **simputation** est un package permettant d'effectuer de l'imputation par prédiction de manière très générique et avec une syntaxe simplifiée. Certaines méthodes de régression y sont pré-implémentées (régression linéaire, régression linéaire robuste, CART, forêts aléatoires, ...) et la fonction `impute_proxy` permet de mettre en œuvre une méthode d'imputation définie après estimation d'une fonction de prédiction arbitraire. Ainsi, par exemple, l'imputation par LOESS peut être réalisée en combinant cette fonction avec un modèle obtenu par le package **locfit**. Le package **VIM** propose également des méthodes d'imputation fondées sur la régression linéaire ou la régression linéaire généralisée (fonction `regressionImp`). L'imputation par régression est aussi utilisée dans le contexte d'études génétiques : dans celles-ci, des marques de mutation (appelées SNP) sont collectées à divers endroits du génome d'individus d'intérêt et ce type de données contient généralement un grand nombre de valeurs manquantes. Dans ce cadre-ci, le package **snpStats** (Bioconductor) propose une imputation qui combine une régression pas à pas pour sélectionner un ensemble de marqueurs permettant de bien expliquer un marqueur d'intérêt et un modèle de régression généralisé utilisant cet ensemble de marqueurs pour la prédiction.

Il existe plusieurs types d'amélioration des méthodes par prédiction :

— l'approche par *régression stochastique* se propose d'injecter un bruit aléatoire lors de l'étape de prédiction [133]. Ceci a pour objectif de limiter la sous-estimation de la variabilité et la sur-corrélation des variables imputées. Cette méthode (prédiction par régression ridge puis injection de bruit) est implémentée dans la fonction `mice.impute.norm` du package **mice** [39] (fonction `mice`) ;

— l'approche par *spécification de lois conditionnelles (FCS)* [18] [38] spécifie, de manière paramétrique et pour toute variable Y_j ayant des valeurs manquantes, la densité conditionnelle des lois $f(Y_j|Y_{-j}, R; \theta_j)$, avec Y_{-j} l'ensemble des variables différentes de Y_j et θ_j le paramètre permettant de spécifier la loi conditionnelle. Après une initialisation de l'imputation (par exemple, une imputation par la moyenne), et pour chaque variable j, traitée par ordre croissant du nombre de valeurs manquantes, deux étapes sont itérées :

  — $\theta_j^{(t)}$ est tirée aléatoirement selon la loi $p(\theta_j|Y_j = \mathbf{y}_{1,\text{obs}}, Y^{-j} = \mathbf{Y}^{-j,(t-1)})$;

  — $\mathbf{y}_1^{(t)}$ est tirée aléatoirement selon la loi $f(Y_{\text{miss}}|Y_j = \mathbf{y}_{1,\text{obs}}, Y^{-j} = \mathbf{Y}^{-j,(t-1)}; \theta_1^{(t)})$.

  L'approche est donc relativement similaire aux approches bayésiennes décrites dans la section 4.3.2 mais permet de créer des modèles de spécification des données plus flexibles, qui prend en compte les spécificités de chaque variable (contraintes de positivité, dépendances conditionnelles entre variables, ...) de manière plus naturelle. Comme les méthodes de la section 4.3.2, elle est fréquemment utilisée pour l'imputation multiple (voir section 4.5.2).

Les approches d'imputation par régression sont très largement utilisées pour produire un jeu de données complet avant analyse. Elles sont relativement flexibles, s'adaptant aux *a priori* sur les données, par l'utilisation de modèles de prédiction paramétriques ou non paramétriques. Leur performance est donc fortement dépendante de deux aspects : le premier est la capacité à pouvoir estimer des valeurs réalistes pour les valeurs manquantes à partir des valeurs observées sur les autres variables. Elles requièrent donc une dépendance entre les variables utilisées pour l'imputation et celles qui sont imputées. Elles ne couvrent donc pas non plus, *a priori*, le cas MNAR. Le deuxième aspect est la nécessité de bien spécifier la méthode de régression (ou le modèle de régression dans un cadre paramétrique) permettant

18. *Fully Conditional Specification*, en anglais.

d'imputer les variables : les approches classiques d'évaluation des méthodes de prédiction (validation croisée, ...) peuvent donc être utiles pour évaluer la fiabilité de l'approche choisie. Par ailleurs, il faut noter que l'approche est difficilement praticable lorsque certaines variables ont un fort ratio de manquants (les modèles de régression, dont la précision dépend directement du nombre de valeurs observées pour la variable à imputer, sont alors difficilement estimables) ou lorsque les valeurs manquantes entre les diverses variables sont fréquemment liées aux mêmes individus (il est alors difficile d'avoir suffisamment de variables observées pour estimer un modèle de régression) : elles sont donc mieux adaptées aux répartitions de données manquantes sans structure. Enfin, les garanties théoriques pour ces méthodes concernent principalement l'erreur commise sur la valeur imputée (par rapport à la valeur réelle non observées, et pas l'inférence statistique qui pourraient être pratiquées sur le tableau de données imputées) et découlent directement des garanties théoriques connues pour les diverses méthodes de régression utilisées.

Conclusion et recommandations :

— *Avantages* : permettent d'obtenir un jeu de données complet sur lequel n'importe quelle analyse statistique peut être pratiquée ; flexibles (large choix d'approches de régression) ;

— *Désavantages* : principalement valables dans le cas MAR ; requièrent une bonne spécification de la méthode de régression ; requièrent une bonne prédictibilité des variables ayant des valeurs manquantes par les autres variables ; cadre théorique lié à l'erreur quadratique sur la valeur imputée (et non aux résultats de l'analyse statistique pratiquée).

4.4.4 Approches factorielles pour l'analyse exploratoire

Il est important de souligner qu'un grand nombre de travaux étudiant le traitement des données manquantes se placent dans un cadre inférentiel (c'est le cas, par exemple, de l'ouvrage de référence de [133]). Ceux-ci peuvent ne pas être bien adaptés à un cadre exploratoire comme l'analyse de données, dans lequel des critères géométriques sont privilégiés par rapport aux hypothèses de nature probabilistes. Parmi les analyses exploratoires, l'Analyse en Composantes Principales (ACP) tient une place importante et son extension en présence de valeurs manquantes a été largement étudiée ([115] et [106]). De nombreux problèmes sont soulignés pour la pratique de l'ACP en présence de manquants : difficulté pour le centrage et la réduction des variables, non unicité de la solution de minimisation de la fonction de coût classique en ACP, extension non triviale de la notion de base de l'ACP, ...

Dans l'étude de l'ACP en présence de valeurs manquantes, deux objectifs complémentaires sont visés : celui de la réalisation d'une ACP en présence de valeurs manquantes et celui de l'utilisation de l'ACP pour imputer des valeurs manquantes. Dans le cadre d'études de simulations où des données manquantes sont produites de manière artificielle pour évaluer la qualité des algorithmes (sur-imputation ; voir section 4.5.1), ces deux objectifs sont évalués par des métriques de performance différentes [115] : coefficient RV [75] entre les coordonnées des individus sur les données complètes par rapport aux coordonnées produites par les approches d'ACP adaptées, d'une part, et erreur de reconstitution entre valeurs initiales et valeurs imputées, d'autre part.

De nombreuses variantes des methodes de prises en compte des valeurs manquantes dans l'ACP ont été proposées dont les principales sont :

— **Nonlinear Iterative Partial Least Squares (NIPALS)** [234]. Le principe de cette méthode est aussi à la base de la régression PLS (*Partial Least Squares* ; [209]). Il permet de réaliser une ACP avec données manquantes sans supprimer les individus i pour lesquelles une valeur y_{ij} est manquante et sans imputer les valeurs manquantes. En ce sens, la méthode se rapproche des méthodes fondées sur l'analyse des cas disponibles, décrites dans la section 4.2.2, mais elle peut, en outre, être utilisée comme base pour l'imputation des valeurs manquantes.

De manière plus précise, si on suppose les variables $(Y_1, ..., Y_p)$ centrées, l'algorithme NIPALS utilise la formule de décomposition de l'ACP suivante :

$$\mathbf{Y} \simeq \sum_{h=1}^{d} \mathbf{t}_h \boldsymbol{\rho}_h^\top$$

où $d \leq p$ est la dimension de projection permettant d'obtenir une « bonne » reconstitution des données et $\{\mathbf{t}_h\}_{h=1,...,d} \subset \mathbb{R}^n$ et $\{\boldsymbol{\rho}_h\}_{h=1,...,d} \subset \mathbb{R}^p$ sont, respectivement, les composantes principales et les vecteurs directeurs des axes principaux de l'ACP. Ceci implique que les observations de la variable Y_j peuvent s'écrire comme une régression linéaire sur les composantes $(\mathbf{t}_h)_h$: $Y_j = \sum_{h=1}^{d} \rho_{hj} \mathbf{t}_h$ (et respectivement pour l'individu i qui peut être écrit comme une régression sur les axes principaux).

L'algorithme NIPALS utilise cette remarque et estime, de manière itérative et jusqu'à convergence, les $(\boldsymbol{\rho}_h)_h$ et les $(\mathbf{t}_h)_h$ par régressions successives sur les valeurs observées, en initialisant les composantes principales, par exemple, à une colonne de $\mathbf{Y}$. Contrairement à l'approche standard de l'ACP où les axes sont déterminés simultanément par décomposition spectrale, l'approche NIPALS calcule les axes successivement en utilisant une étape de déflation.

Une fois les $(\mathbf{t}_h)_{h=1,...,d}$ et les $(\boldsymbol{\rho}_h)_{h=1,...,d}$ estimés, il est possible de proposer une estimation des valeurs manquantes en utilisant la formule de reconstitution des individus :

$$\widehat{y}_{ij} = \sum_{h=1}^{d} t_{hi} \rho_{hj}. \tag{4.4}$$

En pratique, l'approche NIPALS fournit des solutions raisonnables lorsque le taux de manquant est faible mais elle souffre de plusieurs désavantages. Le premier est que lorsqu'une proportion importante de valeurs sont manquantes, la procédure itérative de NIPALS propage les erreurs d'axe en axe et sa convergence n'est pas garantie. Par ailleurs, si l'ACP est pratiquée sur les données centrées et réduites, NIPALS ne peut réaliser une mise à jour de l'écart-type des variables (à cause de la déflation) et produit donc un résultat qui ne correspond pas à une ACP réduite. Enfin, les axes obtenus ne sont pas nécessairement orthogonaux et le critère classique de minimisation de l'erreur de reconstitution de l'ACP,

$$\sum_{i=1}^{n} \left\| \mathbf{y}_i - \sum_{h=1}^{d} t_{hi} \boldsymbol{\rho}_h \right\|^2, \tag{4.5}$$

n'est pas minimisé par la procédure séquentielle.

— **ACP itérative** [120]. L'ACP itérative est une approche itérative qui vise à minimiser l'erreur de reconstitution de l'ACP (équation (4.5)). L'initialisation de la méthode attribue une valeur arbitraire aux données manquantes (souvent la moyenne de la variable considérée). Une ACP est ensuite effectuée sur ce jeu de données rendu complet et les données initialement manquantes sont alors mises à jour via la formule de reconstitution de l'équation (4.4). Les deux étapes d'estimation de l'ACP et d'imputation sont répétées jusqu'à convergence, [120] montrant que la procédure converge nécessairement, éventuellement vers un minimum local.

En raison de l'alternance des étapes d'estimation et d'imputation, similaires aux étapes *Expectation* et *Maximization* des algorithmes EM, l'ACP itérative est souvent appelée ACP-EM. En effet, l'ACP peut être vue comme un modèle statistique dans lequel les données ont une structure dans un espace à faible dimension (d) et sont corrompues par un bruit [40]. Cette formulation se ré-écrit sous la forme d'un modèle à effet fixe [43]

$$y_{ij} = \sum_{h=1}^{d} t_{hi}\rho_{hj} + \epsilon_{ij}, \qquad \epsilon_{ij} \sim \mathcal{N}(0, \sigma^2) \text{ (i.i.d.)}, \tag{4.6}$$

que [115] utilisent pour montrer que l'ACP itérative peut effectivement être vue exactement comme un algorithme EM et bénéficie donc des propriétés et des caractéristiques de ces approches.

Toutefois, l'approche souffre d'un problème de sur-ajustement aux données, particulièrement dans les cas de grande dimension ($p > n$) [115]. Aussi, pour pallier le problème du sur-ajustement, la version régularisée de l'ACP itérative lui est préférée. La régularisation peut être effectuée en choisissant une dimension réduite, $d \ll p$, pour la reconstitution ou bien en ajoutant un terme de pénalité en norme ℓ_2 (*ridge*) lors de l'étape d'imputation. [219] montrent que l'ACP régularisée *ridge* peut être vue comme une extension de l'équation (4.6) au modèle mixte

$$\mathbf{y}_i = \mathbf{R}\mathbf{t}_i + \epsilon_i, \tag{4.7}$$

où $\mathbf{R}$ est une matrice de dimension $p \times d$, $\mathbf{t}_i \sim \mathcal{N}(0, \mathbb{I}_d)$ et $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ (i.i.d.). Ce modèle, connu sous le nom d'« ACP probabiliste », est proposé initialement dans [213].

— **ACP bayésienne** ([106] et [219]). Diverses approches bayésiennes sont proposées dans la littérature pour l'ACP, fondées sur le modèle à effets fixes de l'équation (4.6) ou le modèle d'ACP probabiliste de l'équation (4.7). En particulier, [219] montrent que l'ACP probabiliste peut être vue comme un traitement bayésien des effets fixes du modèle de l'équation (4.6) ou bien comme un traitement bayésien direct des effets fixes avec le modèle

$$\mathbf{y}_i = \widetilde{\mathbf{y}}_i + \epsilon_i, \qquad \widetilde{\mathbf{y}}_i \sim \mathcal{N}(0, \tau_d)$$

où la matrice $\widetilde{\mathbf{Y}} = \begin{pmatrix} \widetilde{\mathbf{y}}_1 \\ \vdots \\ \widetilde{\mathbf{y}}_n \end{pmatrix}$ est de dimension $n \times d$. [106] proposent d'autres *a priori* bayésiens et font le lien entre diverses variantes de l'ACP probabiliste. Ils proposent également des versions rapides de l'estimation, utilisant des approches en ligne ou des approximations variationnelles, qui montrent des résultats encourageants sur les données de la compétition Netflix (2007) (qui consiste à compléter un tableau de notes

de $p = 17\ 770$ films évalués par $n = 480\ 189$ spectateurs et contenant plus de 98% des données manquantes).

Enfin, comme beaucoup de méthodes d'analyse factorielle s'apparentent à l'ACP, il est possible d'étendre l'imputation par ACP à celles-ci. Ainsi, une méthode d'imputation fondée sur l'Analyse des Correspondances Multiples (ACM), proposée par [14], permet de gérer l'imputation de variables catégorielles et une méthode fondée sur l'Analyse Factorielle Multiple (AFM), proposée par [116], permet de prendre en compte la structuration d'un jeu de données en blocs de variables. De même, une méthode fondée sur l'Analyse Factorielle des Données Mixtes (AFDM) de [13] permet d'imputer des données mixtes (catégorielles et numériques). Les approches d'ACP en présence de valeurs manquantes ont également été étendues au cadre de l'imputation multiple (voir section 4.5.2) par [113] pour l'ACP itérative et [15] pour l'ACP bayésienne.

Les méthodes factorielles qui prennent en compte les valeurs manquantes sont implémentées dans plusieurs packages R dont les principaux sont :

— **ade4** [46] qui permet l'analyse exploratoire de données écologiques et environnementales et propose une implémentation de NIPALS ;

— **missMDA** [116] qui propose des implémentations de plusieurs méthodes d'analyse factorielle en présence de valeurs manquantes ;

— **mixOmics** [129] qui propose des méthodes d'analyses multivariées pour l'exploration et l'intégration de données biologiques (en particulier les données 'omiques) et impute les valeurs manquantes avec l'approche NIPALS ;

— **pcaMethods** [200] qui est un package Bioconductor [19] qui propose de nombreuses méthodes d'ACP en présence de valeurs manquantes (dont NIPALS, les méthodes d'ACP probabiliste et d'ACP bayésienne) ainsi que des outils pour la validation croisée et la visualisation des résultats.

Conclusion et recommandations :

— *Avantages* : bien adaptées à l'analyse exploratoire ; garanties théoriques fondées sur les modèles à effets fixes ou mixtes ; variantes adaptées à la grande dimension et au grand volume ;

— *Désavantages* : cadre théorique restreint aux modèles de génération des données fondés sur les modèles à effets fixes ou mixtes décrits plus haut : mêmes limitations que celles décrites dans la section 4.3.

4.4.5 Conclusions sur l'imputation simple

Dans cette section, nous avons présenté les principales méthodes d'imputation simple, en les catégorisant en trois grandes familles : complétion stationnaire, imputation fondée sur des similarités entre individus et méthodes de prédiction. Dans le cadre particulier des analyses factorielles, nous avons aussi présenté les approches développées spécifiquement pour ces cas-ci.

La complétion stationnaire est probablement l'approche la plus simple et la plus rapide. Pour ces raisons, elle peut apparaître comme très attractive. Cependant, même pour des taux de manquants relativement faibles, cette approche n'est pas recommandée car elle ignore

19. `https://www.bioconductor.org`

les relations de corrélation entre variables et entre individus, elle sous-estime fortement la variabilité des variables imputées et en déforme leurs distributions.

Les méthodes qui utilisent une information de ressemblance entre individus (comme les approches hot-deck) sont particulièrement bien appropriées dans le cas de données discrètes (catégorielles ou numériques discrètes). D'une manière générale, toutefois, si elles préservent la distribution univariée des données, elles tendent à fortement déformer les corrélations entre variables. Dans le cas où le jeu de données contient des individus avec un grand nombre de valeurs manquantes, des individus entiers peuvent être utilisés pour imputer toutes les valeurs manquantes comme le suggèrent [226]. Dans ce cas, elles permettent de mieux conserver les relations de corrélation entre variables et sont donc bien adaptées au cas où des analyses factorielles ou une inférence de réseaux sont réalisées après l'imputation comme dans [107]. Toutefois, elles nécessitent de pouvoir obtenir une mesure de ressemblance ou une distance entre individus, ce qui peut être réalisé par l'utilisation de covariables complètement observée. Le choix de la distance et la nécessité d'avoir des données permettant de la calculer sont donc également deux limitations de la méthode.

Les approches d'imputation qui utilisent des méthodes de régression ou une modélisation jointe (comme les approches paramétriques multivariées de la section 4.3 ou les approches factorielles) sont généralement mieux adaptées pour la modélisation de la loi jointe des variables. Elles sont plus difficiles à mettre en œuvre, en général, que les approches précédentes, nécessitent la définition correcte d'un modèle de loi jointe des données ou d'une méthode de régression dont la qualité de l'analyse dépend fortement. Dans le cas d'approches paramétriques, il est parfois possible d'obtenir une estimation de la variabilité du paramètre de la loi (voir section 4.5.3) et elles fournissent donc, par ce biais, une information sur l'incertitude liée à l'imputation.

Néanmoins, au sein d'un même jeu de données, il peut s'avérer utile d'utiliser une combinaison d'approches pour s'adapter au mieux aux spécificités de chaque variable ou chaque individu contenant des valeurs manquantes. La démarche standard consiste à commencer par une analyse exploratoire des valeurs manquantes puis, selon la distribution de celles-ci par variable et par individu, et les corrélations connues entre variables, à supprimer les variables et individus ayant un fort taux de manquants (s'ils sont peu nombreux) puis à combiner diverses méthodes d'imputation (par prédiction, par hot-deck, etc) selon la variable ou l'individu à imputer. Le package **simputation** permet de gérer facilement ce type d'approches en proposant une collection de méthodes standard pour l'analyse exploratoire des données manquantes et leur imputation. Enfin, il est recommandé de chercher à estimer l'incidence de l'imputation sur les analyses pratiquées *a posteriori*, par exemple en estimant l'incertitude liée à l'imputation (voir section 4.5). Des conseils pratiques détaillés sont fournis sur le site décrivant les grandes lignes directrices en matière de qualité dans le traitement des enquêtes de l'organisme public « Statistique Canada » [20] ainsi que par [77].

Enfin, l'imputation doit parfois être adaptée aux particularités du jeu de données. Par exemple, une approche pour l'imputation de variables ordinales est proposée dans [79]. Celle-ci alterne une ACP non linéaire et une imputation par kNN et est implémentée dans le package **ForImp**. Également, l'imputation de séries chronologiques peut être pratiquée en tenant compte de la tendance observée au cours du temps avec des approches par interpolation, par ajustement d'une courbe de lissage ou par estimation d'un modèle de régression longitudinale (ARIMA, par exemple, voir [122]). Les méthodes les plus courantes d'imputation de séries temporelles sont implémentées dans le package **imputeTS** [153]

20. `https://www.statcan.gc.ca/pub/12-539-x/2009001/imputation-fra.htm`

qui, à ce jour, est l'unique package d'imputation de données uniquement dédié aux séries temporelles. D'autres packages dont **zoo** [241] et **forecast** incluent aussi des méthodes d'imputation pour les séries temporelles qui sont relativement sophistiquées. Également, les packages **spacetime** [160], **timeSeries** et **xts** incluent des approches plus basiques pour l'imputation de séries temporelles. Une comparaison des diverses méthodes d'imputation de séries temporelles est effectuée dans [154] qui montrent que les méthodes d'imputation les plus efficaces pour ce type de données sont fondées sur une prise en compte de la saisonnalité de la série temporelle.

4.5 Variabilité et fiabilité de l'imputation

Dans les méthodes d'imputation simple, il est fréquent qu'une valeur manquante soit remplacée par sa valeur imputée et qu'elle joue, dans la suite de l'analyse, le même rôle que les valeurs observées. Le risque est fort de biaiser ces analyses *a posteriori*, sans contrôle de l'incertitude liée à l'imputation. Par exemple, dans le cas de l'estimation d'un paramètre à partir des données, la variance du paramètre est souvent sous-estimée même si le modèle d'imputation est correctement spécifié (voir section 4.3).

On peut distinguer diverses approches pour aborder cette problématique : la première consiste à utiliser des outils diagnostiques destinés à évaluer la fiabilité de l'imputation. Cette question est discutée dans la section 4.5.1 et cherche à identifier des erreurs dans l'estimation de la valeur imputée par rapport à la valeur qui aurait dû être observée.

La seconde se concentre sur l'estimation de la variabilité liée au processus d'imputation. D'une part, elle fournit un diagnostic sur la fiabilité ou le domaine de validité des conclusions de l'analyse et, d'autre part, elle améliore la qualité de l'analyse elle-même (par des méthodes d'agrégation par exemple). Dans ce cadre, une approche fréquemment utilisée est l'*imputation multiple* que nous décrivons dans la section 4.5.2. La section 4.5.3 décrit les alternatives à cette approche dans le cadre particulier de l'algorithme EM et la section 4.5.4 conclut la section par une courte discussion sur ces diverses approches.

4.5.1 Outils de diagnostic

Les valeurs imputées étant des valeurs estimées, il est important de vérifier si elles sont plausibles. Pour cela, il est possible d'utiliser des outils de diagnostic. Cela consiste généralement à comparer les valeurs imputées aux valeurs observées soit à l'aide de graphiques, soit à l'aide de statistiques élémentaires.

Sur-imputation

La première approche pour évaluer la qualité d'une méthode d'imputation est de procéder par sur-imputation [21] en supprimant des données observées et en comparant les valeurs imputées aux valeurs réelles avant suppression, notamment par calcul de l'erreur quadratique moyenne (MSE) ou de sa racine carrée (RMSE), comme proposé dans les

21. *Overimputation* en anglais.

packages **Amelia** et **missMDA**. Cette approche est relativement intéressante pour évaluer la qualité d'une méthode donnée.

Une approche alternative consiste à utiliser uniquement valeurs observées et leur distribution pour évaluer la pertinence des valeurs imputées.

Outils généraux de diagnostic

De manière plus avancée et systématique, [1] et [204] proposent trois types de diagnostic pour des données multivariées. La première approche consiste à représenter, de manière graphique, les données elles-mêmes (au travers, par exemple, de nuages de points) en différenciant valeurs observées et valeurs imputées. Ces graphiques permettent de repérer facilement des valeurs atypiques dans l'imputation, signe par exemple, d'un problème potentiel dans le choix de la méthode d'imputation.

La second type de diagnostic consiste à comparer, pour chaque variable, les densités entre valeurs imputées et celles observées en utilisant un test de Kolmogorov-Smirnov et en réalisant des graphiques diagnostiques (histogramme, courbe de densité, ...). Ceux-ci ont pour but de permettre, pour chaque variable, la comparaison visuelle entre les distributions des valeurs observées et les distributions des valeurs imputées. Les différences entre les valeurs imputées et observées ne sont pas forcément dues à un problème d'imputation. Il est possible qu'un sous-groupe de la population ait plus de données manquantes pour certaines variables. Ainsi, les graphiques diagnostiques permettent de mettre en évidence ces variables pour mieux les étudier.

Le dernier type de diagnostic utilise le fait que les imputations sont générées par des modèles ajustés sur les données observées. Il est donc possible de vérifier la qualité de l'ajustement de ces modèles en comparant la valeur prédite, pour un individu et une variable donnés, à la valeur observée ou bien en utilisant les outils diagnostiques spécifiques d'un modèle donné (graphique des résidus, QQ plot pour un modèle linéaire, par exemple). Ce type de diagnostic se rapproche de la sur-imputation dans la comparaison entre valeur observée et valeur prédite.

Enfin, de manière similaire, et au-delà du cas MAR, [193] étudient le cas de la régression linéaire multiple avec une covariable ayant des valeurs manquantes et proposent des formules explicites pour la dépendance entre le paramètre à estimer ou le coefficient de corrélation de la régression linéaire et les valeurs manquantes. Sous l'hypothèse d'une dépendance linéaire entre la variable contenant des manquants et les autres covariables, les auteurs proposent des graphiques permettant d'étudier l'effet potentiel des valeurs manquantes sur la régression qui peuvent être utilisées comme diagnostics pour évaluer la pertinence de l'imputation dans ce cadre-ci.

Les packages **mi** et **VIM** proposent différents graphiques diagnostiques. Pour comparer les distributions, le package **VIM** fournit divers graphiques uni et bi-variés représentant de manière séparée ou simultanée les valeurs observées et les valeurs imputées. Par exemple, pour les variables « drink_days » (nombre de jours, au cours du dernier mois, où la personne a bu au moins un verre d'alcool) et « health_poor » (nombre de jours, au cours du dernier mois, où la personne n'a pu pratiquer une activité « habituelle » à cause de problèmes de santé), la figure 4.5 montre les distributions univariées des valeurs imputées et observées pour les deux variables et un nuage de points sur lequel les points correspondant à au moins une valeur imputée sont mis en valeur par une couleur distincte. Les densités des valeurs imputées et observées sont similaires et aucune répartition spécifique des points correspondant à des valeurs imputées n'est repérable sur le nuage de points, ce qui est un

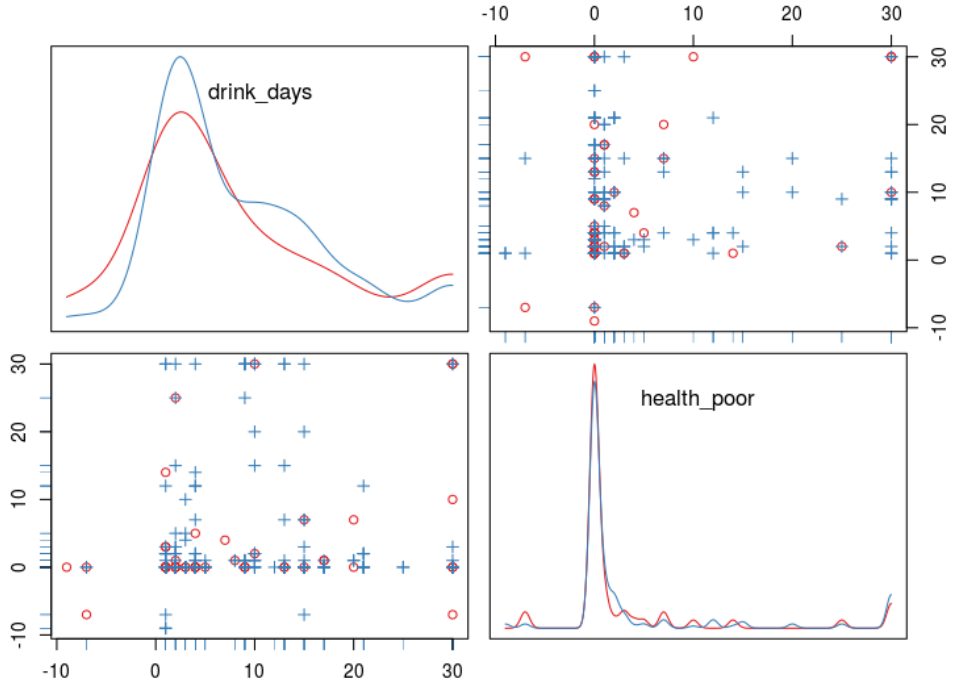


Figure 4.5 Graphiques des distributions univariées (densités) des variables « drink_days » (en haut à gauche) et « health_poor » (en bas à droite) pour les valeurs manquantes (en rouge) ou observées (en bleu). Nuage de points des deux variables (en haut à droite et en bas à gauche).

indicateur positif de la fiabilité de l'imputation. Le package **mi** utilise l'approche d'imputation FCS décrite dans la section 4.4.3 et fournit un graphique contenant distribution du tableau de données imputées et observées (par un histogramme) et graphiques comparant valeurs prédites et résidus aux valeurs observées (voir figure 4.6 pour la variable « weight_lbs »).

Figure 4.6 Exemple de graphiques diagnostiques fourni par le package **mi** (pour la variable « weight_lbs ») : histogramme des valeurs observées et imputées, valeurs imputées (prédites) et résidus en fonction des valeurs observées.

Erreur d'imputation et décomposition dans le cas des k-plus proches voisins

Comme indiqué dans la section 4.5.1, l'estimation de l'erreur d'imputation est souvent limitée à la comparaison entre valeurs observées et valeurs imputées. Dans [201], les auteurs vont au-delà et proposent de décomposer l'erreur d'imputation en :

— *erreur de mesure*, qui est l'erreur commise entre les valeurs observées, y_{ij} et la « vraie » valeur de Y_j pour l'individu i, y^*_{ij} (qui reste inconnue en raison d'erreurs liées aux appareils de mesure ou bien de différences expérimentales incontrôlées entre les mesures par exemple). Contrairement au cadre habituel (qui suppose cette erreur nulle), le cadre de l'article de [201] est celui d'erreurs de mesure non nulles mais qui ne présentent pas de biais et qui sont indépendantes de covariables complètement observées, X ;

— et *erreur pure* (qui peut être vue comme une erreur du modèle d'imputation) qui est spécifiée dans le cadre d'une approche d'imputation dans laquelle la variable avec des valeurs manquantes Y_j est imputée à partir d'un modèle faisant uniquement intervenir des covariables complètement observées X. Dans ce cadre-ci, l'erreur pure s'écrit :

$$y^*_{ij} - g_j(\mathbf{x}_i)$$

où g_j est la fonction de prédiction permettant l'imputation de la valeur de Y_j. C'est cette erreur qui est d'intérêt pour diagnostiquer la méthode d'imputation choisie.

Dans le cadre de l'imputation par la méthode kNN et lorsque $k = 1$, ils montrent que l'on peut estimer l'erreur d'imputation pour la variable à imputer Y_j, à partir de la différence d'erreur quadratique moyenne (MSD) :

$$\text{MSD}_j = \frac{\sum_{i=1}^{n} r_{ij} \left(y_{ij} - y_{\mathcal{N}_1(i),j}\right)^2}{\sum_{i=1}^{n} r_{ij}}$$

où $\mathcal{N}_1(i)$ est le plus proche voisin de i, parmi les individus pour lesquels Y_j est observée, au sens de la distance sur X comme définie dans l'équation (4.3). Enfin, ils proposent d'estimer l'*erreur standard d'imputation* (SEI) par

$$\text{SEI}_j^2 = \text{MSD}_j - \frac{1}{2}\text{MMSD}(0)_j$$

où $\text{MMSD}(0)_j$ est la valeur de MSD obtenue pour une petite fraction des paires d'individus ayant les plus petites distances entre eux, non pas au sens de l'équation (4.3) mais au sens de la distance de Mahalanobis (ces paires étant utilisées pour estimer l'erreur de mesure).

Cette proposition est généralisée aux cas où $k > 1$ en utilisant la valeur moyenne des kNN. Ces erreurs diagnostiques sont proposées dans le package R **yaImpute** [60].

4.5.2 Imputation multiple

Pour tenter de mesurer l'impact de l'imputation et pour quantifier l'erreur commise lors de celle-ci, l'approche la plus répandue consiste à répéter l'imputation plusieurs fois en introduisant de l'aléa. Ces approches sont connues sous le nom d'imputation multiple.

Principe de l'imputation multiple

L'imputation multiple ([183], [182] et [188]) consiste à proposer, pour chaque valeur manquante, non pas une mais plusieurs valeurs plausibles pour l'imputation. Cette méthode permet de mesurer la variabilité, sur le résultat final, du processus d'imputation.

L'imputation multiple se déroule en trois phases, représentées sur la figure 4.7 :

Phase d'imputation Le tableau de données initiales est dupliqué M fois et un modèle d'imputation est appliqué sur chaque nouveau tableau de données. Une part d'aléa est introduite, soit au niveau de la duplication du tableau initial (qui n'est pas reproduit à l'identique), soit au niveau de l'imputation elle-même, ce qui permet l'obtention de M tableaux différents de données complètes ;

Phase d'analyses statistiques L'analyse statistique retenue (régression, ACP, inférence de réseau, ...) pour analyser le tableau de données est mise en œuvre sur chacun des $m = 1, ..., M$ tableaux de données imputées pour obtenir M estimations ;

Phase d'analyse combinée Les M résultats obtenus sont combinés selon les règles définies par [183] pour obtenir une seule estimation finale ou pour estimer la variabilité des résultats par une analyse statistique cible pratiquée sur les données complétées.
Les procédures d'imputation qui incorporent une variabilité appropriée à travers les M jeux de données imputées dans le modèle sont dites « adéquates[22] » au sens de [183] ou [133] : cela signifie que ces méthodes d'imputation reflètent correctement la variabilité de la méthode fondée sur les données imputées, en prenant en compte, à la fois, la variabilité intra-imputation (correspondant à la variabilité due à la méthode elle-même et au bruit dans les données) et la variabilité inter-imputation (attribuable à la présence de données manquantes).

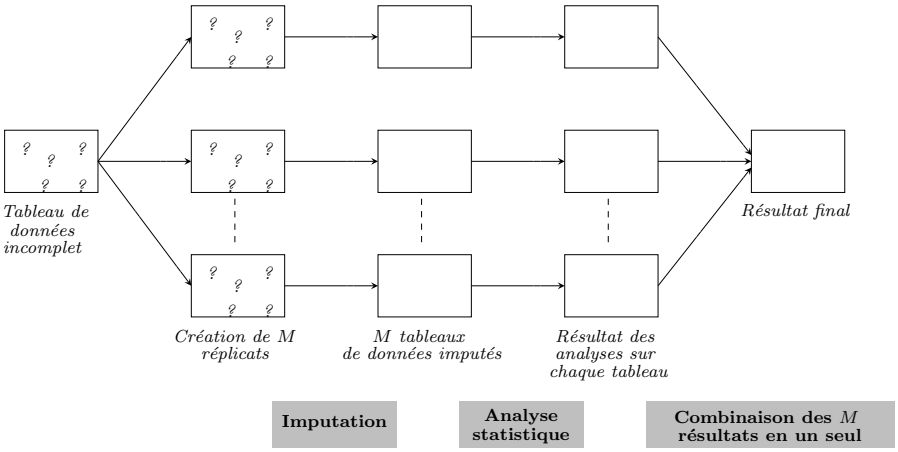


Figure 4.7 Schéma de l'imputation multiple

Ces diverses étapes et les approches principales pour leurs mises en œuvre sont décrites dans les sections suivantes.

Phase d'imputation

Plusieurs approches permettent d'obtenir des tableaux de données imputées différents, fondées soit sur des perturbations de l'échantillon initial, soit sur l'introduction d'un processus aléatoire dans l'imputation elle-même.

Approche par ré-échantillonnage Dans les approches par ré-échantillonnage, l'aléa est introduit au moment de la duplication du tableau de données initiales en M copies. Au

22. *proper* en anglais.

lieu de dupliquer le tableau initial, un sous-échantillonnage ou un ré-échantillonnage sont pratiqués pour obtenir M copies « perturbées » du tableau de données initiales. En pratique, les approches bootstrap ou bien Jackknife (avec $M = n$) sont les plus utilisées. L'analyse statistique conduit donc, par exemple, à l'estimation d'un paramètre θ par M valeurs $\hat{\theta}^{(m)}$ (pour $m = 1, \ldots, M$) qui sont les estimations obtenues par la méthode statistique cible à partir d'un tableau obtenu par ré-échantillonnage ou sous-échantillonnage puis imputation. L'approche par bootstrap est, par exemple, utilisée par [117] et [14] pour estimer la variabilité de la position d'un individu avec des valeurs manquantes dans l'ACP ou l'ACM. Une approche d'imputation multiple Jackknife pour estimer la variabilité d'un estimateur dans le cadre de l'imputation hot-deck est décrite par [34] : [167] montrent toutefois que celle-ci peut entraîner des biais importants et proposent une alternative fondée sur un estimateur Jacknife corrigé qui n'utilise qu'une imputation simple.

Approche de type « hot-deck » Les approches de type « hot-deck » (voir section 4.4.2) conduisent à la création, pour chaque valeur manquante, d'un ensemble de « donneurs » correspondant à un ensemble de valeurs plausibles pour la valeur manquante considérée. En effectuant un tirage aléatoire dans ce ensemble de donneurs pour chaque valeur manquante, M tableaux de données imputées différents sont obtenus [58].

Approche bayésienne Dans les approches bayésiennes (section 4.3.2), la phase d'imputation finale est fondée sur un échantillonnage selon la loi $f(Y_{\text{miss}}|Y_{\text{obs}}, \theta^{(T)})$ où T est le nombre d'itérations de l'algorithme et $\theta^{(T)}$ l'estimation courante du paramètre qui régit la loi jointe, θ. Il est donc possible d'utiliser cette approche pour générer M tableaux de données imputées différents. Cette approche est utilisée dans [39] et [205] pour une imputation fondée sur une méthode FCS (voir section 4.4.3) et par [15] pour une imputation multiple par ACP bayésienne.

Combiner les résultats : cas de l'estimation d'une quantité numérique α et estimation de la variance de l'estimation

Lorsque le but de l'analyse statistique est l'estimation d'une quantité numérique α, l'approche la plus fréquente pour combiner les résultats des M analyses statistiques après imputation est le simple calcul de l'estimateur moyen $\bar{\alpha}$ [133] :

$$\bar{\alpha} = \frac{1}{M} \sum_{m=1}^{M} \hat{\alpha}^{(m)}.$$

Dans le cas d'une approche par imputation multiple fondée sur le Jackknife, l'approche standard consiste à imputer le jeu de données entier avec une approche quelconque puis à obtenir $M = n$ estimateurs $\hat{\alpha}^{(m)}$ à partir des échantillons imputés correspondants aux individus $\{1, \ldots, n\} \setminus \{m\}$. L'estimation de α est alors réalisée de manière standard pour les approches Jackknife, en calculant la moyenne des pseudo-valeurs

$$\hat{\alpha} = \hat{\alpha}^{(0)} + (n-1)(\hat{\alpha}^{(0)} - \bar{\alpha}), \tag{4.8}$$

où $\hat{\alpha}^{(0)}$ est l'estimateur de α obtenu à partir de l'échantillon entier après imputation [183].

La variance de l'estimateur $\bar{\alpha}$ est, quant à elle, obtenue par

$$\text{Var}(\bar{\alpha}) = \underbrace{\frac{1}{M}\sum_{m=1}^{M}\text{Var}(\hat{\alpha}^{(m)})}_{\text{variance intra-imputation: } W} + \underbrace{\frac{1}{M-1}\sum_{m=1}^{M}(\hat{\alpha}^{(m)}-\bar{\alpha})^2}_{\text{variance inter-imputation: } B}$$

où B s'obtient directement à partir des m estimateurs $\hat{\alpha}^m$ et W dépend de la méthode employée pour obtenir cet estimateur (classiquement, par exemple, lorsque $\hat{\alpha}^{(m)}$ est une moyenne empirique, W s'obtient à partir des M variances empiriques des observations des M tableaux de données imputées). L'approximation de la variance peut être améliorée en multipliant B par $(1+\frac{1}{M})$ afin de prendre en compte le fait que les estimations de α ne sont que des approximations obtenues pour un nombre fini de tableaux, M : une variabilité supplémentaire, correspondant à l'erreur de simulation, peut être ajoutée et la variance totale de $\bar{\alpha}$ est alors estimée par

$$W + \frac{M+1}{M}B.$$

Dans le cas où l'imputation multiple est réalisée avec une approche bootstrap ou Jackknife, on peut aussi obtenir une estimation de la variance de l'estimateur sans avoir besoin d'un estimateur de $\text{Var}(\hat{\alpha}^{(m)})$, en utilisant les échantillons dit « *out-of-bag* » (non sélectionnés dans l'échantillon bootstrap courant, pour l'approche bootstrap) ou bien par

$$\frac{1}{n(n-1)}\sum_{m=1}^{n}(\tilde{\alpha}^{(m)}-\hat{\alpha})^2,$$

avec $\tilde{\alpha}^{(m)} = n\hat{\alpha}^{(0)} - (n-1)\hat{\alpha}^{(m)}$ et les autres notations comme dans l'équation (4.8), pour l'approche Jackknife.

Autres approches pour la combinaison

Les approches décrites dans la section précédente ne permettent la combinaison des résultats que dans le cadre de l'estimation d'une quantité numérique. Lorsque les analyses statistiques pratiquées sur les M tableaux de données imputées produisent des résultats sous une forme plus complexe, d'autres approches peuvent être mises en œuvre soit pour visualiser la variabilité due à l'imputation, soit pour combiner les résultats.

[116] proposent l'utilisation de l'imputation multiple en ACP pour obtenir des ellipses de confiance (sous hypothèse de distribution gaussienne) autour de la projection des individus dans l'ACP. Pour cela, une projection de référence est obtenue par ACP itérative et les résultats d'imputations multiples sont utilisées pour représenter les individus imputés comme individus supplémentaires, permettant ainsi l'estimation des contours des ellipses de confiance.

Lorsque le but de l'imputation multiple n'est pas seulement l'estimation de la variabilité de l'imputation mais aussi la définition d'un résultat « combiné » obtenu à partir de plusieurs imputations, diverses stratégies alternatives au calcul de la moyenne sont proposées : dans le cadre d'analyses factorielles, [226] proposent d'utiliser la méthode STATIS [127] pour combiner les différentes configurations obtenues lors d'une AFM (Multiple Factor Analysis, [74]) réalisée par imputation multiple : cette approche recherche une projection consensuelle, c'est-à-dire une projection la plus corrélée aux M projections obtenues à partir des données imputées. Enfin, [107] proposent une approche fondée sur l'analyse de la fréquence de prédiction d'une arête dans le cas où l'analyse statistique est une inférence de réseau : cette

approche permet de ne conserver que les arêtes dont la prédiction est peu affectée par la valeur imputée et, ainsi, de diminuer le taux de faux positifs dans l'inférence.

Packages R

Divers packages proposent des implémentations pour effectuer des imputations multiples avec des approches différentes pour la partie imputation :

— **Amelia** propose une méthode d'imputation multiple fondée sur une approche par modélisation jointe gaussienne (estimée par EM ou par approche bayésienne), combinée à une imputation multiple par bootstrap (dans le cadre EM) ou bayésienne ;

— **hot.deck** propose une version multiple de l'imputation hot-deck fondée sur le score d'affinité proposé par [58] ;

— **jomo** et **pan** sont deux packages qui proposent de nombreux modèles d'imputation par modélisation jointe (approches bayésiennes) dans un cadre d'imputation multiple dit « multi-niveaux », c'est-à-dire lorsque les individus sont stratifiés en classes ;

— **mi** propose des méthodes d'imputation multiple avec une approche dite par « équations chaînées », qui est une approche bayésienne fondée sur la méthode FCS (voir section 4.4.3). Le package contient un grand nombre de modèles pour variables numériques ou catégorielles, des approches par injection de bruit pour limiter les problèmes dûs aux colinéarités entre variables et propose également divers outils de diagnostic pour évaluer la fiabilité du modèle choisi ;

— **mice** est un des packages les plus utilisés pour l'imputation multiple. L'introduction de l'aléa dans l'imputation est réalisée via l'approche par équations chaînées (comme **mi**). Le package permet de traiter des variables de types variés (catégorielles ou numériques) et contient plusieurs outils diagnostiques ;

— **missMDA** propose des méthodes pour l'imputation multiple en analyse factorielle, soit par modélisation bayésienne, soit par approche bootstrap. L'imputation multiple est utilisée ici pour visualiser la variabilité de la projection sur les axes de l'ACP ou de l'AFM obtenus par imputation simple (ACP itérative) ou pour générer des valeurs multiples d'imputation par ACP (section 4.4.4) ;

— **mitools** permet de combiner des résultats d'imputations multiples de manière générique en aggrégeant n'importe quel résultat obtenu en combinant plusieurs imputations obtenues par ailleurs ;

— **MixedDataImpute** et **NPBayesInput** sont deux packages proposant des approches de modélisation jointe (approches bayésiennes) pour l'imputation, respetivement, de variables catégorielles et mixtes.

4.5.3 Estimation de l'incertitude dans les modèles EM

L'approche précédente est fréquemment utilisée pour estimer l'erreur quadratique moyenne du paramètre θ dans les modèles d'imputation EM. Cependant, dans ce cas particulier, une alternative, moins coûteuse en temps de calcul, est proposée dans [147] sous le nom de SEM [23].

Le principe de la méthode consiste à exprimer l'erreur quadratique moyenne de θ en fonction de deux quantités facilement estimable : l'erreur quadratique moyenne de θ sur les

23. *Supplemental EM*, en anglais

données observées et le taux de convergence de l'algorithme EM (qui est la différentielle de la fonction d'évolution de l'estimation du paramètre au cours de l'algorithme EM). Cette remarque permet d'obtenir directement l'erreur quadratique moyenne de θ au cours de l'algorithme EM.

4.5.4 Discussion

L'évaluation de l'incertitude liée à l'imputation est une phase importante pour évaluer la fiabilité des résultats d'une étude. Cette incertitude a diverses composantes, comme le soulignent [201] : l'erreur standard du paramètre estimé ou de la valeur imputée est liée, d'une part, à l'incertitude existant sur les données observées et, d'autre part, à la part d'incertitude provenant de l'imputation elle-même. Dans la plupart des cas, ces deux composantes sont confondues et l'erreur globale est estimée.

Dans les approches EM, l'imputation est prise en charge par une hypothèse paramétrique nécessitant l'estimation d'un paramètre θ. L'incertitude liée à l'imputation est donc directement liée à la valeur de ce paramètre et à son erreur standard. Toutefois, cette dernière n'est obtenue directement que dans la méthode FIML et les autres approches ML requièrent l'insertion d'une étape supplémentaire dans la méthode (SEM ou bien approches par ré-échantillonnage) pour fournir une estimation de l'erreur standard sur l'estimation de θ. Toutefois, ces approches nécessitent d'avoir une taille d'échantillon assez élevée : dans le cas contraire, il est fréquent d'avoir recours à une approche bayésienne.

Enfin, la principale limite des approches EM est qu'elles nécessitent des hypothèses paramétriques et l'adaptation de l'approche pour chaque cadre d'hypothèses. Aussi, l'imputation multiple constitue-t-elle un cadre plus simple pour l'estimation de l'incertitude liée à l'imputation. Dans le cadre standard de l'estimation d'une quantité numérique, la combinaison des différents résultats se fait de manière naturelle par un simple calcul de moyenne même s'il peut être plus compliqué de trouver des règles de combinaison des résultats satisfaisants les propriétés préconisées dans [133] pour des analyses plus complexes. Toutefois, dans le cadre de l'inférence statistique, la supériorité, en terme de puissance statistique, de l'approche EM (en particulier FIML) sur l'imputation multiple est fréquemment soulignée ([55], [189], [93] et [69]).

4.6 Prendre en compte les données manquantes informatives (MNAR)

La plupart des approches présentées dans cette revue et implémentées dans les packages R sont fondées sur l'hypothèse implicite que les données sont manquantes de type MAR. En pratique, cette hypothèse est souvent abusive, particulièrement dans le cas de sondages portant sur des questions sensibles ou d'études cliniques longitudinales (dans lesquelles des patients peuvent sortir de l'étude pour des raisons liées aux variables d'intérêts mesurées : cette question est donc liée à la thématique des données censurées).

Lorsque les données sont manquantes de type MNAR, la loi de Y_{miss} n'est pas indépendante de la loi de R. Dans ce cas, les approches habituelles de traitement des

données manquantes (qui consistent à estimer la loi multivariée $f(Y;\theta)$ à partir des données observées puis à utiliser cette loi pour l'inférence ou l'imputation) produisent des estimateurs ou des valeurs imputées biaisés.

Dans ce cas, l'estimation de la distribution jointe des données et de la probabilité d'absence, $f(Y,R;\theta,\psi)$ (ou $f(X,Y,R;\theta,\psi)$ si des covariables complètement observées sont disponibles), est la clé pour aborder cette question. Une approche courante consiste à proposer une factorisation réaliste de cette loi jointe qui soit estimable à partir des observations [131]. On distingue, en particulier, deux approches principales : les modèles de sélection [24] ([97] et [66], section 4.6.1) et les modèles par mélange de profils [25] [180] (section 4.6.2). Une troisième approche consiste à estimer les dépendances entre Y et R au moyen de variables latentes aléatoires : ce sont les modèles à paramètres partagés [26] ([131] et [100], section 4.6.3).

4.6.1 Modèles de sélection

Dans l'approche par modèle de sélection, la factorisation suivante de la loi jointe est utilisée :

$$f(Y,R;\theta,\psi) = f(Y|\theta)f(R|Y;\psi).$$

Cette factorisation est intuitive car elle modélise directement la distribution d'intérêt en utilisant la probabilité d'absence d'une donnée conditionnellement aux variables d'intérêt Y.

Un exemple typique est le modèle de [96], dans lequel les valeurs d'une variable Y_j sont expliquées par

$$Y_j = X^\top\theta + \epsilon, \tag{4.9}$$

où les erreurs ϵ sont indépendantes de X et suivent une loi gaussienne centrée de variance σ^2. La probabilité d'absence d'une valeur, R, conditionnellement à (X,Y_j) est, dans une première étape, estimée à l'aide (par exemple) d'un modèle PROBIT puis l'espérance conditionnelle $\mathbb{E}(Y|R=1)$, obtenue à partir de cette estimation, est utilisée comme variable explicative supplémentaire dans le modèle de régression de l'équation (4.9).

Des variantes de cette approche existent qui rentrent dans le cadre du modèle de sélection : par exemple, la méthode décrite dans l'article de [66] est une extension du modèle de Heckman au cas multivarié et [173] et [179] proposent des versions semi-paramétriques de ces approches pour la distribution des données complètes $f(Y;\theta)$ et les appliquent pour l'analyse des résultats d'un sondage sur le SIDA.

Une limite de ces approches est qu'elles sont souvent fondées sur des hypothèses paramétriques assez fortes, en particulier sur la spécification du modèle permettant d'obtenir $f(R|Y;\psi)$.

24. *Selection model* en anglais.
25. *Pattern mixture model* en anglais.
26. *Shared-parameter model* en anglais.

4.6.2 Modèles de mélange de profils

Comme les modèles de sélection, les modèles de mélange de profils utilisent une factorisation de la loi jointe $f(Y, R; \theta, \psi)$ pour estimer celle-ci. Dans ce cas-ci, la factorisation utilisée est

$$f(Y, R; \theta, \psi) = f(Y|R; \theta) f(R; \psi).$$

De manière concrète, la distribution conditionnelle décrit des profils distincts d'individus partageant le même profil de valeurs manquantes. Des sous-groupes d'individus, contenant les mêmes variables manquantes et observées, sont donc créés dans une première étape et dans chaque sous-groupe, la distribution, $f(Y|R; \theta)$, est estimée.

Les modèles de mélange de profils sont, par construction, sous-identifiés car, par définition des profils, certaines variables de $f(Y|R; \theta)$ sont toujours manquantes. [132] propose, pour résoudre ce problème, d'utiliser des restrictions identificatrices, c'est-à-dire des contraintes sur les paramètres inestimables de $f(Y|R; \theta)$ pour les profils incomplets. Différentes restrictions sont proposées, comme par exemple :

— valeurs manquantes des cas complets (CCMV[27]) [132] : le paramètre θ de $f(Y|R; \theta)$ est estimé pour le profil des cas complets et supposé identique pour tous les autres profils ;

— valeurs manquantes des cas disponibles (ACMV[28]) [149] : cette approche étend le cas précédent en estimant tous les paramètres estimables de θ directement dans chacun des profils et fixe les autres paramètres non estimables en utilisant un ordonnancement naturel (par exemple, dans le cas de données longitudinales) sur les différents profils.

Dans le cadre d'applications à l'analyse de données de qualité de vie chez des patientes atteintes du cancer du sein (qui sont des données censurées), [210] proposent une alternative aux restrictions identificatrices via des simplifications de modèle qui consistent à diminuer le nombre de paramètres à estimer. Ce principe est illustré par la description d'une stratégie d'estimation hiérarchique des lois dans les profils $f(Y|R; \theta)$ qui s'appuie sur la structuration longitudinale des variables.

4.6.3 Modèles à paramètres partagés

Dans les modèles à paramètres partagés, des variables aléatoires additionnelles, B, non observées, sont introduites pour modéliser la dépendance entre Y et R, qui sont alors supposées indépendantes sachant B. Dans ce cas, on a alors

$$f(Y, R|B; \theta, \psi) = f(Y|B; \theta) f(R|B; \psi)$$

et, par conséquent,

$$f(Y, R; \theta, \psi) = \int f(Y|B = b; \theta) f(R|B = b; \psi) f(b) db.$$

La stratégie standard consiste à faire une hypothèse paramétrique sur la distribution des effets aléatoires B. Un des premiers modèles à effets partagés a été proposé par [236] qui ont introduit cette approche dans le cadre de données longitudinales gaussiennes. $f(Y|B = b, \theta)$

27. *Complete Case Missing Value*, en anglais.
28. *Available Case Missing Value*, en anglais.

est modélisé comme un modèle linéaire avec effet aléatoire qui est combiné à $f(R|B=b,\psi)$, modèle PROBIT ou logistique à effet aléatoire.

[131] explique que les modèles à paramètres partagés peuvent être considérés comme des modèles de sélection à coefficients aléatoires [29] via la factorisation suivante :

$$f(Y,R,B;\theta,\psi) = f(Y|B;\theta)f(R|Y,B;\psi)f(B)$$

et comme des modèles de mélange de profils à coefficients aléatoires [30], via la factorisation suivante :

$$f(Y,R,B;\theta,\psi) = f(Y|R,B;\theta)f(R|B;\psi)f(B).$$

Des extensions de cette approche sont proposées dans [82] qui développent un modèle pour des réponses binaires dans le cadre d'une étude longitudinale et dans [4] qui étendent l'approche initiale à l'analyse de données de comptage longitudinales. [86] proposent également l'extension de l'algorithme EM stochastique pour estimer les paramètres du modèle à paramètres partagés. Ils y ajoutent une étape supplémentaire pour obtenir une erreur standard sur cette estimation.

4.6.4 Limites de ces approches

Le modèle de sélection est fondé sur des hypothèses paramétriques sur $f(R|Y;\psi)$. Cette particularité le rend sensible à une mauvaise spécification de cette loi. Bien que ne reposant pas sur des hypothèses explicites de paramétrage d'une distribution, les modèles de mélanges de profils sont aussi très sensibles aux hypothèses de restriction, qui ne sont pas vérifiables. Par ailleurs, un compromis est à effectuer pour déterminer un nombre de profils de données manquantes adéquat : en effet, un grand nombre de profils améliore la précision du modèle mais en augmentant le nombre de paramètres à estimer et donc en détériorant la qualité de l'estimation de chacun de ces paramètres. Enfin, dans cette approche, la loi marginale de Y n'est pas disponible directement (les paramètres de cette loi sont estimés conditionnellement à un profil donné). Estimer cette loi nécessite donc une marginalisation par rapport aux profils de données manquantes :

$$f(Y;\theta) = \sum_R f(Y|R;\theta_R)f(R;\psi).$$

Ces deux types d'approches sont plus adaptés au cas où la non réponse est directement liée aux variables observées (comme dans l'exemple d'un questionnaire portant sur des réponses sensibles). Par contre, lorsque l'absence d'une donnée est attribuable à un processus sous-jacent, par exemple la progression d'une maladie, il est préférable d'utiliser un modèle à paramètres partagés qui pourra prendre en compte ce processus à l'aide des effets aléatoires B. C'est le cas, par exemple, en présence de données censurées [131].

4.6.5 Analyse de sensibilité

Les approches décrites précédemment sont fondées sur des hypothèses invérifiables sur le lien entre le processus de données manquantes et le processus d'intérêt. [220] et

29. *Random-coefficient selection model*, en anglais.
30. *Random-coefficient pattern-mixture model*, en anglais.

[210] proposent une approche par analyse de sensibilité fondée sur une perturbation des données en direction de l'hypothèse MNAR pour vérifier la pertinence du modèle MAR. L'idée principale est de comparer les résultats obtenus sous ces deux hypothèses pour analyser la sensibilité des résultats à l'hypothèse MNAR.

Il existe différentes manières d'effectuer une analyse de sensibilité en présence de données manquantes. Une analyse de sensibilité relativement simple consiste à étudier les résultats de différents jeux de données imputés issus de modèles d'imputation différents. Ce principe est proposé dans le package **mice** qui met en place un certain nombre de scénarios plausibles et permet d'examiner les conséquences de chacun d'entre eux sur l'inférence finale. Dans le cas où l'hypothèse MAR semble violée, les auteurs proposent de multiplier les imputations par un facteur ou de leur ajouter une valeur fixe, les deux approches étant des formes basiques de modèles à mélange de profils.

Certaines méthodes utilisées pour imputer les données MNAR peuvent également être employées pour effectuer une analyse de sensibilité. [220] proposent ainsi d'utiliser les modèles à mélange de profils pour l'analyse de sensibilité. [210] utilisent cette approche en comparant les résultats obtenus avec chacune des restrictions identificatrices possibles : cet ensemble de conclusions fournit ainsi un aperçu de la sensibilité aux hypothèses émises. Ce type d'approches peut donc s'avérer une première étape très utile pour détecter des évidences en faveur de l'hypothèse MNAR et trouver la stratégie qui semble la plus adéquate à leur prise en compte.

Enfin, notons que, si quelques approches et modèles permettent d'identifier et de prendre en compte les valeurs manquantes MNAR, une limite forte de celles-ci est l'absence d'implémentations dans les outils habituels de traitement des données manquantes. À notre connaissance, par exemple, aucun package R ne propose d'implémentation des modèles décrits plus hauts ni des approches d'analyse de sensibilité qui permettent de les évaluer.

4.7 Conclusion

Les données manquantes sont un problème fréquemment rencontré dans les analyses statistiques, quel que soit le domaine d'étude. La méthode la plus adéquate pour en tenir compte dépend de paramètres multiples comme la typologie des valeurs manquantes, le type de mécanisme qui a conduit à leur génération, leur distribution dans le jeu de données ainsi que les attentes de l'utilisateur en terme d'analyses statistiques. On peut toutefois dégager des recommandations générales en plusieurs étapes :

— la première étape consiste à décrire les données manquantes afin d'émettre des hypothèses sur le mécanisme des données manquantes. Ces hypothèses doivent guider le choix de la stratégie à utiliser pour les traiter, conduire à supprimer des données (individus ou variables) ou bien à compléter simplement certaines valeurs manquantes dont on a identifié l'origine [77] ;

— lorsque le but de l'analyse statistique est l'inférence et que les données manquantes sont supposées MAR, les approches EM et bayésienne fournissent des estimations non biaisées pour lesquelles il est possible d'obtenir une bonne estimation des erreurs standards.

Dans d'autres cas d'analyses statistiques, les approches d'imputation multiple, qui

permettent d'estimer la variabilité liée à l'imputation tout en fournissant un ou des tableaux de données complets, sont recommandées. Selon les hypothèses sur la distribution multivariée des données et selon le type d'analyse à effectuer *a posteriori*, ces imputations multiples pourront être basées sur des approches hot-deck, des approches par prédiction, des approches factorielles ou des approches bayésiennes. Confronter et comparer différents types d'imputation, notamment par analyse de sensibilité, peut permettre d'identifier les limites liées à chaque approche sur un cas d'application donné.

En revanche, si les données sont MNAR, ce qui est particulièrement fréquent dans le cas des études longitudinales, l'imputation doit alors être fondée sur des modèles spécifiques à ce type de données ;

— la dernière étape consiste à essayer d'obtenir une évaluation de la qualité de l'imputation ou de l'estimation statistique, soit en utilisant des outils ou des caractéristiques numériques diagnostiques, soit en procédant par analyse de sensibilité. En particulier, les hypothèses MAR/MNAR étant impossibles à vérifier par définition, il semble judicieux de systématiquement effectuer une analyse de sensibilité des résultats d'imputations sous hypothèses MAR/MNAR en cas de doute (lorsque la distribution des valeurs manquantes n'est pas homogène, par exemple). Toutefois, ces approches ne sont pas, à notre connaissance, implémentées dans les packages R actuellement disponibles.

Cette revue fournit un panorama des grandes familles de méthodes pouvant prendre en compte les données manquantes lors d'analyses statistiques. Nous nous sommes attachées à décrire des solutions logicielles disponibles pour utiliser ces méthodes, en listant les divers packages R dans lesquels elles sont implémentées. Des tableaux récapitulant les différentes méthodes et les packages R associés sont fournis après cette conclusion, organisés de la même manière que les sections de cet article (analyse descriptive, utilisation des données observées, inférence, imputation simple, variabilité liée à l'imputation). La liste des packages ne prétend pas à l'exhaustivité mais propose un panorama réaliste des packages utilisables pour mettre en œuvre une approche donnée.

Méthodes	**Packages R**	**Cadre d'application**
Identification de motifs de données manquantes	**mi** [205]	tableaux de données mixtes
Description des données manquantes	**naniar** ; **VIM** ([208] et [123])	tableaux de données mixtes
Test MAR/MCAR	**BaylorEdPsych** ; **missMech** [110]	numériques et catégorielles

Tableau 4.1 Packages permettant l'analyse descriptive des données manquantes

Méthodes	Packages R	Cadre d'application
Analyse des cas complets	option disponible dans de nombreuses fonctions : `na.action=na.omit`	numériques et catégorielles
Analyses des cas disponibles	**regtools** ; option disponible dans certaines fonctions (par exemple, `method="pairwise"` dans la fonction `cor`)	numériques et catégorielles
Pondération par probabilité inverse (IPW)	**ipw** [227]	numériques et catégorielles

Tableau 4.2 Récapitulatif des méthodes fondées uniquement sur les données observées

Méthodes	Packages R	Cadre d'application
FIML	**lavaan** [178]	modèle à équations structurelles
Approche EM avec un modèle multivarié normal	**norm** [190]	données multivariées gaussiennes
Approche EM avec un modèle log-linéaire	**cat** [190]	données multivariées catégorielles
Équivalent du package **norm** pour des données mixtes	**mix** [190]	données multivariées mixtes
EM avec approche bayésienne ou bootstrap	**Amelia** [102]	variables numériques

Tableau 4.3 Packages implémentant les approches paramétriques d'inférence statistique (EM ou bayésiennes)

Méthodes	Packages R	Cadre d'application
Moyenne, médiane	**ForImp** ; **Hmisc** ; **simputation**	variables numériques
Mode	**ForImp** ; **Hmisc**	variables catégorielles
LOCF	**zoo**	données longitudinales
k-plus proches voisins	**DMwR** [214] ; **impute** [215] ; **VIM** ([208] et [123]) ; **yaImpute** [60]	variables numériques et/ou catégorielles, selon la distance choisie
Hot-deck	**hot.deck** [58] ; **HotDeckImputation** ; **simputation** ; **VIM** ([208] et [123])	tableaux de données mixtes
Régression	**simputation** ; **snpStats** (Bioconductor) ; **VIM** ([208] et [123])	variables numériques pour **simputation** et **VIM** ; données SNP pour **snpStats**
Régression LOESS	**locfit**	variables numériques
Régression stochastique	**mice** ($m = 1$) [39]	variables numériques
Arbres et forêts aléatoires	**missForest** [202]	tableaux de données mixtes
NIPALS	**ade4** [46] ; **pcaMethods** (Bioconductor, [200]) ; **mixOmics** [129]	variables numériques
Analyses factorielles	**missMDA** [116]	variables catégorielles et/ou numériques, selon la méthode choisie
Procédure d'imputation « en avant »	**ForImp** [79]	variables ordinales
Interpolation, ajustement d'une courbe de lissage, estimation de régression longitudinales	**forecast** ; **imputeTS** [153] ; **spacetime** [160] ; **timeSeries** ; **xts** ; **zoo** [241]	séries temporelles

Tableau 4.4 Packages contenant des approches d'imputation simple

Méthodes	Packages R	Cadre d'application
	Outils de diagnostic	
Calcul d'erreurs	**Amelia** [102] ; **missMDA** [116] ; **yaImpute** [60]	tableaux de données mixtes
Graphiques	**mi** [205] ; **VIM** ([208] et [123])	tableaux de données mixtes
	Imputation multiple	
Équations chaînées	**mi** [205] ; **mice** [39]	tableaux de données mixtes
Hot-deck	**hot.deck** [58]	tableaux de données mixtes
Analyses factorielles (MIPCA, MIMCA)	**missMDA** [116]	tableaux de données mixtes
Approche de modélisation jointe (EM et bayésienne)	**Amelia** [102]	variables numériques
Approche de modélisation jointe (bayésienne)	**MixedDataImpute** ; **NPBayesInput**	variables catégorielles et mixtes, respetivement
Approches de modélisation jointe (bayésiennes) multi-niveaux	**jomo** ; **pan**	tableaux de données mixtes
Combinaison générique	**mitools**	tableaux de données mixtes stratifiés en classes

Tableau 4.5 Packages incluant des approches d'évaluation de la variabilité due en présence de données manquantes ou due à l'imputation

Données manquantes et inférence de réseau

Chapitre 5

Imputation multiple hot-deck pour l'inférence de réseau à partir de données RNA-Seq

Ce chapitre est la traduction d'un article publié dans Bioinformatics [107].

5.1 Introduction

Ces dernières années, les domaines de la biologie et de la médecine ont connu une avancée importante en ce qui concerne les techniques de séquençage, permettant d'accéder à une large quantité de données 'omiques, à différents niveaux de l'échelle du vivant. Parmi les techniques de séquençage à haut débit, la technologie RNA-Seq permet de mesurer simultanément l'expression de plusieurs milliers de gènes, même inconnus, pour un tissu donné. Ces larges quantités de données générées ont créé un besoin dans le traitement et l'analyse, au niveau bioinformatique et statistique, pour ce type de données expérimentales. Une attention particulière a été portée à la recherche des différents types de relations (co-expression ou régulation) entre les gènes [242, 152]. Mieux comprendre ces relations permet de donner un aperçu du fonctionnement global de la cellule (ou du tissu) dans un environnement donné. Il s'agit d'un point important pour mettre en évidence des voies de signalisation et pour identifier des gènes cibles pour un problème biologique donné. Afin de faciliter l'analyse globale de ce type de données, il peut être intéressant de visualiser ces relations sous la forme d'un réseau de gènes.

Les modèles habituellement utilisés pour inférer des réseaux à partir de données d'expression de gènes sont les modèles graphiques gaussiens (GGM). Ces modèles supposent que les données suivent des distributions normales. Cependant, les données d'expression RNA-Seq sont des données de comptage et sont, par conséquent, de nature discrète. Les modèles GGM ne s'avèrent donc pas adaptés pour ce type de données d'expression. Des travaux récents proposent des modèles graphiques prenant en compte la particularité des données d'expression de type RNA-Seq, en se basant sur des modèles utilisant des distributions de Poisson. Parmi ces modèles, nous avons des modèles linéaires généralisés basés sur des distributions de Poisson (modèle log-linéaire de Poisson proposé par [6]) ou des modèles hiérarchiques log-normal de Poisson [87]. Comme pour les modèles GGM, une pénalisation L_1 est ajoutée au modèle pour sélectionner les arêtes et obtenir un réseau parcimonieux. L'inférence de réseau reste néanmoins un problème difficile [221] puisque le nombre d'échantillons disponibles (n) est généralement plus faible que le nombre de paramètres à estimer (de l'ordre de p^2 où p correspond au nombre de gènes). En outre, l'inférence de réseau s'avère être sensible aux individus manquants dans les gènes clés [163]

ou à la présence d'individus « influents » [19]. Avoir un nombre élevé d'échantillons reste un moyen sûr pour obtenir des résultats fiables dans l'analyse statistique de données RNA-Seq [136].

Nous proposons ici une méthode pour augmenter la robustesse des réseaux de gènes lorsque des données d'expression RNA-Seq ont été mesurées avec d'autres données biologiques choisies avec précaution. Ces données doivent être choisies de telle sorte que des individus similaires pour ces données biologiques doivent également être similaires pour les données RNA-Seq. Divers cas de figures sont possibles : le premier cas typique est une étude dans laquelle des données d'expression de gènes ont été mesurées via diverses technologies. En effet, il est fréquent de mesurer, simultanément aux données RNA-Seq, d'autres types de données d'expression de gènes. D'un coût généralement moins élevé, ces données sont disponibles pour un nombre plus important d'échantillons. Un autre cas d'étude est lorsque les données d'expression RNA-Seq sont mesurées sur deux tissus différents. Il est possible qu'un des deux tissus soit plus difficile à récupérer et qu'il y ait par conséquent moins d'échantillons pour un des deux tissus. Dans ces deux cas, le nombre d'échantillons peut facilement être plus assez important et permet de fournir de l'information supplémentaire sur la relation entre les individus et la variabilité biologique. Il serait logique d'utiliser cette information pour améliorer la qualité de l'inférence de réseau construit à partir des données RNA-Seq.

Nous avons conçu une approche basée sur l'imputation dans laquelle des individus non observés pour les données RNA-Seq (mais observés dans un autre jeu de données) sont considérés comme manquants. Cette nouvelle méthode d'imputation, l'imputation multiple hot-deck (**hd-MI**), est présentée dans la section 5.2. Le choix des hyperparamètres est discuté dans la section 5.3. L'approche **hd-Mi** est évaluée sur deux jeux de données réelles RNA-Seq : un provenant d'une étude sur l'expression des gènes dans divers tissus humains et un provenant de l'étude longitudinale DiOGenes mesurant l'expression des gènes du tissu adipeux. Les jeux de données et la méthodologie utilisée pour l'évaluation sont présentés dans la section 5.4. La section 5.5 présente les résultats.

5.2 Présentation de la méthode

5.2.1 Notations

Dans la suite de ce chapitre, $\mathbf{X}$ correspond au jeu de données RNA-Seq avec n_1 lignes (n_1 individus) et p colonnes (p gènes). Le comptage du gène j, $j \in \{1, \ldots, p\}$, pour l'individu i, $i \in \{1, \ldots, n_1\}$ est noté x_{ij}. Des données auxiliaires ont également été obtenues sur ces n_1 individus et sur d'autres individus. On note $\mathbf{Y}$ la matrice de dimension $n \times q$ avec $n > n_1$ contenant ces données. y_{ij} correspond à l'observation de la variable j, $j \in \{1, \ldots, q\}$, pour l'individu i, $i \in \{1, \ldots, n\}$. Sans perte de généralité, les individus communs entre $\mathbf{X}$ et $\mathbf{Y}$ sont supposés correspondre aux n_1 premières lignes de $\mathbf{Y}$. Ce problème peut alors être vu

comme un problème de données manquantes dans la matrice $\left[\tilde{\mathbf{X}}, \mathbf{Y}\right]$ de dimension $n \times (p+q)$ dans laquelle, $\tilde{x}_i$ est défini par :

$$\begin{cases} x_i & \forall i = 1, \ldots, n_1 \\ \tilde{x}_i \text{ est manquant} & \forall i \geq n_1 + 1 \end{cases}.$$

Une telle structure de données manquantes est appelée non-réponse totale (*unit non-response*, en anglais) puisque les valeurs manquantes correspondent à l'absence totale d'un individu (autrement dit aucune variable n'est observée pour cet individu). La figure 5.1 permet de représenter schématiquement cette structure de données manquantes.

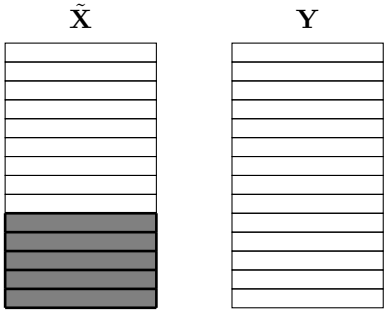


Figure 5.1 Schéma des données manquantes dans le jeu de données d'expression RNA-Seq ($\tilde{\mathbf{X}}$) et dans le jeu auxiliaire ($\mathbf{Y}$).

Les individus manquants $i \in \{n_1 + 1, \ldots, n\}$ du jeu de données RNA-Seq sont supposés être MCAR. Il s'agit d'une hypothèse standard si les individus n'ont pas été choisis selon une caractéristique spécifique parmi $\{1, \ldots, n\}$ mais qu'ils ont été sélectionnés aléatoirement ou à cause de contraintes techniques comme une expérience ratée, un manque de tissu ou encore à cause de contraintes financières.

5.2.2 Imputation multiple hot-deck (hd-MI)

Une grande variété de méthodes permet d'imputer des valeurs manquantes [73, 133]. Néanmoins, la plupart de ces méthodes impute les valeurs manquantes indépendamment les unes des autres. Dans notre cas, nous devons faire face à deux problématiques : tout d'abord, ce sont des individus en entier (et non quelques variables) qui sont considérés comme manquants. Deuxièmement, dans le contexte de l'inférence de réseau, il est important de préserver, lors de l'imputation, la structure de corrélation entre les variables. Or, les méthodes usuelles d'imputation ne remplissent pas ce critère.

L'imputation hot-deck est souvent utilisée pour imputer des problèmes de non-réponse totale dans les sondages [11]. Cette méthode est basée sur le concept de donneurs. Pour chaque individu, i, appelé « récipient », avec une valeur manquante $\tilde{x}_{ij}$, un groupe d'individus similaires (appelés « donneurs ») est créé à partir des individus pour lesquels cette variable $\tilde{x}_{i'j}$ est observée : $\{i' : i' \neq i$ tel que $\tilde{x}_{i'j}$ n'est pas manquant$\}$. Cet ensemble de donneurs dépend de l'individu i lui-même. Il est appelé groupe de donneurs et est noté $\mathcal{D}(i)$. Un des donneurs est finalement choisi aléatoirement parmi les individus appartenant à $\mathcal{D}(i)$. La valeur de $\tilde{x}_{i'j}$ est utilisée pour imputer $\tilde{x}_{ij}$. L'imputation hot-deck permet généralement de préserver la distribution des variables et ne sous-estime pas la variance [73]. Les valeurs imputées étant des valeurs observées, elles s'avèrent donc être réalistes et respectent

les spécificités et caractéristiques des variables (par exemple, la positivité ou le caractère discret).

Cependant, dans une utilisation basique de l'imputation hot-deck, la structure de corrélation entre les variables est modifiée durant l'imputation puisque les différentes variables manquantes pour un individu i sont imputées indépendamment les unes des autres. Pour pallier ce problème dans le cas de non-réponse totale, [226] ont proposé d'imputer simultanément toutes les variables $(\tilde{x}_{ij})_{j=1,\dots,p}$ par les valeurs provenant d'un seul et même donneur $i' \in \mathcal{D}(i)$ dans le cadre d'un problème d'intégration de données 'omiques.

Notre approche **hd-MI**, schématisée par la figure 5.2, est donc relativement proche des travaux de [226]. Nous l'adaptons pour un problème d'inférence de réseau avec un jeu de données auxiliaire. Ainsi, une approche de type « hot-deck » est utilisée pour imputer des lignes entières de $\mathbf{X}$ en utilisant de l'information de proximité entre individus mesurée avec les données $\mathbf{Y}$. Cette méthode a l'avantage de respecter les caractéristiques initiales des données (caractère discret et positivité) et de conserver la structure de corrélation entre les variables imputées. Ce dernier point est primordial pour l'inférence de réseau. La méthode est mise en œuvre dans un cadre d'imputation multiple permettant d'observer la stabilité des arêtes inférées.

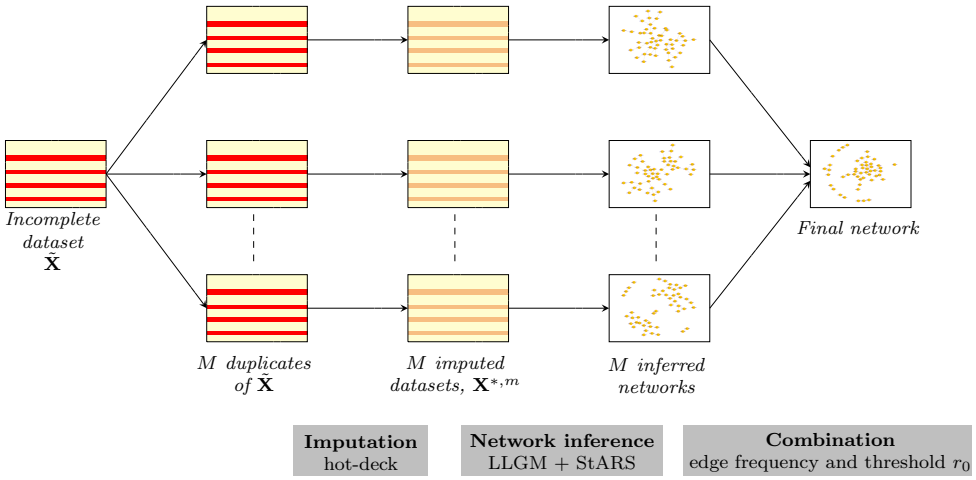


Figure 5.2 **Aperçu de la méthode hd-MI.** Les données initiales $\tilde{X}$ (première colonne) sont dupliquées M fois (seconde colonne). Pour chaque jeu dupliqué, chaque ligne manquante est imputée via l'approche hot-deck (troisième colonne, $X^{*,m}$). Un réseau est inféré pour chaque jeu imputé (quatrième colonne) avec la méthode LLGM (le critère StARS est utilisé pour choisir le paramètre de régularisation, ρ). Pour finir, les réseaux sont combinés en un seul en utilisant le seuil r_0 pour sélectionner les arêtes les plus fréquentes parmi les M réseaux obtenus (cinquième colonne).

Pour résumer, une méthode d'imputation multiple hot-deck est mis en œuvre :

1. Dans un premier temps, pour tous les individus manquants dans $\tilde{\mathbf{X}}, i = n_1 + 1, \dots, n$, le groupe de donneurs $\mathcal{D}(i)$ est créé et contient tous les individus $i' \leq n_1$ qui sont similaires à l'individu i. Pour estimer cette similarité entre les individus, les données auxiliaires $\mathbf{Y}$ sont utilisées. Différentes similarités peuvent être calculées entre les individus sur ce jeu de données $\mathbf{Y}$. Parmi elles, nous proposons d'utiliser un score

d'affinité, proposé dans [58]. Ce score d'affinité est calculé pour tous les individus i' de la façon suivante :

$$s(i,i') = \frac{1}{q}\sum_{j=1}^{q} \mathbb{I}_{\{|y_{ij}-y_{i'j}|<\sigma\}}$$

dans laquelle σ est un seuil fixé. Le groupe de donneurs est alors défini par $\mathcal{D}(i) = \{i' : s(i,i') = \max_{l=1,\ldots,n_1} s(i,l)\}$. Le score correspond au nombre moyen de variables observées pour lesquelles les individus i et i' sont « proches » ;

2. Dans un second temps, un individu i' est choisi aléatoirement dans le groupe de donneurs $\mathcal{D}(i)$. La ligne entière i de $\tilde{\mathbf{X}}$ est imputée par la ligne i' de $\tilde{\mathbf{X}}$. Cette étape est répétée pour tout $i = n_1 + 1, \ldots, n$ pour produire un jeu de donnée complet, noté $\mathbf{X}^*$.

Dans le contexte de l'imputation multiple, cette procédure est répétée M fois afin de produire M jeux de données complets $\mathbf{X}^{*,m}$. La seconde étape de l'analyse consiste alors à inférer un réseau pour chacun de ces jeux de données complets en utilisant le modèle log-linéaire de Poisson (LLGM), proposé par [6]. Les M réseaux sont finalement combinés en un seul réseau. Pour cela, nous étudions le nombre de fois où une arête est prédite parmi ces M réseaux :

$$r(e) = \frac{\text{nombre de fois où l'arête } e \text{ est prédite}}{M}.$$

Un seuil de fiabilité, r_0, est finalement choisi et le réseau final est composé des arêtes e tel que $r(e) \geq r_0$. Cette approche est similaire au critère de stabilité décrit par [143]. L'incertitude de l'imputation est traitée de la même façon que les approches standards pour améliorer la qualité de l'inférence de réseau [8, 17]. Ces travaux utilisent pour cela des poids moyens ou des rangs moyens entre plusieurs réseaux provenant de différents ré-échantillonnages ou d'expérimentations indépendantes.

Finalement, **hd-MI** ne requiert pas d'ajuster un nombre trop important d'hyperparamètres. Un paramètre, σ, est à définir pour le groupe de donneur $\mathcal{D}(i)$. Il convient également de choisir M, le nombre de fois où l'imputation hot-deck est appliquée. L'étape de combinaison requiert également de définir le seuil de fiabilité r_0. Un paramètre est également à choisir lors de l'inférence du réseau : le paramètre de régularisation ρ. Les choix de ces paramètres sont discutés dans la section suivante.

5.3 Choix des hyperparamètres

Le modèle d'imputation hot-deck nécessite de définir un seuil σ. Nous proposons une méthode afin de choisir une valeur de σ. L'objectif est d'obtenir un bon compromis entre l'homogénéité (soit une variabilité faible au sein d'un groupe de donneurs) et la variété (soit un nombre de donneurs assez important) des individus dans le groupe de donneurs. Nous proposons donc de calculer l'inertie moyenne intra-$\mathcal{D}(i)$ pour tout $i = n_1 + 1, \ldots, n$:

$$V_{\text{intra}}(\sigma) = \frac{\sum_i \frac{\sum_{i' \in} ||x_i - x_{i'}||^2}{d_i}}{n - n_1}$$

où d_i correspond au nombre de donneurs pour l'individu i. Lorsque la valeur de $V_{\text{intra}}(\sigma)$ est petite, l'homogénéité des individus au sein des groupes de donneur est importante.

En revanche, la valeur de $V_{\text{intra}}(\sigma)$ augmente avec la valeur de σ. De ce fait, nous proposons d'étudier l'évolution de V_{intra} en fonction des valeurs de σ et d'utiliser la « règle du coude » pour sélectionner une valeur pour le seuil σ.

Le nombre de jeux de données imputés, M, doit être assez important afin de limiter l'influence de certains individus spécifiques dans l'imputation et afin d'estimer la variabilité induite par l'imputation sur les résultats de l'analyse (dans notre cas, l'inférence de réseau). Dans notre cas, nous avons choisi la valeur $M = 100$ qui s'est montrée satisfaisante en pratique.

En ce qui concerne le paramètre de régularisation ρ, utilisé dans le modèle d'inférence de réseau, nous avons opté pour le choix du critère StARS [135].

Finalement, le seuil r_0 est choisi conformément à la distribution de $r(e)$ parmi toutes les paires de gènes, e. On s'attend à ce qu'une forte proportion des valeurs de $r(e)$ soit petite (c'est-à-dire que la plupart des paires de gènes sont dans l'ensemble d'arêtes prédites 1-5 fois parmi les M réseaux inférés) et à ce que seulement une petite fraction d'arêtes soit présente dans la plupart des M réseaux inférés. En pratique, de tels comportements ont été observés et des valeurs de r_0 comprises entre 90-100 $\%$ produisent des résultats intéressants.

5.4 Évaluation de la méthode

Pour évaluer les performances de la méthode **hd-MI**, des données réelles provenant de deux projets distincts ont été utilisées : *Genotype-Tissue Expression* (GTEx) [138] et DiOGenes [125].

L'évaluation, consistant à tester l'efficacité de notre méthode, se décompose en deux parties :

— une partie « évaluation » où nous comparons notre méthode avec d'autres méthodes d'imputation ;

— une partie « application » où nous testons notre méthode sur des données réelles afin de voir si les résultats obtenus sont cohérents avec les a priori biologiques.

Dans la partie « évaluation », nous avons voulu évaluer l'efficacité de notre méthode à améliorer la qualité de l'inférence de réseau en la comparant avec d'autres méthodes d'imputation. Pour cela, nous nous sommes restreints aux échantillons communs entre les données initiales $\mathbf{X}$ et les données auxiliaires $\mathbf{Y}$. De cette façon, les données ne comportent pas d'individus manquants. Nous pouvons alors supprimer artificiellement un pourcentage d'individus de $\mathbf{X}$.

Dans la partie « application », la méthode **hd-MI** est appliquée sur l'ensemble des individus disponibles pour le projet DiOGenes : les individus communs entre CID1 et CID2 ainsi que les individus présents à un seul pas de temps. Les réseaux obtenus avec **hd-MI** seront comparés à de précédents réseaux de gènes inférés sur d'autres types d'expression et/ou sur des sous-groupes d'individus différents, provenant du même projet.

5.4.1 Description des données

Description des données provenant de GTEx

GTEx est un projet qui a collecté et analysé différents tissus humains à partir de 544 donneurs. Les données d'expression, mesurées par RNA-Seq, ont été acquises sur plus de 53 tissus humains. Les données sont disponibles sur `https://gtexportal.org`. Dans le cadre de notre analyse, nous nous sommes restreints à deux tissus pour lesquels [145] ont mis en évidence des profils similaires d'expression de gènes. Les deux tissus sélectionnés sont le poumon et la thyroïde.

Les expressions de gènes obtenues avec le tissu du poumon sont utilisées comme données initiales $\mathbf{X}_0$ et celles obtenues avec le tissu de la thyroïde sont utilisées comme données auxiliaires $\mathbf{Y}$. Seuls les 221 individus communs entre les deux jeux de données sont gardés dans l'analyse. Les données RNA-Seq ont été normalisées avec la normalisation TMM (disponible dans le package **edgeR**). L'inférence de réseau a été réalisée en utilisant les gènes les plus variables ($p = 100$ gènes pour $\mathbf{X}_0$ et $q = 50$ pour $\mathbf{Y}$). Entre les deux jeux de données, 36 gènes sont en commun.

Description des données provenant de DiOGenes

DiOGenes est une étude d'intervention diététique contrôlée, réalisée dans huit pays européens. Nous nous intéressons ici à la première phase de l'étude : des individus obèses ont suivi un régime hypocalorique durant huit semaines avec pour objectif de perdre au moins 8% de leur poids initial. Parmi les nombreuses mesures biologiques, des mesures transcriptomiques ont été obtenues à partir de biopsies de tissu adipeux avant le régime (CID1) et à la fin de celui-ci (CID2). L'expression des gènes a été quantifiée via différentes techniques, comme le RNA-Seq et la RT-qPCR. Nous abordons ici le problème de l'imputation des données RNA-Seq afin de mieux comprendre l'impact d'une restriction calorique sur la relation entre les gènes.

Les données d'expression de gènes ont été mesurées par RNA-Seq sur 433 individus en CID1 et 307 en CID2 avec seulement 189 individus en commun entre les deux pas de temps. Pour l'inférence de réseaux, seulement 317 gènes ont été retenus dans l'analyse. Ces gènes ont été sélectionnés selon des critères (en fonction de l'obésité, du changement de poids, de l'adaptation métabolique ou des différents types de cellules adipeuses) définis par les biologistes ou la littérature [222, 22]. Les données RNA-Seq ont été normalisées via la méthode *upper-quartile* grâce au package **edgeR**.

Les données d'expression RT-qPCR (données auxiliaires $\mathbf{Y}$) ont été mesurées sur 413 individus en CID1 et 428 en CID2 pour 272 gènes. Ces 272 gènes sont inclus dans la liste des 317 gènes sélectionnés pour les données RNA-Seq. La figure 5.4 est un diagramme de Venn permettant de voir les individus communs entre les différents jeux de données. Les données RT-qPCR ont été normalisées en utilisant le gène de référence $2^{-\Delta C_t}$ [222, 137].

Pour la partie « évaluation », les données à CID1 et CID2 ont été étudiées indépendamment en utilisant seulement les 189 individus communs. Pour $\mathbf{Y}$ (RT-qPCR), les échantillons sont restreints aux échantillons en commun avec les 189 échantillons définis précédemment. Nous avons ainsi 166 échantillons RT-qPCR pour CID1 et 172 pour CID2.

Pour la partie « application », l'ensemble des individus en RNA-Seq est utilisé pour l'inférence de réseau. Les données auxiliaires contiennent 413 individus en CID1 et 428 en CID2. Les données auxiliaires pour CID1 sont composées de 313 individus également

observés et de 100 individus non observés en RNA-Seq à CID1. De même pour CID2, 276 individus sont observés dans les jeux RNA-Seq et RT-qPCR à CID2 et 152 sont seulement observés dans le jeu RT-qPCR.

La différence entre les parties « évaluation » et « application » (sur les données provenant de DiOGenes) est illustrée par la figure 5.3.

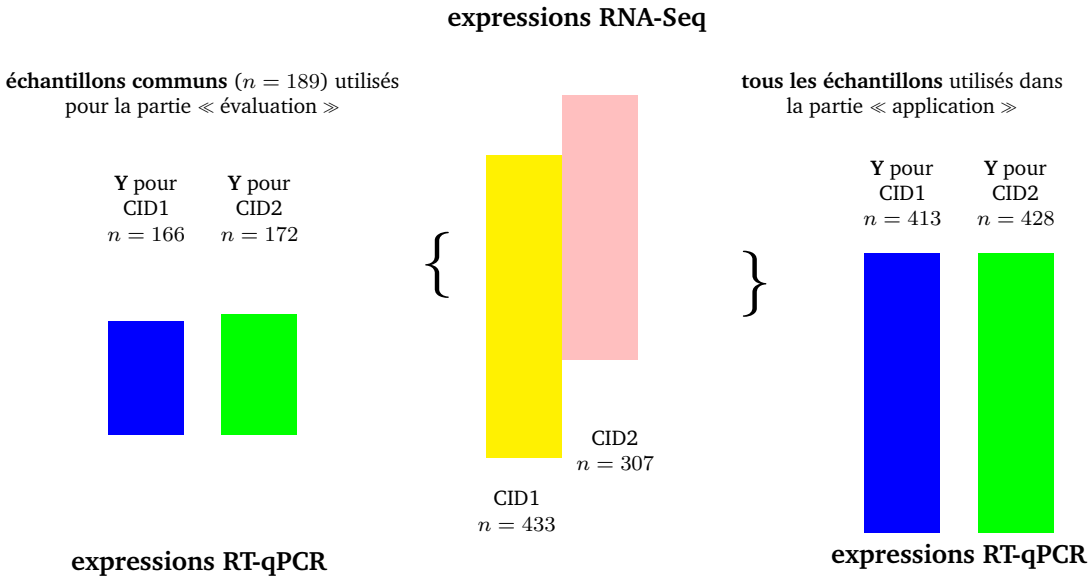


Figure 5.3 **Organigramme pour les jeux de données DiOGenes.** La partie gauche illustre la partie « évaluation » de la méthode basée sur ces données. La partie droite illustre la parte « application » de la méthode sur l'ensemble des données pour obtenir deux réseaux : un à CID1 et un à CID2. Pour plus d'information sur le nombre d'échantillons utilisés dans chaque cas, voir Figure 5.4.

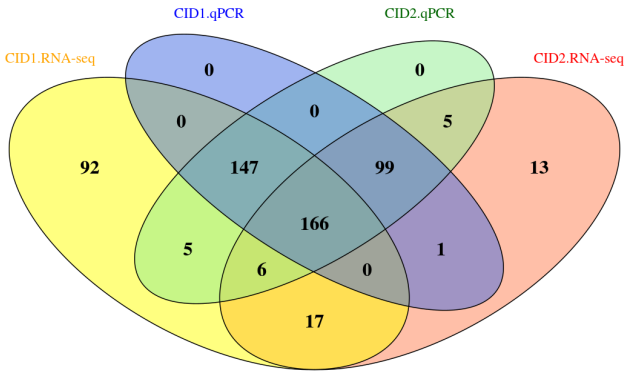


Figure 5.4 DiOGenes : diagramme de Venn (nombre d'échantillons communs) CID1 contre CID2 et RNA-Seq contre RT-qPCR.

5.4.2 Évaluation et comparaison avec d'autres méthodes

Les jeux de données complets, composés des échantillons communs avec les données auxiliaires, sont utilisés pour inférer les réseaux dits « de référence ». Ils ne contiennent pas d'individus manquants et sont notés $\mathbf{X_0}$.

Les données avec des individus manquants sont ensuite générées en supprimant aléatoirement un certain nombre d'individus dans $\mathbf{X_0}$. Plus précisément, à partir du jeu complet $\mathbf{X_0}$, un pourcentage donné, f, de lignes est supprimé aléatoirement (avec $f \in \{10, 20, 30 \text{ et } 40\%\}$) pour produire un jeu de données $\tilde{\mathbf{X}}$ avec des individus manquants. Le jeu de données des cas complets correspondant est noté $\mathbf{X}$. La méthode d'imputation **hd-MI** ainsi que deux autres méthodes sont alors utilisées pour imputer le jeu incomplet. Les deux autres approches sont une méthode d'imputation simple et naïve, l'imputation par la moyenne et une méthode d'imputation multiple par PCA (MIPCA) [117].

Nous inférons alors des réseaux à partir :

— du jeu de données initial $\mathbf{X_0}$. Les réseaux obtenus à partir des données sont référencés par le terme « **reference** » dans la suite de ce chapitre. Le réseau obtenu est utilisé comme référence absolue pour la comparaison ;

— des jeux de données avec un certain pourcentage d'individus manquants. Les réseaux sont alors inférés à partir des cas complets $\mathbf{X}$. Les réseaux sont référencés par le terme « **missing** » (potentiellement suivi du taux d'individus manquants) ;

— des jeux de données imputés par la moyenne. Les réseaux sont référencés par le terme « **mean** » (potentiellement suivi par le taux d'individus manquants) ;

— des M jeux de données imputés via l'approche MIPCA. Pour DiOGenes, les données mesurées par RT-qPCR (données auxiliaires $\mathbf{Y}$) sont réduites avant de calculer les similarités. Certaines valeurs imputées pour les données d'expression RNA-Seq sont négatives. Or, les valeurs étant des données de comptage (données positives), les valeurs négatives sont remplacées par 0. Les réseaux obtenus avec ces données sont référencés par le terme « **MIPCA** » (potentiellement suivi par le taux d'individus manquants) dans la suite ;

— des M jeux de données imputés avec notre approche **hd-MI**. Les réseaux sont référencés par le terme « **hd-MI** » (potentiellement suivi par le taux d'individus manquants).

Les réseaux sont inférés de la façon suivante :

— Pour les cas où une seule inférence est effectuée (les cas **reference**, **missing** et **mean**), le modèle LLGM est estimé pour chaque valeur ρ du chemin de régularisation. Pour chaque jeu de données, nous obtenons donc un réseau par valeur de ρ. Nous calculons le critère StARS afin d'obtenir la valeur ρ liée au réseau le plus stable.

— Pour les cas où l'imputation multiple est utilisée (**MIPCA** et **hd-MI**), un réseau est inféré pour chaque jeu $X^{*,m}$ en utilisant le paramètre de régularisation ρ choisi par le critère StARS. Nous obtenons alors M réseaux qui sont combinés en un seul en testant différentes valeurs pour le seuil de fiabilité r_0. Pour chaque seuil r_0 testé, un réseau final est obtenu.

Le processus d'évaluation est représenté par la figure 5.5. Les résultats, obtenus avec ces différentes méthodes, ont été évalués via des comparaisons globales de la structure du réseau et des comparaisons locales en utilisant le réseau **reference** comme référence absolue.

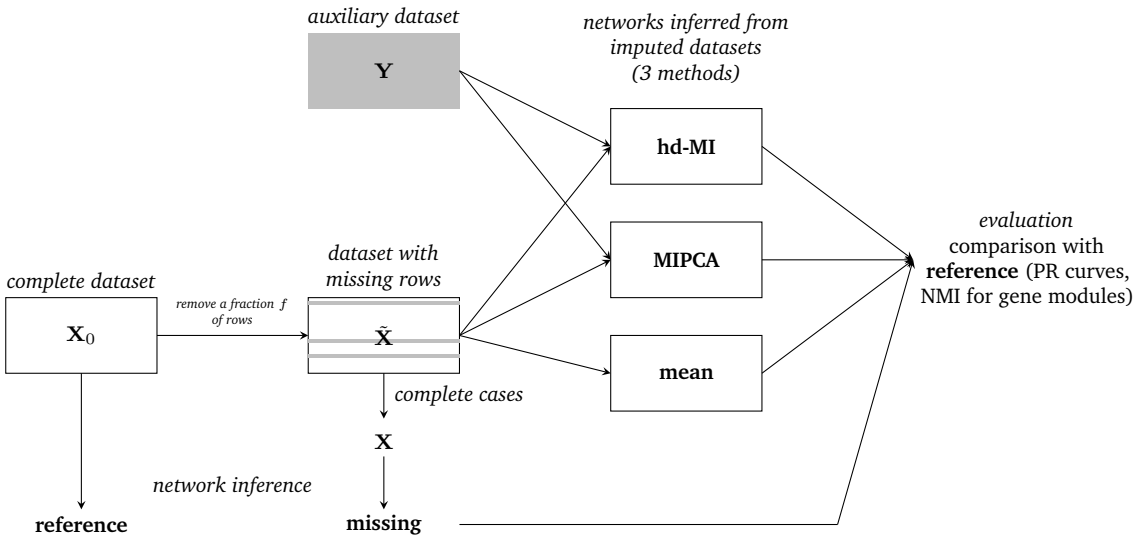


Figure 5.5 Schéma du processus d'évaluation.

De plus, pour les données issues de DiOGenes, 20 réplicats ont été réalisés pour évaluer la stabilité de nos conclusions.

Illustration sur la partie ≪ application ≫ de DiOGenes Finalement, deux réseaux (un pour CID1 et un pour CID2) ont été inférés en utilisant l'ensemble des individus disponibles des données d'expression RNA-Seq pour l'inférence et les données d'expression RT-qPCR comme données auxiliaires.

5.5 Résultats

5.5.1 Évaluation et comparaison avec d'autres méthodes existantes

Choix du paramètre σ

Pour choisir le seuil σ, nécessaire au calcul du score d'affinité, nous utilisons l'évolution de la variance moyenne intra-$\mathcal{D}(i)$ (voir section 5.3) en fonction de la valeur de σ. La figure 5.6 fournit un exemple du choix de cet hyperparamètre σ pour DiOGenes à CID1 avec 20% d'individus manquants.

Dans l'exemple illustré par la figure 5.6, la valeur choisie pour σ est donc 3.

Distribution des arêtes

La distribution d'apparition d'une arête dans les M ($M = 100$) réseaux inférés à partir des données imputées par **hd-MI** est représentée dans la figure 5.7 pour DiOGenes à CID1 avec 20% d'individus manquants. Les résultats obtenus pour GTEx sont similaires. Une large proportion d'arêtes est présente dans moins de 10% des réseaux. Il s'agit d'arêtes peu stables qui permettent de montrer la sensibilité de l'inférence de réseau à certains individus.

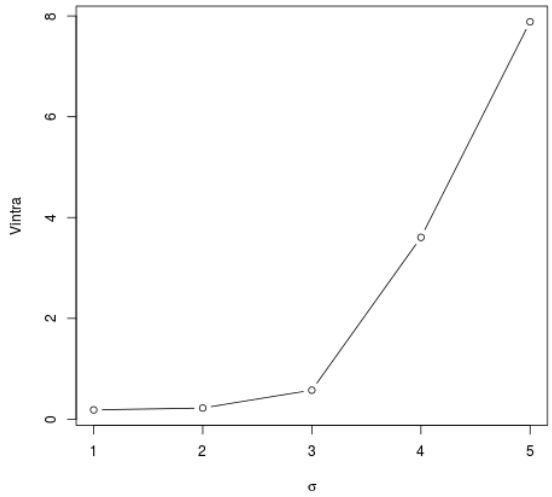


Figure 5.6 Choix de la valeur σ pour DiOGenes à CID1, 20% d'individus manquants.

Cependant, nous pouvons noter la présence d'une petite fraction d'arêtes toujours stables (c'est-à-dire présentes dans plus de 90% des réseaux).

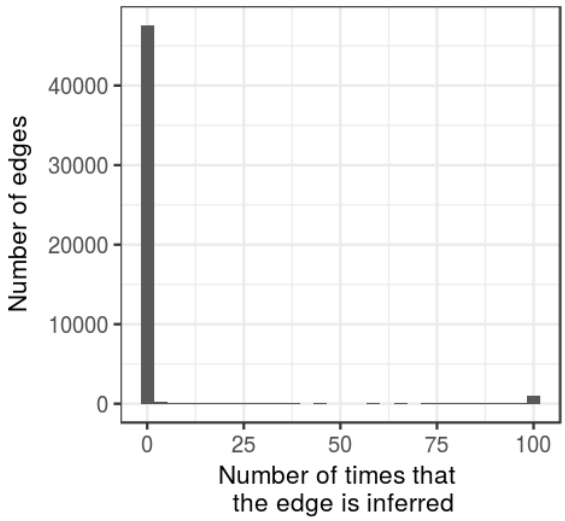


Figure 5.7 Distribution de l'apparition d'une arête dans les M ($M = 100$) réseaux pour les données imputées par **hd-MI**, 20% d'individus manquants, DiOGenes, CID1.

Lors de l'évaluation de la qualité des réseaux inférés, la valeur 0,9 pour le seuil r_0 a été choisie.

Propriétés globales des réseaux inférés

Les tableaux 5.1 (a) et 5.1 (b) donnent les propriétés globales de chaque réseau inféré, respectivement pour GTEx et DiOGenes à CID1 pour 20% d'individus manquants. Ces résultats montrent que la méthode **hd-MI** est en accord avec le réseau de référence en ce qui concerne ces mesures, même si le nombre d'arêtes inférées est légèrement inférieur par

rapport aux réseaux **reference**, **missing** et **mean**. En ce qui concerne l'approche **MIPCA**, les réseaux obtenus sont généralement peu denses, constitués de peu d'arêtes par rapport aux réseaux obtenus avec les autres approches.

(a) GTEx, 20% d'individus manquants

Réseau	♯ d'arêtes	Densité	Transitivité	Diamètre	♯ de la plus large composante connexe
reference	287	0.058	0.177	6	100
missing	292	0.059	0.169	7	100
mean	302	0.061	0.159	7	100
MIPCA	23	0.005	0.6	2	3
hd-MI	228	0.046	0.153	10	100

(b) DiOGenes, CID1, 20% d'individus manquants

Réseau	♯ d'arêtes	Densité	Transitivité	Diamètre	♯ de la plus large composante connexe
reference	1540	0.031	0.104	6	299
missing	1616	0.032	0.0997	6	300
mean	1609	0.032	0.099	6	299
MIPCA	158	0.003	0.040	20	103
hd-MI	1120	0.022	0.078	7	295

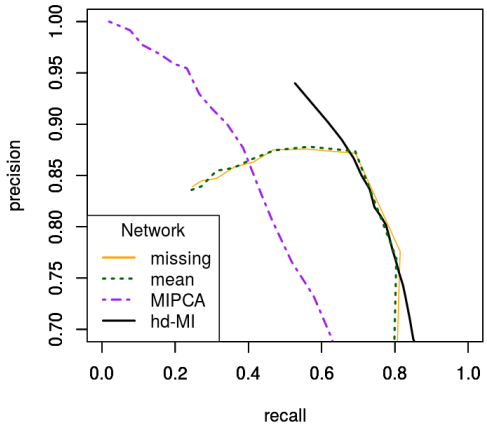
Tableau 5.1 **Propriétés globales des réseaux inférés pour GTEx (a) et DiOGenes à CID1 (b) pour 20% d'individus manquants.** Le nombre d'arêtes, la densité, la transitivité, le diamètre et la taille de la plus grande composante connexe sont donnés pour chaque réseau (**reference**, **missing**, **mean**, **MIPCA** and **hd-MI**). Pour **reference**, **missing** et **mean**, le réseau choisi est celui associé au ρ sélectionné par StARS. Pour **MIPCA** et **hd-MI**, le réseau choisi est celui obtenu avec un seuil de fiabilité r_0 fixé à 0,9.

Courbes de précision/rappel

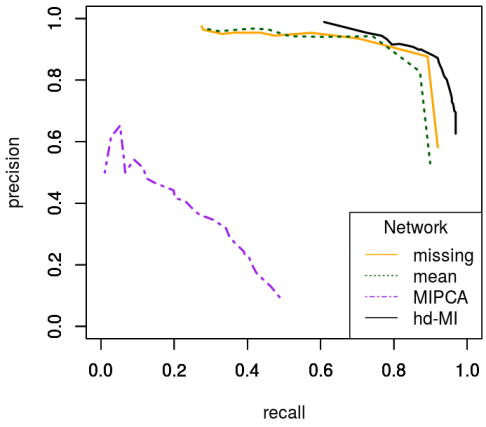
Les courbes de précision/rappel (PR) sont affichées dans la figure 5.8. Les courbes pour l'imputation par la moyenne et les données manquantes sont obtenues en variant le paramètre de régularisation ρ. Les courbes pour **MIPCA** et **hd-MI** sont obtenues en variant le seuil de fiabilité r_0. Les figures du haut montrent les résultats pour GTEx (figure 5.8 (a)) et DiOGenes à CID1 (figure 5.8 (b)) pour 20% d'individus manquants et les figures du bas (figures 5.8 (c), (d) et (e)) permettent de visualiser l'impact du taux d'individus manquants pour les données DiOGenes à CID1. Les courbes pour **mean** et **missing** sont similaires, montrant ainsi que des méthodes d'imputation naïve comme la moyenne ne permettent pas d'améliorer la qualité de l'inférence par rapport aux cas complets. Au contraire, **hd-MI** a le meilleur rappel pour les précisions les plus élevées.

MIPCA a les plus mauvaises performances : quel que soit le niveau de précision, **MIPCA** a la plus mauvaise courbe PR pour les données GTEx et DiOGenes. **hd-MI** est donc plus adaptée que **MIPCA** pour le cas de l'inférence de réseau. Deux raisons peuvent expliquer ceci. La première est que les co-apparitions des valeurs entre les différentes variables sont moins réalistes avec MIPCA puisque les données imputées, contrairement à l'approche **hd-MI** ne sont pas des données observées. La seconde, et la principale, est que **MIPCA** ne permet pas de poser des contraintes sur certaines caractéristiques des variables pour les valeurs imputées. En effet, des valeurs non pertinentes (par exemple, des valeurs négatives) sont parfois imputées et peuvent affecter fortement les résultats.

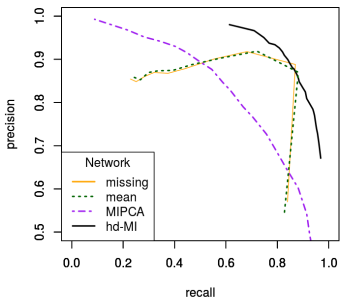
Pour les autres pourcentages d'individus manquants, les résultats restent similaires même si, lorsque le taux d'individus manquants augmente, les performances globales de toutes les méthodes se détériorent et les différences entre les méthodes tendent à diminuer.



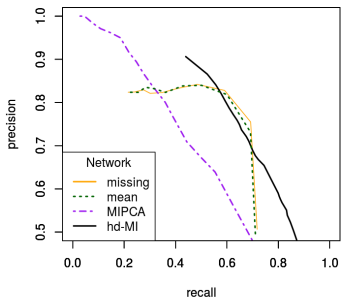
(a) DiOGenes - 20%



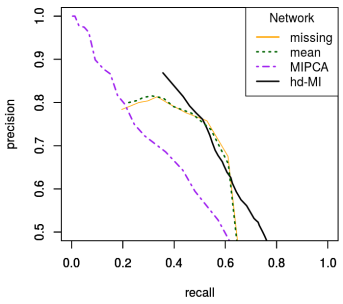
(b) GTEx - 20%



(c) DiOGenes - 10%



(d) DiOGenes - 30%



(e) DiOGenes - 40%

Figure 5.8 **Courbes Précision/Rappel pour chaque méthode et chaque jeu de données.** Les figures du haut montrent les résultats obtenus pour DiOGenes à CID1 (a) et GTEx (b) pour 20% d'individus manquants. Les figures du bas montrent les effets dus à la variation du taux d'individus manquants sur les données DiOGenes à CID1. Pour des précisions élevées, **hd-MI** a de meilleures valeurs pour le rappel, en particulier pour DiOGenes et avec le plus petit taux d'individus manquants.

Pour évaluer la stabilité des résultats, la procédure de simulation est répétée 20 fois pour le jeu de données DiOGenes, à CID1 avec 20% d'individus manquants. Les résultats montrent que seule la méthode **hd-MI** est capable d'obtenir systématiquement un bon rappel pour les précisions les plus élevées tandis que **missing** et **mean** réussissent à atteindre un taux de précision de 85% pour 18 courbes sur les 20 mais n'atteignent jamais un taux de précision de 90%. L'approche **MIPCA** atteint certes toujours le taux de précision ciblée mais avec des valeurs de rappel très faibles. Certaines statistiques de base (moyenne, minimum et maximum) pour le rappel pour ces deux taux de précision (85 et 90%) sont données respectivement dans les tableaux 5.2 (a) et 5.2 (b) . Ces résultats montrent que notre méthode **hd-MI** permet d'obtenir des réseaux plus stables et plus fiables puisqu'elle a une

(a) Précision fixée à 85%

Méthode	Min.	Moyenne	Max.
missing	0.352	0.649	0.746
mean	0.487	0.641	0.733
MIPCA	0.324	0.355	0.397
hd-MI	0.580	0.658	0.729

(b) Précision fixée à 90%

Méthode	Min.	Moyenne	Max.
MIPCA	0.227	0.277	0.310
hd-MI	0.545	0.593	0.655

Tableau 5.2 Statistiques des valeurs pour le rappel pour une précision fixée à 85% (a) et 90% (b). Pour des précisions fixées à 85%, pour les approches **missing** et **mean**, deux courbes n'ont pas atteint la valeur fixée pour la précision. Les valeurs dans ce cas ont été remplacées par le rappel ayant eu la précision la plus élevée.

variabilité nettement inférieure et un meilleur rappel (en moyenne) par rapport aux autres méthodes.

Impact du choix de la similarité dans la création du groupe de donneurs

Dans cette partie, nous évaluons l'impact de choix de la similarité utilisée pour créer le groupe de donneurs. Différentes similarités sont comparées :

— le score d'affinité décrit dans [58], calculé sur les données « brutes » ou sur les données réduites ;

— l'approche par k plus proches voisins (k-NN) pour laquelle le groupe de donneurs est composé des $k = 5$ plus proches voisins calculés sur le jeu $\mathbf{Y}$ pour un individu manquant dans $\mathbf{X}$. Le choix $k = 5$ est un choix habituel pour ce type de problème. De plus, pour ne pas être influencé par des corrélations possibles entre les variables, deux distances ont été testées : la distance euclidienne et la distance de Mahalanobis ;

— pour finir, une approche k-NN basée sur une analyse canonique des corrélations (CCA). Cette méthode s'appelle « voisin le plus similaire » (*most similar neighbor*, en anglais, noté MSN) et est disponible dans le package R **yaImpute** [60]. Cependant, cette méthode ne peut pas être directement appliquée sur nos données puisque le nombre d'échantillons est plus faible que celui des variables. Dans ce cas ($n \ll p$), la CCA est un problème mal posé. Pour pallier ce problème, il est nécessaire d'ajouter une pénalité au modèle. Nous avons donc adapté l'approche et utilisé une méthode k-NN basée sur une analyse canonique des corrélations régularisée avec une pénalité de type L_2 [225].

La comparaison entre ces approches est évaluée via les courbes de précision/rappel. Les courbes, pour le jeu de données DiOGenes à CID1 pour 20% d'individus manquants, sont illustrées par la figure 5.9. Elles montrent que ces différentes méthodes mènent à des résultats similaires en terme de qualité de l'inférence de réseau.

Modules de gènes

Comme expliqué dans [195] et [224], utiliser les réseaux de gènes pour identifier les groupes de gènes associés (modules de gènes) est plus pertinent que d'étudier les relations deux à deux entre les gènes. Pour évaluer la préservation des modules de gènes, une classification des sommets est réalisée en maximisant un critère de qualité, la modularité [157], d'une part dans le réseau de référence (afin d'obtenir des modules considérés comme modules de référence) et d'autre part dans tous les autres réseaux inférés (**missing, mean,**

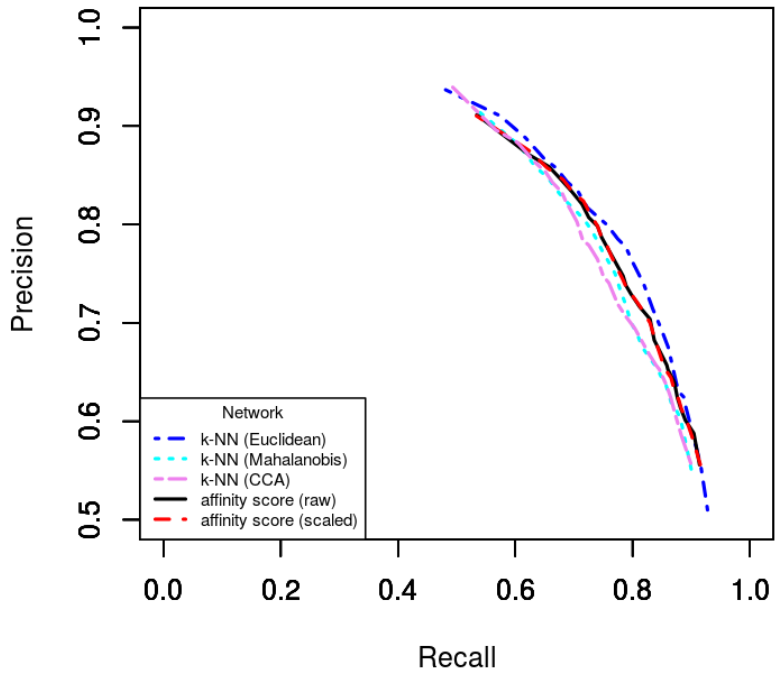


Figure 5.9 **Courbes précision/rappel obtenues avec différentes approches pour créer le groupe de donneurs.** Les différentes approches pour calculer la similarité sont le score d'affinité (sur données « brutes » ou données réduites) et les approches k-NN (avec une distance euclidienne, de Mahalanobis ou basée sur une CCA régularisée).

MIPCA, hd-MI). Pour éviter un nombre de groupes aberrant, la classification n'est effectuée que sur la plus large composante du réseau. Les ressemblances de structure des modules entre le réseau de référence et les autres réseaux sont ensuite évaluées en utilisant l'information mutuelle normalisée (NMI [1]) [63]. Le NMI est un critère de qualité prenant ses valeurs dans $[0, 1]$: il vaut 0 lorsque les modules entre deux réseaux sont totalement indépendants et 1 lorsque les modules sont identiques.

Le nombre de modules de gènes et les valeurs de NMI sont donnés dans les tableaux 5.3 et 5.4 pour GTEx et DiOGenes à CID1 respectivement, pour 20% d'individus manquants. Généralement, les modules de gènes sont mieux préservés avec la méthode **hd-MI** pour les données issues de GTEx quel que soit le taux d'individus manquants. Ce n'est pas le cas avec les données DiOGenes. Ceci peut s'expliquer par le fait que, dans ce dernier cas, **hd-MI** a des performances plus proches des résultats obtenus avec **missing** et **mean** que dans GTEX et que la ressemblance avec les modules du réseau de référence est artificiellement favorisée par la sélection des arêtes. En effet, la sélection des arêtes pour le réseau final est effectuée avec StARS dans le réseau de référence, celui avec les données manquantes (**missing**) et celui imputé par la moyenne (**mean**) et avec le seuil de fiabilité r_0 pour **hd-MI**. Cela correspond à deux niveaux différents de précision dans les courbes de précision/rappel. Cependant, même dans cette configuration, les modules obtenus avec **hd-MI** sont plutôt

1. *normalized mutual information*

similaires à ceux du réseaux de référence, illustrant le fait que notre méthode préserve bien la structure du réseau.

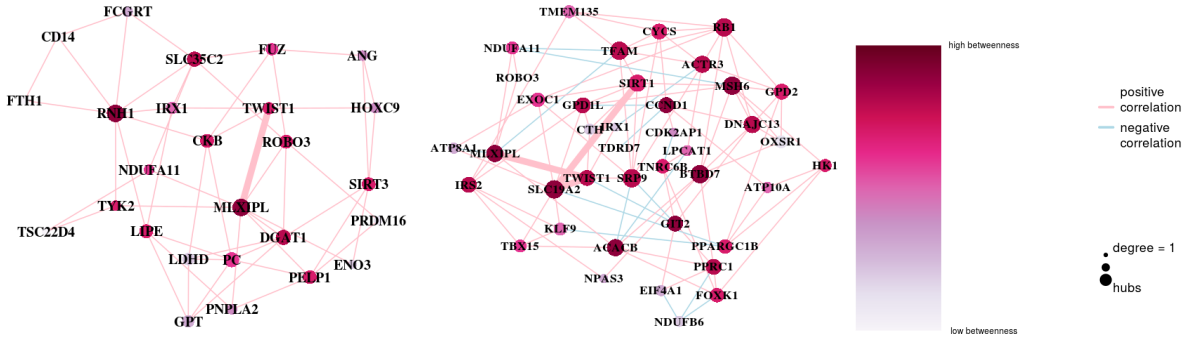
Réseau	reference	missing	mean	MIPCA	hd-MI
♯ modules de gènes	7	7	7	1	8
NMI		0.557	0.545	1 (3 gènes)	0.638

Tableau 5.3 GTEx : nombre de modules de gènes et NMI (comparaison avec les modules obtenus avec le réseau de référence), 20% d'individus manquants.

Réseau	reference	missing	mean	MIPCA	hd-MI
♯ modules de gènes	7	7	7	10	8
NMI		0.526	0.612	0.346	0.493

Tableau 5.4 DiOGenes : Nombre de modules de gènes et NMI (comparaison avec les modules obtenus avec le réseau de référence), CID1, 20% d'individus manquants.

5.5.2 Application sur l'ensemble des données du projet DiOGenes



(a) CID1 - Module 1 **(b)** CID2 - module 1 **(c)** Légende

Figure 5.10 **Module 1 pour (respectivement) CID1 (a) et CID2 (b) obtenu après classification des nœuds dans les réseaux obtenus avec notre méthode hd-MI.** La figure (c) correspond à la légende. Ces modules montrent les liens directs entre *TWIST1* et *MLX1PL* (a et b) et un nouveau lien de *TWIST1* à *SIRT1* via *SLC19A2* à CID2 (b).

Nous avons appliqué la méthode **hd-MI** sur des données d'expression du tissu adipeux mesurées via la technologie RNA-Seq. Les modules de gènes ont été extraits en utilisant l'optimisation de la modularité. Huit et sept modules de gènes ont été trouvés respectivement pour CID1 et CID2.

Certaines liaisons entre les gènes se retrouvent à la fois dans le réseau à CID1 et celui à CID2. Par exemple, l'arête présente à CID1 entre les gènes *MLX1PL* et *TWIST1* est également présente à CID2 (voir la figure 5.10).

Cependant, les deux réseaux montrent également des relations différentes entre les gènes. Ces différences montrent l'impact du régime hypocalorique sur les associations entre les gènes. Le gène humain *TWIST1* code pour un facteur de transcription qui s'exprime fortement dans les adipocytes chez des personnes minces. Il est positivement corrélé à la sensibilité à l'insuline et est un potentiel régulateur du remodelage du tissu adipeux [162].

À CID1, *TWIST1* est connecté à *PNPLA2*. À CID2, il est toujours connecté à *MLX1PL* mais aussi à *SLC19A2* qui est lui même lié à *SIRT1*. Le gène *SLC19A2* code pour un transporteur d'une vitamine, la thiamine, hTHTR-1, qui joue un rôle essentiel dans la glycolyse. *SLC19A2* est une des 41 régions de gènes candidats associées à la sélection naturelle positive et impliqués dans l'absorption et le métabolisme des nutriments [184]. Le gène *SIRT1* code une désacétylase qui régule diverses voies métaboliques [41]. La restriction calorique est connue pour favoriser l'expression de l'histone désacétylase. *MLX1PL* code pour le facteur de transcription, ChREBP, dont l'activité est induite par le glucose [80]. En coopération avec les gènes *TWIST1* et *SIRT1* après la restriction calorique (soit à CID2), il peut être considéré comme un capteur de glucose et un sensibilisateur à l'insuline.

Les réseaux inférés sont cohérents avec de précédents travaux effectués sur l'expression de ces gènes. La méthode **hd-MI** n'entraîne donc pas de distorsion dans la relation entre les gènes. Les échantillons d'ARNm provenant du tissu adipeux de ce projet DiOGenes ont déjà été utilisés dans de précédentes analyses à partir de données RT-qPCR [152, 222] et plus récemment avec des données RNA-Seq [12]. Des réseaux ont été inférés à partir des données RT-qPCR sur des sous-groupes composés d'hommes et de femmes [222] ou sur des groupes de femmes uniquement [152]. Plusieurs structures présentes dans ces deux études ont été retrouvées dans nos réseaux inférés avec **hd-MI**. En particulier, un module contenant le même groupe de gènes corrélés qui codent pour des enzymes impliquées dans la lipogénèse, incluant les gènes *FADS1*, *FADS2* et *AACS*, a été trouvé dans tous les réseaux à CID1. Plus intéressant encore, les liens entre *FADS1* et *AACS* et entre *FADS2* et *AACS* persistant dans le réseau à CID2 seulement dans le réseau inféré avec notre approche **hd-MI**. La persistance de ces arêtes peut être un effet positif de notre méthode d'imputation. En effet, dans l'optique de comparer des réseaux à différents pas de temps, il est préférable de les construire à partir des mêmes individus. Or, notre méthode permet d'imputer des individus manquants à un des deux pas de temps et donc d'augmenter le nombre d'individus communs (pour l'inférence de réseau) entre les deux temps.

L'étude présente, utilisant des données d'expression RNA-Seq, révèle de nouvelles structures par rapport aux analyses de réseaux des précédents travaux. Par exemple, un lien persistent a été trouvé entre deux facteurs de transcription impliqués dans la sensibilité à l'insuline, *TWIST1* et *MLX1PL* (ce dernier gène n'était pas présent dans la liste des gènes sélectionnés pour l'inférence de réseau lors des analyses précédentes) et de nouvelles connections entre gènes, comme *SIRT1* et *SLC19A2*, sont visibles après la restriction calorique (CID2).

5.6 Conclusion

Nous avons mis au point une méthode pour améliorer l'inférence de réseau effectuée à partir de données d'expression RNA-Seq en utilisant de l'information supplémentaire provenant d'un jeu auxiliaire. Celui-ci permet de calculer des similarités entre individus. La méthode **hd-MI** est basée sur deux méthodes d'imputation : l'imputation hot-deck et l'imputation multiple. Elle permet de préserver la structure de corrélation entre les variables en utilisant une approche hot-deck adaptée aux cas de non-réponse totale. De plus, de par son approche multiple, elle peut estimer l'incertitude liée à l'imputation. **hd-MI** permet

d'obtenir une meilleure précision pour la détection des arêtes que le cas complet ou que les méthodes d'imputation naïves. De plus, contrairement à ces approches, elle fournit une information sur la fiabilité d'une arête et sa sensibilité à l'absence d'individus.

hd-MI a été appliquée sur des données provenant d'un projet réel portant sur l'impact d'un régime hypocalorique sur l'expression des gènes du tissu adipeux. La méthode a réussi à fournir des réseaux pertinents. En effet, les résultats obtenus sur ces réseaux sont similaires à ceux obtenus sur de précédents réseaux, inférés à partir de données obtenues avec d'autres techniques de mesure d'expression et/ou avec d'autres sous-groupes d'échantillons. De plus, les réseaux obtenus avec **hd-MI** ont prédit la persistance des liens entre les *AACS*, *FADS1* et *FADS2* à CID2 (c'est-à-dire que les arêtes entre ces gènes sont encore présentes après le régime hypocalorique). Ils ont également permis de mettre en évidence un nouveau partenaire dans l'homéostasie glucidique dans le tissu adipeux : SLC19A2 (en plus des gènes *TWIST1* et *MLX1PL*). Son rôle précis dans ce phénomène biologique reste encore à étudier.

Intégration de différents jeux de données et inférence de réseau

Chapitre 6

Association de données transcriptomiques et de données cliniques

6.1 Introduction

L'obésité est caractérisée par un excès de masse grasse qui a des conséquences néfastes sur la santé (augmentation du risque de maladie cardiovasculaire, résistance à l'insuline, diabète de type II et cancers). Les interventions diététiques ont pour objectif de réduire la masse grasse, de restaurer la fonction du tissu adipeux.

La plupart des troubles métaboliques liés à l'obésité sont réversibles avec une perte de poids. Par exemple, l'amaigrissement permet généralement d'améliorer le contrôle glycémique. Cependant, une très grande variabilité entre les individus est observée dans leur capacité à perdre et à maintenir leur poids suite à un régime. En outre, chez les personnes obèses, les fluctuations du poids sont fréquentes puisque la perte de poids, due à un régime, conduit souvent à une reprise de poids à long terme [232, 194].

Les études transcriptomiques ont permis de mieux comprendre le fonctionnement du tissu adipeux lors d'essais contrôlés. Parmi les méthodes pour quantifier l'expression des gènes, la technique de séquençage haut débit RNA-Seq est de plus en plus utilisée. Cependant, cette méthode est encore relativement coûteuse. Une nouvelle technique de séquençage à haut débit, moins onéreuse mais aussi performante, a été mis au point récemment : la technique QuantSeq [150].

Les analyses basées seulement sur l'étude des données transcriptomiques ne fournissent pas assez d'informations pour comprendre tous les mécanismes biologiques impliqués. L'intégration de différentes sources d'informations permet d'obtenir une meilleure compréhension du système biologique dans son ensemble.

L'étude présentée ici cherche à déceler des caractéristiques du réseau d'expression des gènes dans le tissu adipeux qui soient cohérentes avec certains traits cliniques mesurés pendant une intervention diététique incluant une phase de restriction calorique suivie par une phase de suivi pondéral. L'objectif est de trouver quels changements d'expression de gènes sont associés à des changements de variables cliniques entre chaque phase (la phase hypocalorique et la phase de suivi pondéral) et entre le début et la fin de l'étude clinique. Connaître les liens entre les gènes et les données cliniques permettrait d'améliorer les connaissances sur les mécanismes biologiques conduisant aux pathologies associées à l'obésité.

L'intégration de différents types de données représente cependant un défi en biologie des systèmes. Une des questions soulevées par l'analyse intégrative est le problème des échelles qui sont différentes entre les ensembles de données. Autrement dit, l'intensité du lien de dépendance entre différents types de données (entre un gène et une variable

clinique ou entre deux gènes) a des échelles très différentes. Généralement, les relations entre un gène et une variable clinique se retrouvent masquées par les relations (d'intensité plus forte) entre deux gènes. Des approches statistiques multivariées ont été développées pour analyser conjointement différents ensembles de données 'omiques, tout en traitant le problème de grande dimension et en utilisant des méthodes de sélection de variables [91, 177] ou conjointement avec d'autres types de données [152] (pour plus d'informations sur les méthodes intégratives, voir les deux revues suivantes : [124, 25]).

Pour obtenir une modélisation globale du système, nous proposons une approche basée sur de l'inférence de réseau pour intégrer deux types de données : des données cliniques et des données transcriptomiques. Cette approche permet de donner un aperçu des interactions entre ces différents éléments.

La section 6.2 de ce chapitre présente les données, le plan d'analyse ainsi que la méthode d'inférence de réseau mis en place pour l'analyse intégrative. La section 6.3 présente les résultats obtenus avec les données provenant de l'étude DiOGenes.

6.2 Matériel et méthodes

6.2.1 Description des données

Les données présentées dans ce chapitre font partie des données collectées durant l'étude DiOGenes (pour une description plus détaillée de l'étude voir [125, 152]). DiOGenes est une étude d'intervention diététique menée dans huit pays européens.

Elle se déroule en deux phases, illustrées par la figure 6.1. Durant la première phase, les sujets obèses suivent un régime hypocalorique (800 kcal par jour) pendant huit semaines. Les sujets ayant réussi à perdre au moins 8% de leur poids initial participent à la seconde phase. Cette seconde phase est une phase de suivi pondéral de six mois. Durant cette seconde partie de l'étude, les sujets sont répartis aléatoirement dans cinq groupes de régime *ad libitum* : quatre régimes avec différentes teneurs en protéines et indice glycémique et un régime témoin (soit le régime type et équilibré du pays).

De nombreuses mesures biologiques ont été réalisées avant l'étude (CID1), après le régime hypocalorique (CID2) et après la phase de suivi pondéral (CID3) (voir figure 6.1). Parmi ces mesures, nous avons des mesures phénotypiques et cliniques ainsi que des mesures transcriptomiques du tissu adipeux.

Données cliniques À chaque pas de temps, les informations suivantes sur les individus sont disponibles :

— des données « d'identification » : le sexe, l'âge, le centre (pays où l'individu a suivi l'étude), etc. ;

— des données cliniques mesurant différents paramètres cliniques liés à l'obésité ou à des risques de maladies associées.

Parmi l'ensemble des données cliniques disponibles, nous avons sélectionné douze variables, pouvant être réparties dans trois groupes :

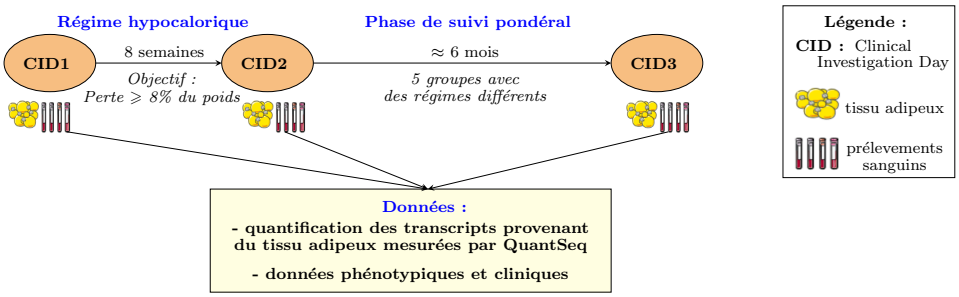


Figure 6.1 Schéma du protocole DiOGenes et données mesurées.

— **en relation avec le poids** : l'indice VAI [1], le tour de taille, l'IMC et le poids ;

— **en relation avec la sensibilité à l'insuline** : Matsuda, HOMA, le glucose et l'insuline plasmatiques ;

— **en relation avec les lipides circulants** : LDL, HDL, le cholestérol et la triglycéridémie (taux de triglycérides dans le sang).

Ces variables cliniques ont été choisies car elles sont indicatrices des risques associés à l'obésité (le diabète pour les variables en relation avec la sensibilité à l'insuline et les maladies cardia-vasculaires pour les variables en relation avec les lipides circulants). Par exemple, les variables HOMA et Matsuda ont été choisies puisqu'elles permettent de mesurer l'insulinorésistance.

Expressions de gènes L'ARN a été extrait à partir de biopsies de tissu adipeux sur 1051 échantillons correspondant à 599 individus. Le diagramme de Venn de la figure 6.2 précise le nombre d'échantillons disponibles pour chaque pas de temps. Certains individus ne sont pas observés à tous les temps. Une nouvelle technique de séquençage a été utilisée pour quantifier l'expression des gènes : le QuantSeq [150]. Pour le séquençage, les échantillons ont été placés aléatoirement sur 12 plaques constituées de 96 puits.

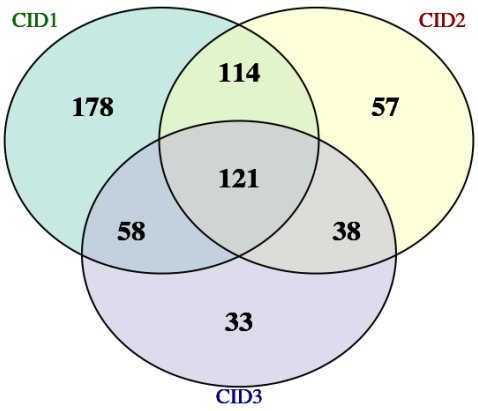


Figure 6.2 Nombre d'échantillons disponibles pour l'analyse QuantSeq pour chaque pas de temps de l'étude DiOGenes.

1. *Visceral Adiposity Index*

Lors de l'analyse des données QuantSeq, une fois les *reads* alignés, le nombre de *reads* est compté. Cinquante échantillons dont le nombre total de *reads* alignés était inférieur à 3 millions de *reads* (seuil minimal recommandé par le fournisseur Lexogen) ont été séquencés une seconde fois. Pour ces échantillons, de nouvelles librairies ont été générées avec une profondeur de séquençage plus élevée. Ces 50 échantillons ont été traités avec les 7 échantillons restants sur la dernière plaque (la plaque 12).

6.2.2 Analyse exploratoire

Les données ont tout d'abord été analysées par des méthodes statistiques univariées et multivariées par analyses en composantes principales (ACP). Le premier objectif est de rechercher la présence d'échantillons atypiques afin de les corriger ou de les supprimer pour la suite de l'analyse. Les analyses exploratoires ont également pour but de détecter des biais tels que des effets centre ou sexe.

L'analyse exploratoire a également permis d'apprendre à connaître les données de comptage obtenues avec la technique QuantSeq. Cette technique étant relativement récente, il est important d'étudier et de comprendre les caractéristiques de ces données afin de choisir les méthodes statistiques les plus adaptées pour les analyser.

6.2.3 Analyse différentielle

Modèle

Les données QuantSeq sont des données discrètes et hétérogènes. Comme pour les données RNA-Seq, nous avons pu constater une surdipersion, c'est-à-dire que la variance des gènes est généralement supérieure à la moyenne. Nous utilisons donc les méthodes d'analyse différentielle développées pour les données RNA-Seq. Ces méthodes sont basées sur des distribution binomiales négatives qui tiennent compte de la surdispersion.

Les données ont tout d'abord été normalisées avec la méthode de normalisation TMM [175]. Pour chaque contraste, seuls les individus appariés ont été utilisés. Cette approche permet de prendre en compte, dans le modèle, l'effet longitudinal (ou du moins l'appariement entre observations collectées à différents pas de temps) via l'introduction d'un effet individu. Pour rechercher les gènes différentiellement exprimés entre deux temps, nous utilisons un modèle linéaire généralisé (GLM).

Les données de comptage de l'échantillon i pour le gène j sont modélisées par une loi binomiale négative $NB(\mu_{ji}, \phi_j)$ (voir chapitre 2.1.4). Les covariables utilisées pour définir le plan expérimental sont l'individu et le pas de temps (CID). Utiliser l'individu comme covariable permet de corriger l'effet individuel et permet également de prendre en compte l'aspect temporel. Le modèle GLM utilisé est donc le suivant :

$$\log(\mu_{ji}) = \beta_{2,j,\text{CID}(i)} + \beta_{1,j,i} + \log(N_i) \tag{6.1}$$

où $\beta_{1,j,i}$ est l'effet (fixe) de l'individu i et β_{2,j,CID_t} est l'effet (fixe) du pas de temps, $\text{CID}_t = \text{CID}(i)$, de l'individu i. Nous cherchons à savoir si le gène j s'exprime différemment entre deux temps (soit entre deux CID). L'hypothèse à tester est donc la suivante :

$$\mathcal{H}_{0j} : \{\beta_{2,j,\text{CID}_t} = \beta_{2,j,\text{CID}_{t'}}\}.$$

qui est réalisé par un test du rapport de vraisemblance.

Cette analyse a été effectuée avec le package **edgeR** [174]. Une correction pour les tests multiples a été effectuée en utilisant l'approche proposée par [24]. Les gènes ont été considérés comme différentiellement exprimés entre deux temps si leur p-valeur ajustée était inférieure à 5%.

Sélection des gènes

Le nombre d'observations collectées (n) est généralement très faible devant le nombre de gènes (p). [221] a montré qu'on ne peut espérer une estimation satisfaisante d'un réseau d'interactions géniques à l'aide de modèles gaussiens parcimonieux dans le cadre de l'ultra haute dimension. Il est donc important de restreindre le nombre de gènes pour pouvoir inférer le réseau. Pour cela, divers critères ont été utilisés pour réduire et obtenir un nombre acceptable de gènes qui permet d'estimer convenablement les paramètres avec les modèles graphiques.

Un filtre a été appliqué sur les données afin de supprimer les gènes comprenant trop de comptages nuls ou dont le logarithme du changement d'expression (log Fold-Change, noté logFC dans la suite) est manquant. Le seuil de données manquantes ou indiquant une expression très faible a été fixé à 25%. Pour la suite des analyses, deux listes de gènes sont créées :

— La première liste correspond aux gènes différentiellement exprimés pour les différents contrastes, soit les gènes dont la p-valeur ajustée par la méthode de [24] est inférieure à 5%. Cependant, le nombre de gènes dans cette liste est encore trop élevé pour les modèles d'inférence de réseaux.

— Un second seuil en fonction du niveau de changement d'expression (Fold-Change, FC) a donc été utilisé pour restreindre le nombre de gènes dans chaque contraste. Après avoir testé plusieurs valeurs, le seuil pour le FC a été fixé à 1,3. Ce seuil a été choisi pour maximiser le nombre de gènes retenus tout en restant dans un rapport nombre de gènes / nombre d'observations compatible avec l'inférence de réseau par approche GGM. La seconde liste est donc composée des gènes différentiellement exprimés (p-valeur ajustée inférieure à 5%) et dont l'expression est suffisamment régulée entre entre conditions ($|\text{FC}| > 1.3$)

6.2.4 Intégration des données et inférence de réseau

Les relations entre les variables cliniques d'intérêt et les expressions ont été étudiées en utilisant deux approches différentes : des modèles linéaires mixtes et une approche basée sur de l'inférence de réseau. Nous présentons ici plus particulièrement l'approche d'inférence de réseau utilisée.

Pour chaque contraste (CID1/CID2, CID2/CID3 et CID1/CID3), un réseau est inféré. Seuls les individus avec des données appariées pour deux temps (CID) sont utilisés. Les données utilisées dans cette partie sont :

— les logFC de l'expression des gènes ;

— les variables cliniques d'intérêt sélectionnées ;

— quelques variables pour ajuster les modèles : le sexe, l'âge et le centre.

Association gènes et variables cliniques

Nous cherchons à voir quels changements d'expression de gènes sont associés avec un changement de valeurs des variables cliniques entre deux pas de temps de l'étude (soit pour chaque contraste).

Les relations entre les gènes et une variable clinique Y ont été estimées à l'aide du modèle linaire mixte [165] suivant :

$$Y_{t'} = \beta_0 + \beta_1 \text{logFC}_{t't} + \beta_2 Y_t + \beta_3 X_{\text{sexe}} + \beta_4 X_{\text{âge}} + u_1 Z_{\text{centre}} + \epsilon \tag{6.2}$$

où $Y_{t'}$ est la variable clinique d'intérêt Y mesurée au temps t' ($t' \in \{CID2, CID3\}$). Les variables explicatives sont Y_t la variable clinique Y mesurée au temps t ($t \in \{CID1, CID2\}$, $t < t'$) et $\text{logFC}_{tt'}$ le logFC de l'expression d'un gène entre les deux temps t' et t. Le modèle est ajusté par trois variables : l'âge et le sexe qui sont considérés comme des effets fixes et le centre considéré comme un effet aléatoire. u représente l'effet aléatoire et suit une loi normale $\mathcal{N}(0, \sigma^2)$. ϵ est le terme d'erreur du modèle et suit une loi normale $\mathcal{N}(0, \sigma_\epsilon^2)$.

Un modèle mixte est ajusté pour chaque variable clinique, pour chaque gène et pour les différents contrastes. Pour chaque contraste, les effets de changements d'expression de gènes sont évalués en testant $\beta_1 = 0$. Des corrections de tests multiples [24] sont ensuite réalisées. L'association entre un gène et une variable clinique est considérée comme significative si sa p-valeur ajustée est inférieure à 10%.

Inférence de réseau

Une première approche naïve est de combiner les deux ensembles de données (expression de gènes et variables cliniques d'intérêt) et d'inférer un réseau à partir de l'ensemble de ces données. Néanmoins, les relations entre les gènes dominent et masquent les relations entre les gènes et les variables cliniques. Le résultat obtenu est un réseau où les variables cliniques n'ont aucun lien avec les gènes. Il convient donc de construire les liens entre gènes et variables cliniques via une autre méthode.

Nous nous inspirons de l'approche proposée dans [152] en proposant une approche d'inférence de réseau en plusieurs étapes.

La première étape consiste à inférer un réseau de gènes pour chaque contraste. Pour chaque contraste, nous travaillons seulement avec les échantillons appariés entre les deux temps et calculons les logFC de l'expression des gènes différentiellement exprimés pour ce contraste.

Des données manquantes sont présentes parmi les logFC calculés. Une méthode d'imputation multiple basée sur l'ACP (MIPCA) [117] est alors employée. Nous avons utilisé le package R **missMDA** [114].

Contrairement aux données d'expression QuantSeq qui sont discrètes, les données logFC sont des données continues dont la distribution est proche d'une distribution normale. Pour cette raison, nous utilisons un modèle d'inférence de réseau adapté à ce type de données : le modèle graphique gaussien [85] (voir Section 2.2.2 du chapitre 2). Un réseau de gènes est inféré pour chaque contraste. Le nombre d'arêtes sélectionnées est choisi via une pénalisation de type lasso ajoutée au problème. Le paramètre de régularisation est choisi avec le critère RIC [140, 244].

La seconde étape consiste à estimer les relations significatives entre les gènes et les variables cliniques. Un modèle mixte (voir section 6.2.4) est ajusté pour chaque variable clinique, pour chaque gène et pour les différents contrastes. Des corrections de tests multiples ([24], BH) sont ensuite réalisées pour chaque contraste. L'association entre un gène et une variable clinique est considérée comme significative si sa p-valeur ajustée est inférieure à 10%.

La dernière étape consiste à construire le réseau final mettant en relation les expressions de gènes avec des variables cliniques d'intérêt. Les arêtes entre les variables (cliniques et gènes) considérées comme significatives lors de l'étape précédente sont ajoutées au réseau de gènes inféré à la première étape. La figure 6.3 résume les différentes étapes pour l'analyse de réseau.

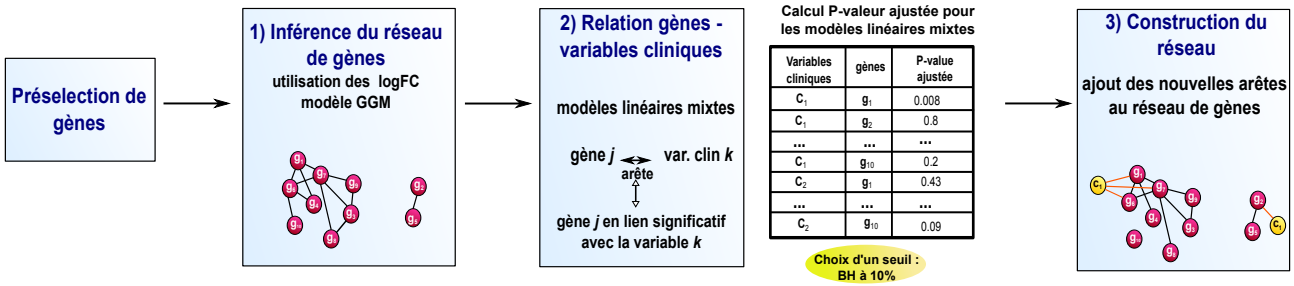


Figure 6.3 **Les différentes étapes de l'analyse de réseaux.** 1) Un réseau de gènes est inféré. 2) Les associations entre variables cliniques et expression de gènes sont ensuite évaluées avec des modèles linéaires mixtes. 3) Une arête entre une variable clinique et un gène est ajoutée sur le réseau initial si leur association est considérée comme significative dans les modèles linéaires mixtes (après correction de tests multiples en utilisant la méthode BH).

Les réseaux ont été inférés avec le modèle graphique lasso implémenté dans le package R **huge** [244]. Nous avons utilisé le le package R **nlme** [166] pour les modèles mixtes.

6.3 Résultats

6.3.1 Description des données et analyse exploratoire

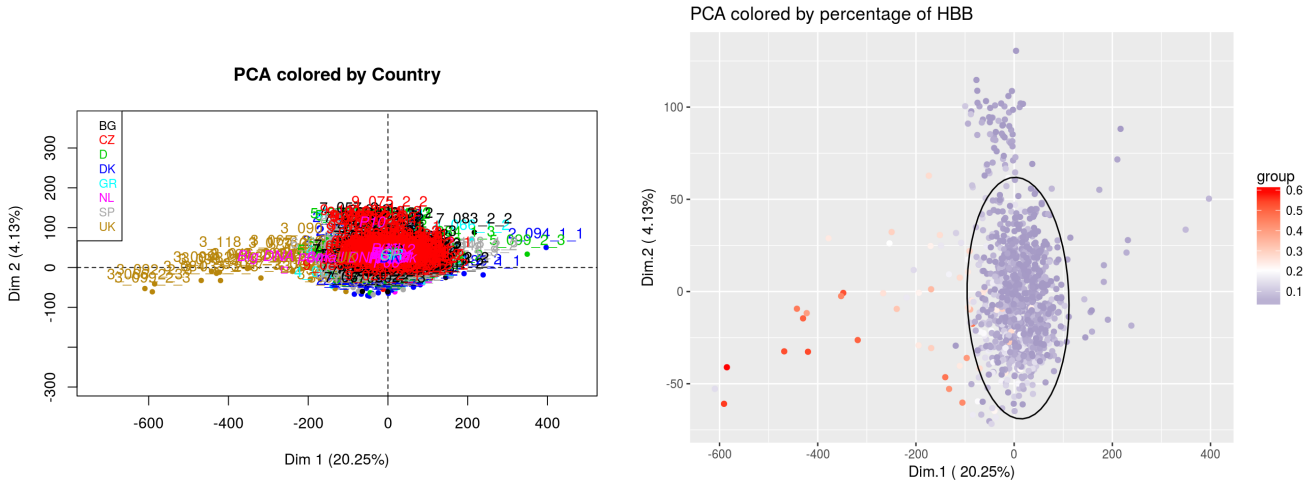
Diverses ACP ont été effectuées sur les données afin d'en vérifier la qualité et de procéder à des pré-traitements permettant d'obtenir des données épurées ; c'est-à-dire des données directement utilisables et interprétables pour les analyses statistiques. L'objectif des ACP est donc d'identifier des biais expérimentaux afin de les corriger, de repérer des valeurs atypiques pour les corriger, voire les supprimer.

La figure 6.4 (a) fournit la projection des individus sur les deux premiers axes de l'ACP de l'ensemble des données (après suppression de trois individus aberrants), dans laquelle chaque échantillon est colorié par le centre qui l'a collecté. Nous remarquons que les

échantillons provenant du centre UK se distinguent des autres échantillons. La figure 6.4 (b) correspond à la même représentation des individus mais colorés par le taux de contamination sanguine de la biopsie (ratio du comptage du gène HBB qui code pour la β-globine sur le comptage total). Les deux graphiques de la figure 6.4 mettent en évidence un effet centre dont une cause probable est la qualité d'exécution de la biopsie (protocole différent selon les pays).

Pour la suite des analyses statistiques, nous avons choisi de supprimer les échantillons dont les biopsies ont été trop contaminées par du sang. En effet, le transcriptome des cellules sanguines et celui du tissu adipeux sont différents et garder de tels échantillons risque d'introduire de faux positifs dans l'analyse différentielle des gènes du tissu adipeux.

L'information de contamination par le sang est accessible via une variable additionnelle qui a été utilisée pour supprimer les échantillons contaminés, le HBB. La figure 6.5 représente la distribution du pourcentage en HBB en fonction des plaques. Quelle que soit la plaque, des échantillons avec des taux en HBB assez élevés sont présents. À l'aide de ce graphique, le seuil arbitraire de 20% pour le pourcentage en HBB a été choisi.



(a) Coloration en fonction du centre

(b) Coloration en fonction du % HBB

Figure 6.4 **Projection des individus sur les deux premiers axes de l'ACP.** (a) Selon le centre, représenté par la couleur (Pays-Bas (NL), Danemark (DK), Royaume-Uni (UK),Grèce (GR), Bulgarie (BG), Allemagne (D), Espagne (SP) et République Tchèque (CZ)). (b) Selon la contamination de l'échantillon en cellules sanguine. La couleur représente le pourcentage de la variable HBB (β-globine). Plus la couleur est rouge, plus la contamination sanguine est importante.

Les échantillons provenant de deux plaques (10 et 12) semblent différents des échantillons provenant des autres plaques (figure 6.8 (a)). Pour la plaque 12, cette différence s'explique par des librairies différentes. Nous pouvons donc espérer que normaliser les données résoudra le problème. En revanche, pour la plaque 10, les causes de cette différence n'ont pas été identifiées.

La figure 6.6 montre l'évolution du nombre d'échantillons au cours du nettoyage des données : le nombre d'échantillons initial est égal à 1051 au début de l'analyse, puis à 997 après l'analyse exploratoire et se réduit à 918 lorsque les échantillons provenant de la plaque 10 sont retirés de l'étude.

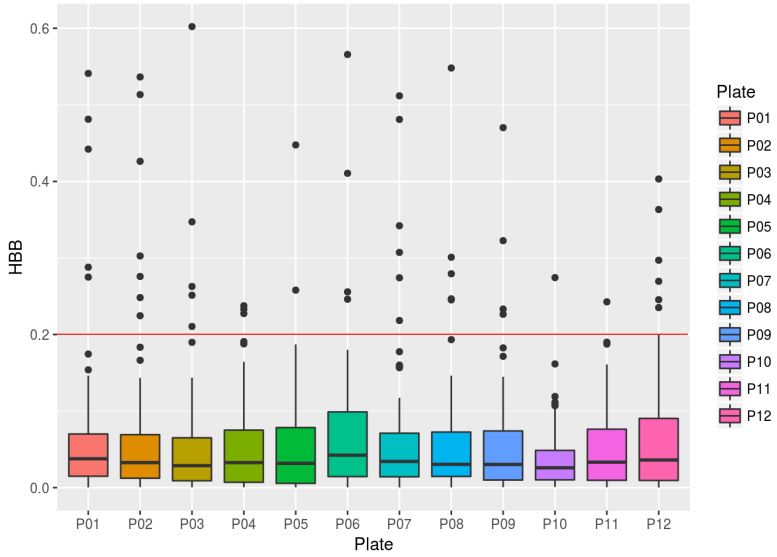
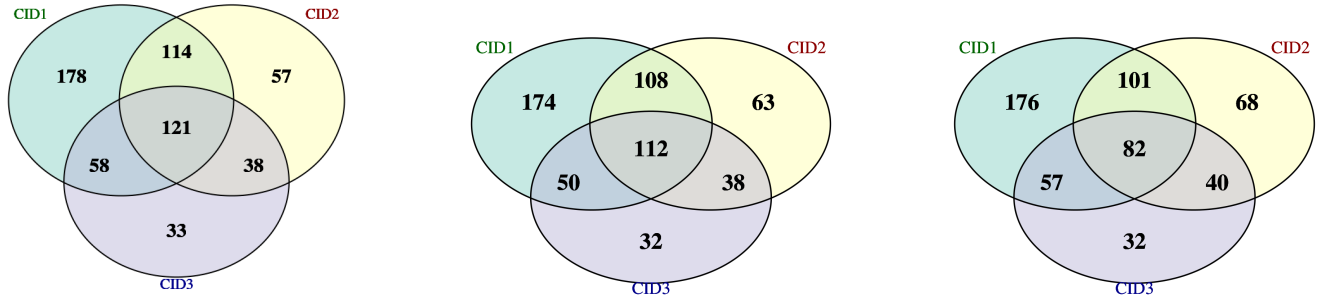


Figure 6.5 **Choix du seuil pour le % de contamination en sang.** La ligne rouge correspond au seuil fixé (20%).



(a) Tous les échantillons disponibles **(b)** Après suppression des échantillons atypiques **(c)** Sans les échantillons de la plaque 10

Figure 6.6 **Évolution du nombre d'échantillons pour les trois temps (CID) lors de l'analyse exploratoire :** (a) au début de l'étude : 1051 échantillons (599 individus), (b) après la suppression des échantillons aberrants : 997 échantillons (577 individus) et (c) sans les échantillons provenant de la plaque 10 : 918 échantillons (556 individus).

En conclusion, 54 échantillons ont été supprimés : trois dont l'ARN était trop dégradé, 50 échantillons dont la biopsie du tissu adipeux était trop contaminée par du sang et un échantillon aberrant (après suppression des 53 échantillons). Deux plaques se sont également révélées atypiques. Pour pallier ce problème, nous avons adopté la stratégie suivante :

— normaliser les données pour réduire l'atypicité des plaques ;

— incorporer un effet plaque dans les modèles estimés afin de corriger l'effet des plaques restées atypiques après normalisation.

Cette stratégie a été comparée avec la suppression des plaques des données.

La figure 6.7 résume les différentes étapes de l'analyse des données d'expression de gènes : de l'analyse exploratoire des données transcriptomiques mesurées avec la nouvelle

technique QuantSeq jusqu'à l'analyse intégrative de ces données avec des données cliniques en utilisant une approche basée sur des réseaux.

6.3.2 Analyse différentielle

Les données ont été normalisées avec l'approche TMM. La figure 6.8 montre le résultat d'une ACP avant et après normalisation, avec une coloration en fonction de la covariable plaque. Comme attendu, après normalisation, les échantillons provenant de la plaque 12 ne diffèrent pas des autres échantillons. Cependant, la normalisation n'a pas pu corriger le biais des échantillons provenant de la plaque 10 (voir figure 6.8 (b)).

Une analyse différentielle a été réalisée également avec l'ensemble des plaques. Pour ce faire, le modèle GLM (équation (6.1), section 6.2.3) a été légèrement modifié afin de corriger le biais technique dû à la plaque. Plus précisément, la covariable « plaque » a été ajoutée au modèle GLM. Les résultats (non montrés) se sont avérés similaires à ceux présentés ici (en supprimant la plaque atypique et en utilisant le modèle de l'équation (6.1)). Nous avons donc opté pour le choix de ne pas garder la plaque 10 dans la suite de l'analyse (résultats plus conservatifs).

Nous observons des changements d'expression de gènes dans le tissu adipeux entre les différentes phases de l'intervention. Quel que soit le contraste étudié, nous avons environ 5000 gènes différentiellement exprimés. Plus précisément, 6648 gènes sont différentiellement exprimés entre CID1 et CID2, 5474 entre CID1 et CID3 et 4691 entre CID2 et CID3. Nous pouvons remarquer que 1257 gènes sont différentiellement exprimés par les trois contrastes.

Le diagramme de Venn de la figure 6.9 récapitule le nombre de gènes différentiellement exprimés entre les différents contrastes.

6.3.3 Intégration des données et inférence de réseau

Pour chaque contraste, l'inférence de réseau a été réalisée en suivant les différentes étapes de la méthode d'inférence proposée dans la section 6.2.4. La figure 6.10 donne le nombre de gènes utilisés pour l'inférence de réseau pour chaque contraste. Ces gènes correspondent à la seconde liste des gènes sélectionnés pour les analyses statistiques (voir section 6.2.3). Nous avons retenus 541 gènes pour le contraste CID1/CID2, 661 pour le contraste CID2/CID3 et 470 pour le contraste CID1/CID3. Les nombres de gènes et d'observations utilisés pour l'inférence de réseau pour chaque contraste sont résumés dans le tableau 6.1.

Contraste	Nombre d'observations (individus appariés)	Nombre de gènes
CID1/CID2	183	541
CID2/CID3	122	661
CID1/CID3	139	470

Tableau 6.1 Nombre d'observations et de gènes utilisés pour l'inférence de réseau pour chaque contraste.

Le tableau 6.2 résume les résultats obtenus avec les modèles linéaires mixtes en indiquant le nombre de gènes significativement associés à chaque variable clinique pour les différents contrastes. Nous remarquons un nombre plus important de gènes dont les change-

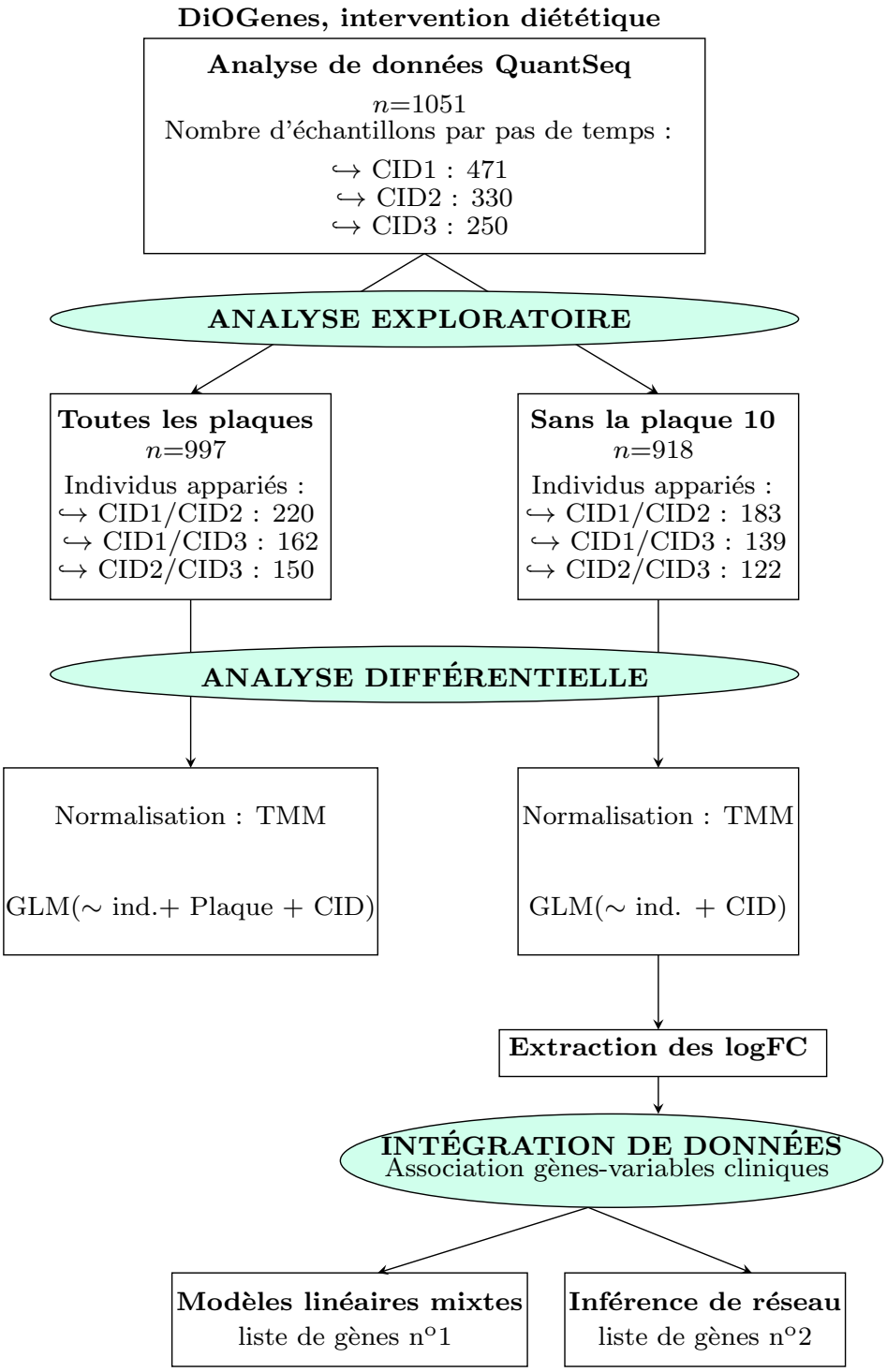
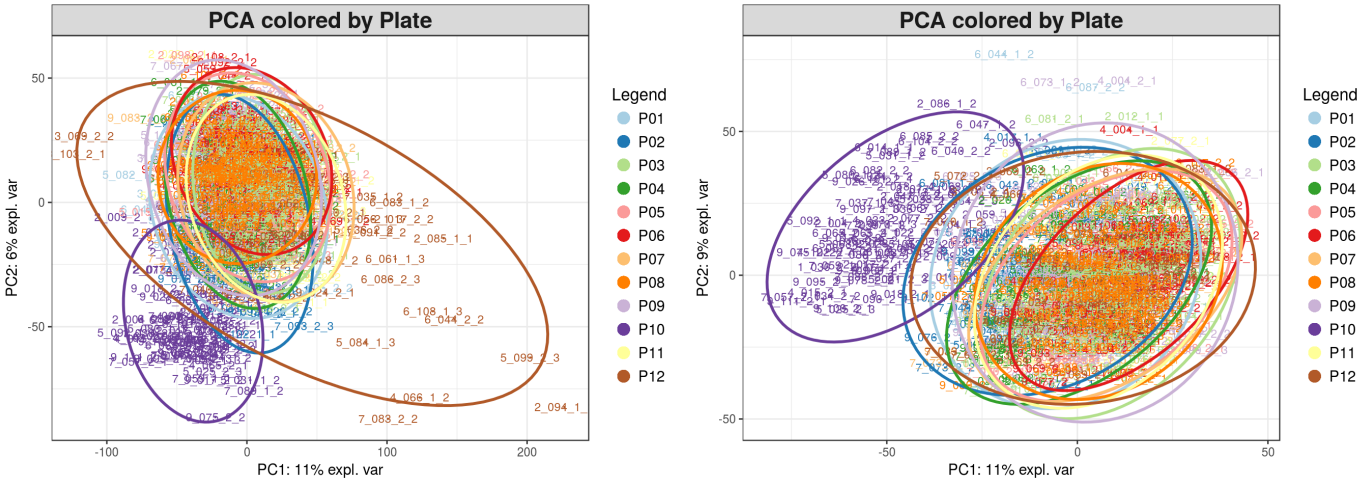


Figure 6.7 **Organigramme de l'analyse des données Quantseq.** L'analyse exploratoire a permis de supprimer 54 échantillons aberrants et a mis en évidence l'atypicité de deux plaques. Pour l'analyse différentielle, deux modèles GLM ont donc été testés. Nous avons choisi de ne pas garder la plaque 10 pour la fin de l'analyse. Nous avons ensuite extrait les logFC des gènes différentiellement exprimés pour effectuer deux approches intégratives : des modèles linéaires mixte et l'approche basée sur de l'inférence réseau (présentée dans la section 6.2.4).



(a) Avant normalisation TMM

(b) Après normalisation TMM

Figure 6.8 **ACP des échantillons avant (a) et après normalisation (b) TMM des données d'expression.** Les couleurs correspondent aux plaques.

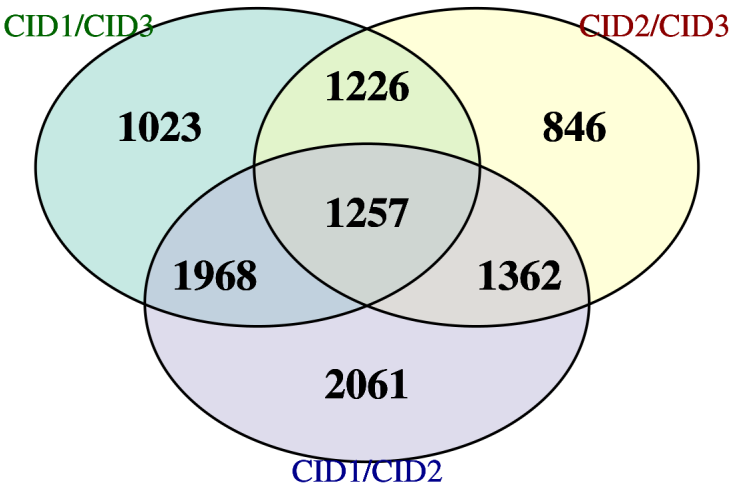


Figure 6.9 Diagramme de Venn du nombre de gènes différentiellement exprimés entre les différents contrastes.

ments d'expression sont liés aux changements de variables cliniques avec le contraste CID1/CID2 (première phase, régime hypocalorique), comparativement aux contrastes CID2/CID3 (deuxième phase, suivi post régime hypocalorique) et CID1/CID3 qui récapitule l'effet à long terme (6 mois) de l'amaigrissement provoqué par le régime hypocalorique.

La figure 6.11 permet de voir le nombre de gènes associés au poids et à l'IMC pour le poids (figure 6.11 (a)) et pour la variable IMC (figure 6.11 (b)), entre les différents contrastes.

Pour souligner la structure du réseau, une classification des nœuds a été effectuée en maximisant le critère suivant : la modularité [157]. Nous avons utilisé le modèle « spin-

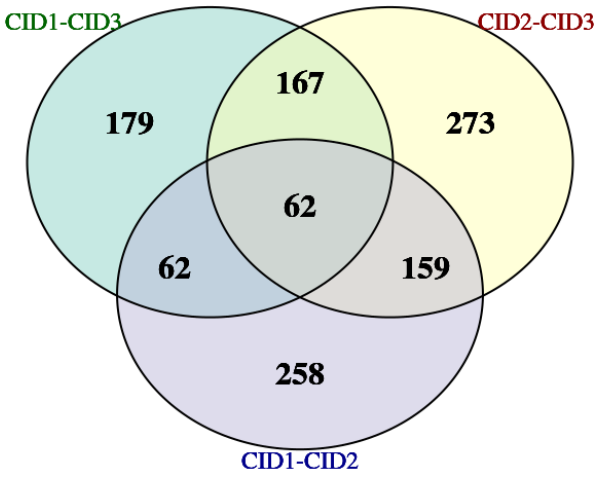


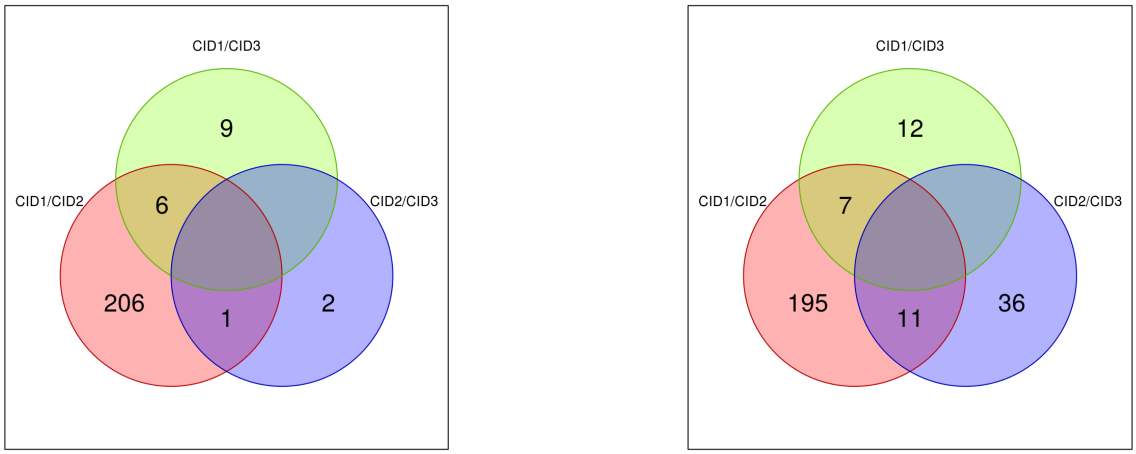
Figure 6.10 Diagramme de Venn des gènes sélectionnés avec la p-valeur ajustée (BH 5%) et $|FC| > 1.3$.

Variables	Contraste CID1/CID2	Contraste CID2/CID3	Contraste CID1/CID3
Nombre total de gènes testés	541	661	470
VAI (VAI)	0	0	0
Tour de taille (waist)	59	0	0
IMC (BMI)	213	47	19
Poids (weight)	213	3	15
Matsuda (dmatsu)	0	0	0
HOMA (dhomares)	1	0	0
Glucose plasmatique (dglu0)	33	0	0
Insuline plasmatique (dins)	0	0	0
LDL (dldl)	6	0	1
HDL (hdl)	0	10	0
Cholestérol (chol)	74	0	0
Triglycéridémie (trig)	0	0	0

Tableau 6.2 Nombre de gènes associés avec une variable clinique pour chaque contraste et nombre de gènes testés par contraste.

glass » qui est une approximation de l'optimisation de la modularité (méthode disponible dans le package R **igraph** [61]). Les figures 6.12, 6.13 et 6.14 représentent respectivement les réseaux obtenus pour les contrastes CID1/CID2, CID2/CID3 et CID1/CID3.

La fonction biologique des protéines codées par les gènes de chaque module a été recherchée à l'aide du logiciel Ingenuity Pathways Analysis (IPA, Ingenuity Systems ; QIAGEN, Inc., Valencia, CA, USA, `https://analysis.ingenuity.com/pa`) qui permet d'interroger la base de connaissances Ingenuity Pathways Knowledge Base pour en extraire des informations sur les annotations fonctionnelles et les interactions entre gènes et/ou protéines. Lors de la première phase, les gènes reliés aux indices de corpulence (poids et IMC) codent pour des protéines participant à l'immunité et à la réponse inflammatoire, alors que ceux associés aux indices de sensibilité à l'insuline (glycémie et HOMA) codent pour des protéines



(a) Avec la variable poids

(b) Avec la variable IMC

Figure 6.11 Diagramme de Venn entre les différents contrastes du nombre de gènes associés à la variable clinique (a) poids et (b) IMC.

du métabolisme lipidique et de la survie cellulaire. La voie du métabolisme lipidique est ensuite associée au poids et l'IMC pendant la deuxième phase (contraste CID2/CID3). Cette association persiste lors de la comparaison entre le début et la fin de l'intervention (contraste CID1/CID3). Ces observations indiquent que le régime hypocalorique impacte fortement le tissu adipeux dont la réponse transcriptionnelle évolue ensuite sur d'autres voies biologiques

6.4 Conclusion

Le QuantSeq est une technologie récente. Elle s'avère être intéressante pour les analyses biologiques puisqu'elle est moins onéreuse que la technologie RNA-Seq, moins sensible à la qualité de l'ARN mais tout aussi performante. Elle permet donc d'augmenter le nombre d'échantillons (et par conséquent la puissance). Ces premières analyses QuantSeq permettent de montrer que les problématiques rencontrées pour ces données sont similaires à celles des données RNA-Seq. Nous avons en effet montré que les données QuantSeq sont des données discrètes, hétérogènes et surdispersées. Il semble donc logique d'employer les méthodes développées pour l'analyse de données RNA-Seq.

Les réseaux sont des modèles utiles pour étudier les relations entre les variables. Cependant, les forces des relations entre différents types de données ont des niveaux très différents. Ce problème est contrôlé ici en utilisant une approche d'inférence non globale afin d'obtenir un modèle global des interactions entre les différents ensembles de données. Pour cela, nous proposons une méthode d'inférence en trois étapes pour inférer un réseau global à partir des différents types de données : tout d'abord, pour chaque contraste, un réseau est inféré à l'aide d'un modèle graphique gaussien à partir des logFC d'expression. Dans un second temps, nous cherchons les liens forts entre les expressions de gènes et les variables cliniques en utilisant des modèles linéaires mixtes. Pour finir, à partir des résultats obtenus dans la

seconde étape, nous ajoutons au réseau de gènes les variables cliniques et les arêtes entre gènes et variables cliniques.

L'approche proposée permet ainsi d'intégrer à la fois des données transcriptomiques et des données cliniques afin d'avoir une vue globale et exhaustive sur les interactions entre les différents types de données et de les relier à des événements cliniques, tels les variations de poids.

Utiliser des modèles linéaires mixtes a permis de trouver les relations entre gènes et variables cliniques. En outre, cette méthode a l'avantage d'ajuster le modèle par des effets fixes et aléatoires. Les variables utilisées pour l'ajustement sont soit connues pour être une source importante de variabilité entre les individus (sexe, âge), soit pouvant introduire des biais importants (centre). L'analyse intégrative basée sur une approche réseau permet de faciliter l'analyse globale du système biologique grâce à sa visualisation sous la forme d'un réseau.

Cette analyse présente cependant des limites. Seuls les individus appariés entre deux contrastes ont été utilisés. Ceci restreint énormément le nombre d'observations utilisées pour l'inférence de réseau. En outre, lors de la deuxième phase, il y a une forte hétérogénéité de comportement des sujets. En effet, certains ont volontairement maintenu une restriction calorique et ont donc continué à perdre du poids, alors que d'autres ont repris tout le poids perdu lors de la première phase. De la même manière, la majorité des sujets (mais pas tous) résistants à l'action de l'insuline lors de l'inclusion dans le programme ont amélioré ce paramètre avec le régime hypocalorique. Parmi eux, certains ont maintenu l'amélioration alors que d'autres sont redevenus résistants à l'insuline. Ces aspects n'ont pas été pris en compte dans les analyses pour les contrastes CID2/CID3 et CID1/CID3.

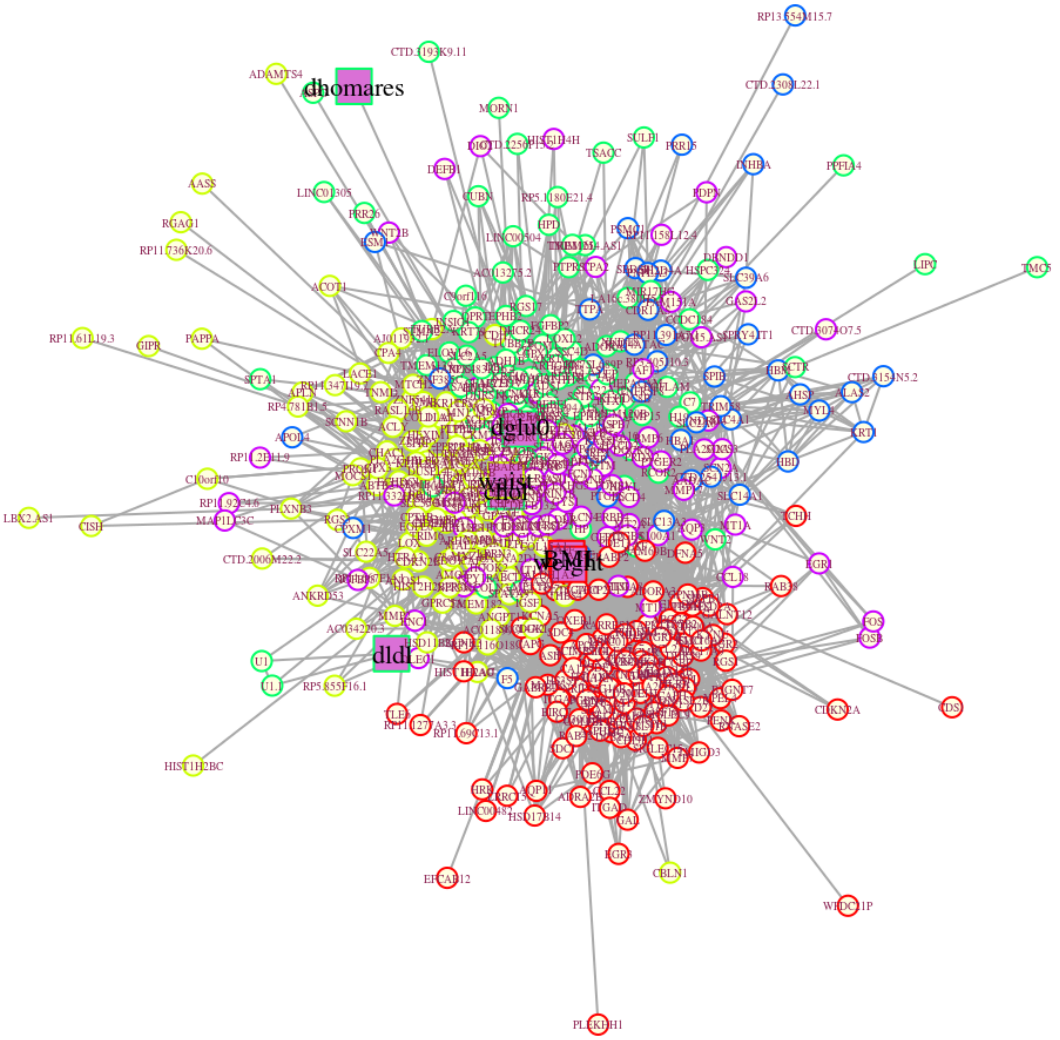


Figure 6.12 **Réseau obtenu pour le contraste CID1/CID2.** La couleur du contour du sommet correspond au module : les sommets avec la même couleur appartiennent aux mêmes modules. Les gènes sont représentés par des sommets ronds de couleur jaune clair et les variables cliniques par des carrés violets.

Figure 6.13 **Réseau obtenu pour le contraste CID2/CID3**. La couleur du contour du sommet correspond au module : les sommets avec la même couleur appartiennent aux mêmes modules. Les gènes sont représentés par des sommets ronds de couleur jaune clair et les variables cliniques par des carrés violets.

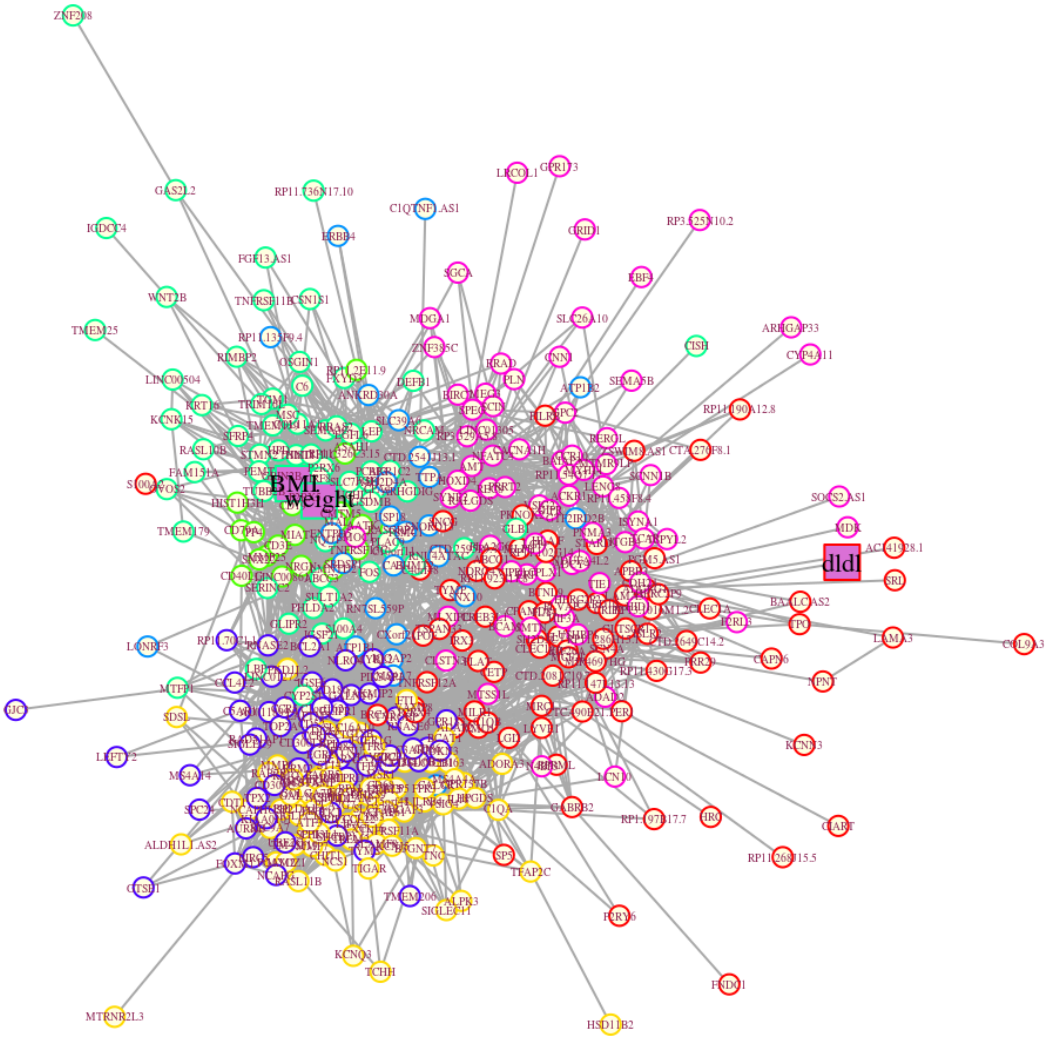


Figure 6.14 **Réseau obtenu pour le contraste CID1/CID3**. La couleur du contour du sommet correspond au module : les sommets avec la même couleur appartiennent aux mêmes modules. Les gènes sont représentés par des sommets ronds de couleur jaune clair et les variables cliniques par des carrés violets.

Chapitre 7

Conclusion et perspectives

Cette thèse regroupe des contributions méthodologiques utiles à l'analyse de données transcriptomiques, en particulier des données d'expression discrètes obtenues via des méthodes de séquençage à haut débit (RNA-Seq et QuantSeq). Nous avons concentré notre attention sur la gestion des données manquantes, l'inférence de réseau et l'analyse intégrative de données de différents types (données cliniques et données transcriptomiques).

Le problème des données manquantes est intimement lié à l'analyse statistique, au fait de collecter et préparer les données pour l'analyse statistique. Dans un premier temps, nous avons donc écrit une revue qui offre une vue d'ensemble des grandes familles de méthodes pouvant gérer les données manquantes ainsi que des recommandations d'utilisation de ces méthodes. Nous avons renseigné les solutions logicielles disponibles, notamment en précisant les packages R où sont implémentés les méthodes décrites. Cette liste n'est évidemment pas exhaustive mais fournit un panorama réaliste des solutions actuellement disponibles.

Cependant, choisir la méthode adéquate pour faire face à la problématique rencontrée et/ou aux types de données manquantes peut s'avérer difficile. Il est donc important de poursuivre la recherche dans ce domaine afin de proposer des méthodes adaptées aux différentes problématiques et aux données manquantes (en fonction du type des données, de leur typologie, du type de mécanisme qui a mené à leur présence, etc.).

Dans un second temps, nous nous sommes intéressées à l'inférence de réseau à partir de données RNA-Seq pour des données longitudinales. La technologie RNA-Seq étant relativement coûteuse, le nombre d'échantillons reste très faible comparé au nombre de variables (gènes). L'inférence de réseau reste donc irréalisable pour la plupart des expérimentations RNA-Seq. Nous avons donc proposé une méthode d'imputation permettant d'obtenir des réseaux de gènes plus robustes à partir de données d'expression RNA-Seq. Cette approche permet d'augmenter artificiellement le nombre d'individus en utilisant de l'information externe mesurée en même temps que les données RNA-Seq. Elle s'avère très utile pour augmenter le nombre d'individus commun entre deux pas de temps afin de comparer les réseaux de gènes obtenus à chaque pas de temps.

Cependant, le modèle graphique log-linéaire de Poisson présente des inconvénients. En effet, [26] a montré que les modèles graphiques de Poisson ne peuvent que capturer les dépendances négatives pour assurer la cohérence de la distribution jointe. Le modèle graphique [6] utilisé est un modèle local et ne permet pas d'avoir un modèle graphique joint cohérent global. De nouveaux modèles graphiques ont récemment été développés. La plupart de ces modèles sont basés sur un modèle log-normal de Poisson [51, 49]. Une perspective d'amélioration de notre méthode serait de remplacer le modèle graphique log-linéaire de Poisson par une de ces approches.

La méthode hd-MI a été mise au point pour des données d'expression discrètes. Il est néanmoins facile de l'adapter pour des données continues en utilisant par exemple des modèles graphiques gaussiens au lieu de modèles graphiques basés sur des distributions de Poisson. Une autre perspective serait d'utiliser l'imputation multiple hot-deck pour d'autres problématiques où le nombre d'observations est faible et où des données auxiliaires sont disponibles. Selon la problématique rencontrée, il faudra changer les dernières étapes de l'imputation multiple, c'est-a-dire la méthode statistique pour analyser chaque jeu imputé et la façon de combiner les résultats. A titre d'exemple, [226] utilisent une autre version de l'imputation multiple hot-deck (hot-deck par classes) afin d'effectuer des analyses factorielles multiples (AFM). Pour combiner les résultats en une configuration compromis, ils utilisent STATIS.

Dans la dernière partie, nous avons commencé par analyser des données d'expression discrètes mesurées avec une nouvelle technique de séquençage à haut débit, le QuantSeq. Cette technologie a l'avantage d'être moins onéreuse que la technique RNA-Seq permettant d'augmenter le nombre d'échantillons séquencés. Elle est également plus rapide et est plus tolérante à la mauvaise qualité de l'ARN. Cette méthode a donc de fortes chances d'être de plus en plus utilisée dans les années à venir pour la recherche de gènes différentiellement exprimés. Comme les données RNA-Seq, ces données sont des données de comptage surdispersées. Il semble donc logique d'utiliser les modèles développées dans le cadre de l'analyse des données RNA-Seq.

Nous avons ensuite cherché à étudier les interactions entre différents ensembles de données et voir leur évolution au cours des différentes phases du protocole DiOGenes. Nous avons donc proposé une méthode intégrative. Pour cela, nous avons utilisé deux types de données (les données transcriptomiques mesurées par QuantSeq et des données biocliniques) afin d'étudier le système biologique dans son ensemble. Afin de faciliter l'analyse globale du système, nous avons choisi une approche basée sur des modèles graphiques, permettant une visualisation de l'ensemble des relations sous la forme d'un réseau. Pour résoudre le problème de l'intensité des liens entre différents types de données, nous ajustons des modèles mixtes pour estimer les relations entre différents types de variables. Un avantage de cette approche est de pouvoir ajuster le modèle avec des variables apportant de la variabilité entre les individus.

Néanmoins, cette approche présente des limites : seuls les individus appariés ont été utilisés, réduisant de façon conséquente le nombre d'observations pour l'inférence de réseau. Il pourrait être intéressant ici d'utiliser l'imputation multiple hot-deck afin d'augmenter le nombre d'individus en commun entre les différents pas de temps. Une autre voie aurait été d'utiliser un modèle graphique joint pour explicitement modéliser la ressemblance entre les différents pas de temps, pour prendre en compte l'appariemment de certains individus et pour inclure de manière explicite l'information sur les individus manquants à certains pas de temps. Des approches de ce type ont déjà été proposées dans le cas gaussien [95, 47, 62, 223, 237]. En effet, généralement, les données d'expression sont collectées dans diverses conditions expérimentales et les réseaux obtenus pour chaque condition sont supposés partager des caractéristiques communes et d'autres différentes. Le principe est donc d'utiliser l'ensemble des observations disponibles pour inférer un réseau dans chaque condition en utilisant un modèle graphique joint pour modéliser explicitement la ressemblance entre les réseaux. Il serait donc intéressant de pouvoir adapter ces modèles pour tenir compte de l'appariemment et des données manquantes, d'une part, et pour l'utilisation avec des

données d'expression discrètes en proposant des modèles graphiques joints basés sur des distributions de Poisson.

Deux ensembles de variables ont été utilisés dans l'analyse intégrative. Il pourrait être intéressant de voir comment construire le modèle en présence de trois types de données, notamment pour sélectionner les arêtes dans la dernière étape afin de ne pas obtenir des réseaux trop denses. [152] ont proposé une approche intégrative basée sur de l'inférence de réseau afin d'étudier les relations entre trois types de données (données d'expression, données de lipidomique et variables bio-cliniques). Les relations entre les variables issues de deux ensembles de données différents sont établies à l'aide d'analyses canoniques régularisées. Une approche intégrative est ensuite utilisée pour fusionner les différents types de relations au sein d'un même réseau. Cette approche ne permet pas d'ajuster le modèle en fonction de variables connues pour apporter de la variabilité entre les individus (par exemple le sexe). Ils ont donc opté pour appliquer la méthode sur des sous-échantillons (groupes constitués uniquement de femmes). Comme pour notre méthode, cela restreint énormément le nombre d'observations utilisées dans l'inférence.

Enfin, les données DiOgenes sont riches et encore incomplètement exploitées. En particulier, ce protocole est peut-être l'unique protocole existant dans lequel il existe des données d'expression obtenues sur un grand nombre d'individus avec quatre technologies différentes (RT-qPCR, puces à ADN, RNA-Seq, QuantSeq). Il serait donc intéressant d'étudier l'impact de la technologie sur l'analyse différentielle et l'inférence de réseau. Une perspective de cette analyse comparative pourrait être l'utilisation de ces données pour comparer les différents modèles graphiques sur la base d'un critère de qualité qui serait la cohérence entre les réseaux inférés sur chacun des types de données.

Bibliographie

[1]K. Abayomi, A. Gelman et M. Levy. “Diagnostics for multivariate imputations”. In : *Journal of the Royal Statistical Society, Series C (Applied Statistics)* 57.3 (2008), p. 273–291 (cf. p. 95).

[2]J. Aitchison et C. H. Ho. “The multivariate Poisson-log normal distribution”. In : *Biometrika* 76.4 (1989), p. 643–653 (cf. p. 50).

[3]H. Akaike. “Information theory and an extension of the maximum likelihood principle”. In : *Second International Symposium on Information Theory*. Sous la dir. de B N Petrov et F Csaki. Budapest : Akadémiai Kiado, 1973, p. 267–281 (cf. p. 51).

[4]P.S. Albert et D.A. Follmann. “Modeling repeated count data subject to informative dropout”. In : *Biometrics* 56.3 (2000), p. 667–677 (cf. p. 105).

[5]K.G. Alberti, R.H. Eckel et S.M. Grundy. “Harmonizing the metabolic syndrome : a joint interim statement of the International Diabetes Federation Task Force on Epidemiology and Prevention National Heart, Lung, and Blood Institute”. In : *Circulation* 120.16 (2009), p. 1640–1645 (cf. p. 30).

[6]G. Allen et Z. Liu. “A log-linear graphical model for inferring genetic networks from high-throughput sequencing data”. In : *Proceedings of IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. 2012 (cf. p. 38, 49, 50, 57, 113, 117, 151).

[7]P.D. Allison. *Missing Data*. Quantitative Applications in the Social Sciences. Thousand Oaks, CA, USA : Sage Publications, 2001 (cf. p. 63, 73).

[8]D. Allouche, C. Cierco-Ayrolles, S. de Givry et al. “A panel of learning methods for the reconstruction of gene regulatory networks in a systems genetics context”. In : *Verification of Methods for Gene Network Inference from Systems Genetics Data*. Sous la dir. d’A. de la Fuente. Springer, 2013 (cf. p. 117).

[9]C. Ambroise, J. Chiquet et C. Matias. “Inferring sparse gaussian graphical models with latent structure”. In : *Electronic Journal of Statistics* 3.21 (2009), p. 205–238 (cf. p. 48).

[10]S Anders et W Huber. “Differential expression analysis for sequence count data”. In : *Genome Biology* 11.10 (2010), R106 (cf. p. 34, 36, 41).

[11]R. Andridge et R.J.A. Little. “A review of hot deck imputation for survey non-response”. In : *International Statistical Review* 78.1 (2010), p. 40–64 (cf. p. 84, 86, 115).

[12]C. Armenise, G. Lefebvre, J. Carayol et al. “Transcriptome profiling from adipose tissue during a low-calorie diet reveals predictors of weight and glycemic outcomes in obese, nondiabetic subjects”. In : *The American Journal of Clinical Nutrition* 106.3 (2017), p. 736–746 (cf. p. 129).

[13]V. Audigier, F. Husson et J. Josse. “A principal component method to impute missing values for mixed data”. In : *Advances in Data Analysis and Classification* 10.1 (2016), p. 5–26 (cf. p. 92).

[14]V. Audigier, F. Husson et J. Josse. “MIMCA : multiple imputation for categorical variables with multiple correspondence analysis”. In : *Statistics and Computing* 27.2 (2016), p. 1–18. arXiv : 1505.08116 (cf. p. 92, 99).

[15]V. Audigier, F. Husson et J. Josse. “Multiple imputation for continuous variables using a Bayesian principal component analysis”. In : *Journal of Statistical Computation and Simulation* 86.11 (2015), p. 2140–2156 (cf. p. 92, 99).

[16]P. L. Auer et R. W. Doerge. “A two-stage poisson model for testing RNA-Seq data”. In : *Statistical Applications in Genetics and Molecular Biology* 10.1 (2011), p. 1–26 (cf. p. 34).

[17]S. Ballouz, W. Verleyen et J. Gillis. “Guidance for RNA-seq co-expression network construction and analysis : safety in numbers”. In : *Bioinformatics* 31.13 (2015), p. 2123–2130 (cf. p. 117).

[18]O. Banerjee, L. El Ghaoui et A. d’Aspremont. “Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data”. In : *Journal of Machine Learning Resaearch* 9 (2008), p. 485–516 (cf. p. 48).

[19]A. Bar-Hen et J.M. Poggi. “Influence measures and stability for graphical models”. In : *Journal of Multivariate Analysis* 147 (2016), p. 145–154 (cf. p. 58, 114).

[20]A.N. Baraldi et C.K. Enders. “An introduction to modern missing data analysis”. In : *Journal of School Psychology* 48.1 (2010), p. 5–37 (cf. p. 63).

[21]L. Baretta et A. Santaniello. “Nearest neighbor imputation algorithms : a critical evaluation”. In : *BMC Medical Informatics and Decision Making*. Proceedings of the 5th Translational Bioinformatics Conference (TBC 2015) : medical informatics and decision making 16.Supp. 3 (2016), p. 74 (cf. p. 83, 86).

[22]V. Barquissau, D. Beuzelin, D.F. Pisani et al. “White-to-brite conversion in human adipocytes promotes metabolic reprogramming towards fatty acid anabolic and catabolic pathways”. In : *Molecular Metabolism* 5.5 (2016), p. 352–365 (cf. p. 119).

[23]T. M. Beasley, S. Erickson et D. B. Allison. “Rank-based inverse normal transformations are increasingly used, but are they merited ?” In : *Behavior Genetics* 39.5 (2009), p. 580 (cf. p. 36).

[24]Y. Benjamini et Y. Hochberg. “Controlling the false discovery rate : a practical and powerful approach to multiple testing”. In : *Journal of the Royal Statistical Society Series B* 57.1 (1995), p. 289–300. arXiv : 95/57289 [0035-9246] (cf. p. 45, 137–139).

[25]M. Bersanelli, E. Mosca, D. Remondini et al. “Methods for the integration of multi-omics data : mathematical aspects”. In : *BMC Bioinformatics* 17.2 (2016) (cf. p. 134).

[26]J. Besag. “Spatial Interaction and the Statistical Analysis of Lattice Systems Spatial Interaction and the Statistical Analysis of Lattice Systems”. In : *Journal of the Royal Statistical Society. Series B (Methodological)* 36.2 (1974), p. 192–236 (cf. p. 49, 151).

[27]G. Blom. *Statistical estimates and transformed beta-variables*. Sous la dir. de New York. Wiley, 1958 (cf. p. 36).

[28]G.E.P. Box et D. Cox. “An analysis of transformations”. In : *Journal of the Royal Statistical Society Series B* 26.2 (1964), p. 211–252 (cf. p. 35).

[29]S. Boyd, N. Parikh, E. Chu, B. Peleato et J. Eckstein. “Distributed optimization and statistical learning via the alternating direction method of multipliers”. In : *Foundations and Trends in Machine Learning* 3.1 (jan. 2011), p. 1–122 (cf. p. 50).

[30]L. Breiman. “Bagging predictors”. In : *Machine Learning* 24.2 (1996), p. 123–140 (cf. p. 52).

[31]L. Breiman. “Random Forests”. In : *Machine Learning* 45.1 (2001), p. 5–32 (cf. p. 87).

[32]L. Breiman, J. Friedman, R. Olsen et C. Stone. *Classification and Regression Trees*. Boca Raton, Florida, USA : Chapman et Hall, 1984 (cf. p. 73–75, 87).

[33]J. H. Bullard, E. Purdom, K. D. Hansen et S. Dudoit. “Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments”. In : *BMC Bioinformatics* 11.1 (2010), p. 94 (cf. p. 40).

[34]R.M. Burns. “Multiple and replicate item imputation in a complex sample survey”. In : *Proceedings of the 6th Annual Research Conference*. Sous la dir. de Bureau of the Census. Washington DC, USA, 1990, p. 655–665 (cf. p. 99).

[35]A. Butte et I. Kohane. “Mutual information relevance networks : functional genomic clustering using pairwise entropy measurements”. In : *Proceedings of the Pacific Symposium on Biocomputing*. T. 426. 2000, p. 418–429 (cf. p. 45).

[36]A. J. Butte et I. S. Kohane. “Unsupervised knowledge discovery in medical databases using relevance networks.” In : *Proceedings of the AMIA Symposium*. 1999, p. 711–715 (cf. p. 45).

[37]S. van Buuren. *Flexible Imputation of Missing Data*. Leiden, The Netherlands : Chapman et Hall/CRC, 2012 (cf. p. 63, 71, 74).

[38]S. van Buuren. “Multiple imputation of discrete and continuous data by fully conditional specification”. In : *Statistical Methods in Medical Research* 16 (2007), p. 219–242 (cf. p. 88).

[39]S. van Buuren et K. Groothuis-Oudshoorn. “MICE : multivariate imputation by chained equations in R”. In : *Journal of Statistical Software* 45 (2011), p. 3. arXiv : NIHMS150003 (cf. p. 88, 99, 109).

[40]E.J. Candès, C.A. Sing-Long et J.D. Trzasko. “Unbiased risk estimates for singular value thresholding and spectral estimators”. In : *IEEE Transactions on Signal Processing* 61.19 (2013), p. 4643–4657 (cf. p. 91).

[41]Y. Cao, X. Jiang, H. Ma et al. “SIRT1 and insulin resistance”. In : *Journal of Diabetes and its Complications* 30.1 (2016), p. 178–183 (cf. p. 129).

[42]J. Carpenter et M. Kenward. *Multiple Imputation and its Application*. Wiley, 2013 (cf. p. 63).

[43]H. Caussinus. “Models and uses of principal component analysis (with discussion)”. In : *Multidimensional Data Analysis. Proceedings of a Workshop, Pembroke College, Cambridge University, England*. Sous la dir. de J. de Leeuw, W.J. Heiser, J.J. Meulman et F. Critchley. Leiden, The Netherlands : DSWO Press, 1986, p. 149–178 (cf. p. 91).

[44]J. Chen et Z. Chen. “Extended Bayesian information criteria for model selection with large model spaces”. In : *Biometrika* 95.3 (2008), p. 759–771 (cf. p. 52).

[45]J. Chen et J. Shao. “Nearest neighbor imputation for survey data”. In : *Journal of Official Statistics* 16.2 (2000), p. 113–131 (cf. p. 83).

[46]D. Chessel, A.B. Dufour et J. Thioulouse. “The ade4 package – I : one-table methods”. In : *R News* 4.1 (2004), p. 5–10 (cf. p. 92, 109).

[47]J. Chiquet, Y. Grandvalet et C. Ambroise. “Inferring multiple graphical structures”. In : *Statistics and Computing* 21.4 (2011), p. 537–553. arXiv : 0912.4434 (cf. p. 152).

[48]J. Chiquet, M. Mariadassou et S. Robin. “Variational inference for probabilistic Poisson PCA”. In : *The Annals of Applied Statistics* 11.2 (2017), p. 655–679. arXiv : 1703.06633 (cf. p. 50).

[49]J. Chiquet, M. Mariadassous et S. Robin. “Variational inference for sparse network reconstruction from count data”. submitted to JCGS. 2018 (cf. p. 50, 51, 151).

[50]H. Choi, J. Gim, S. Won et al. “Network analysis for count data with excess zeros”. In : *BMC Genetics* 18.1 (2017) (cf. p. 50).

[51]Y. Choi, M. Coram, J. Peng et H. Tang. “A Poisson log-normal model for constructing gene covariation network using RNA-seq data”. In : *Journal of Computational Biology* 24.7 (2017), p. 721–731 (cf. p. 50, 151).

[52]W.S. Cleveland et S.J. Devlin. “Locally weighted regression : an approach to regression analysis by local fitting”. In : *Journal of the American Statistical Association* 83.403 (1988), p. 596–610 (cf. p. 87).

[53]N. Cloonan et S. M. Grimmond. “Transcriptome content and dynamics at single-nucleotide resolution”. In : *Genome Biology* 9.9 (2008), p. 234 (cf. p. 26).

[54]J. Cohen, P. Cohen, S.G. West et L.S. Aiken. *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*. 2nd. Mahwah, NJ, USA : Lawrence Erlbaum Associates, 1985 (cf. p. 75).

[55]Linda M. COLLINS, Joseph L. SCHAFER et Karn CHI-MING. “A comparison of inclusive and restrictive strategies in modern missing data procedures”. In : *Psychological Methods* 6.4 (2007), p. 330–351 (cf. p. 102).

[56]D. COOK et D.F. SWAYNE. *Interactive and Dynamic Graphics for Data Analysis*. Use R ! New York, NY, USA : Springer-Verlag, 2007 (cf. p. 67).

[57]J. COSTA-SILVA, D. DOMINGUES et F.M. LOPES. “RNA-Seq differential expression analysis : an extended review and a software tool”. In : *PLOS ONE* 12.12 (déc. 2017), p. 1–18 (cf. p. 44).

[58]S.J. CRANMER et J. GILL. “We have to be discrete about this : a non-parametric imputation technique for missing categorical data”. In : *British Journal of Political Science* 43 (2012), p. 425–449 (cf. p. 84–86, 99, 101, 109, 117, 126).

[59]P. CRAVEN et G. WAHBA. “Smoothing noisy data with spline functions - Estimating the correct degree of smoothing by the method of generalized cross-validation”. In : *Numerische Mathematik* 31.4 (1978), p. 377–403 (cf. p. 51).

[60]N.L. CROOKSTON et A.O. FINLEY. “yaImpute : an R package for kNN imputation”. In : *Journal of Statistical Software* 23 (2008), p. 10 (cf. p. 83, 97, 109, 126).

[61]G. CSARDI et T. NEPUSZ. “The igraph software package for complex network research”. In : *InterJournal* Complex Systems (2006) (cf. p. 145).

[62]P. DANAHER, P. WANG et D. M. WITTEN. “The joint graphical lasso for inverse covariance estimation across multiple classes”. In : *Journal of the Royal Statistical Society. Series B : Statistical Methodology* 76.2 (2012), p. 373–397. arXiv : 1111.0324 (cf. p. 152).

[63]L. DANON, A. DIAZ-GUILERA, J. DUCH et A. ARENAS. “Comparing community structure identification”. In : *Journal of Statistical Mechanics* 2005 (2005), P09008 (cf. p. 127).

[64]A. P. DEMPSTER. “Covariance Selection”. In : *Biometrics* 28.1 (1972), p. 157–175 (cf. p. 47).

[65]A.P. DEMPSTER, N.M. LAIRD et D.B. RUBIN. “Maximum likelihood from incomplete data via the EM algorithm”. In : *Journal of the Royal Statistical Society, Series B (Methodological)* 39.1 (1977), p. 1–38 (cf. p. 76, 77).

[66]P. DIGGLE et M.G. KENWARD. “Informative drop-out in longitudinal data analysis”. In : *Journal of the Royal Statistical Society, Series C (Applied Statistics)* 43.1 (1994), p. 49–93 (cf. p. 103).

[67]M. A. DILLIES, A. RAU, J. AUBERT et al. “A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis”. In : *Briefings in Bioinformatics* 14.6 (2013), p. 671–683 (cf. p. 41, 42).

[68]Y. DING et J.S. SIMONOFF. “An investigation of missing data methods for classification trees applied to binary response data”. In : *Journal of Machine Learning Research* 11 (2010), p. 131–170 (cf. p. 75).

[69]Yiran DONG et Chao-Ying Joanne PENG. “Principled missing data methods for researchers”. In : *SpringerPlus* 2 (2013), p. 222 (cf. p. 102).

[70]B. DUBERN et K. CLÉMENT. “Genetic aspects of obesity”. In : *Presse Médicale* 36.11 (2007), p. 1598–1605 (cf. p. 31).

[71]D. EDWARDS. *Introduction to Graphical Modelling*. Springer, New York, 2000 (cf. p. 45).

[72]C.K. ENDERS. “A primer on maximum likelihood algorithms available for use with missing data”. In : *Structural Equation Modeling* 8.1 (2001), p. 128–141 (cf. p. 77).

[73]C.K. ENDERS. *Applied Missing Data Analysis*. Guilford Press, 2010, p. 401 (cf. p. 86, 115).

[74]B. ESCOFIER et J. PAGÈS. “Multiple factor analysis (AFMULT package)”. In : *Computational Statistics and Data Analysis* 18.1 (1994), p. 121–140 (cf. p. 100).

[75]Y. ESCOUFIER. “Le traitement des variables vectorielles”. In : *Biometrics* 29.4 (1973), p. 751–760 (cf. p. 89).

[76]R.E. FAY. “Alternative paradigms for the analysis of imputed survey data”. In : *Journal of the American Statistical Association* 91.434 (1996), p. 490–498 (cf. p. 86).

[77]I.P. FELLEGI et D. HOLT. “A systematic approach to automatic edit and imputation”. In : *Journal of the American Statistical Association* 71.353 (1976), p. 17–35 (cf. p. 93, 106).

[78]Ch. FENG, H. WANG, N. LU et al. “Log-transformation and its implications for data analysis.” In : *Shanghai archives of psychiatry* 26.2 (2014), p. 105–109 (cf. p. 35).

[79]Pier Alda FERRARI, Paola ANNONI, Alessandro BARBIERO et Giancario MANZI. “An imputation method for categorical variables with application to nonlinear principal component analysis”. In : *Computational Statistics & Data Analysis* 55.7 (2011), p. 2410–2420 (cf. p. 93, 109).

[80]G. FILHOULAUD, S. GUILMEAN, R. DENTIN, J. GIRARD et C. POSTIC. “Novel insights into ChREBP regulation and function”. In : *Trends in Endocrinology and Metabolism* 24.5 (2013), p. 257–268 (cf. p. 129).

[81]C. FINKBEINER. “Estimation for the multiple factor model when data are missing”. In : *Psychometrika* 44.4 (1979), p. 409–420 (cf. p. 77).

[82]D. FOLLMANN et M. WU. “An approximate generalized linear model with random effects for informative missing data”. In : *Biometrics* 51.1 (1995), p. 151–168 (cf. p. 105).

[83]R. FOYGEL et M. DRTON. “Extended bayesian information criteria for Gaussian graphical models”. In : *Advances in Neural Information Processing Systems 23*. Sous la dir. de J D LAFFERTY, C K I WILLIAMS, J SHAWE-TAYLOR, R S ZEMEL et A CULOTTA. Curran Associates, Inc., 2010, p. 604–612 (cf. p. 52).

[84]J. FRIEDMAN. “A recursive partitioning decision rule for nonparametric classification”. In : *IEEE Transactions on Computers* C-26.4 (1977), p. 404–408 (cf. p. 73).

[85]J. FRIEDMAN, T. HASTIE et R. TIBSHIRANI. “Sparse inverse covariance estimation with the graphical lasso”. In : *Biostatistics* 9.3 (2008), p. 432–441 (cf. p. 48, 51, 138).

[86]A.M. GAD et N.M.M. DARWISH. “A shared parameter model for longitudinal data with missing values”. In : *American Journal of Applied Mathematics and Statistics* 1.2 (2013), p. 30–35 (cf. p. 105).

[87]M. GALLOPIN, A. RAU et F. JAFFRÉZIC. “A hierarchical Poisson log-normal model for network inference from RNA sequencing data”. In : *PLoS ONE* 8.10 (2013) (cf. p. 50, 51, 113).

[88]M. GALLOPIN, A. RAU, G. CELEUX et F. JAFFRÉZIC. “Transformation des données et comparaison de modèles pour la classification des données RNA-seq”. In : *47èmes Journées de Statistique de la SFdS*. Lille, France, juin 2015 (cf. p. 37, 39).

[89]A. GELMAN et J. HILL. *Data Analysis Using Regression and Multilevel/Hierachical Models*. New York, NY, USA : Cambridge University Press, 2007 (cf. p. 63).

[90]A. GELMAN, J.B. CARLIN, H.S. STERN et D.B. RUBIN. *Bayesian Data Analysis*. 3rd edition. Boca Raton, FL, USA : Chapman et Hall/CRC, 2013. arXiv : 1001.4656v2 (cf. p. 79).

[91]Ignacio GONZÁLEZ, Kim Anh Lê CAO, Melissa J. DAVIS et Sébastien DÉJEAN. “Visualising associations between paired ’omics’ data sets”. In : *BioData Mining* 5.1 (2012), p. 19 (cf. p. 134).

[92]J.C. GOWER. “A general coefficient of similarity and some of its properties”. In : *Biometrics* 27.4 (1971), p. 857–874. arXiv : arXiv:1011.1669v3 (cf. p. 83).

[93]John W. GRAHAM, Allison E. OLCHOWSKI et Tamika E. GILREATH. “How many imputations are really needed ? Some practical clarifications of multiple imputation theory”. In : *Prevention Science* 8.3 (2007), p. 206–213 (cf. p. 102).

[94]J.W. GRAHAM. “Missing data analysis : making it work in the real world”. In : *Annual Review of Psychology* 60 (2009), p. 549–576 (cf. p. 72, 77).

[95]J. GUO, E. LEVINA, G. MICHAILIDIS et J. ZHU. “Joint estimation of multiple graphical models”. In : *Biometrika* 98.1 (2011), p. 1–15. arXiv : `arXiv:0811.1239` (cf. p. 152).

[96]J. HECKMAN. “Sample selection bias as a specification error”. In : *Econometrica* 47.1 (1979), p. 153–161 (cf. p. 103).

[97]J.J. HECKMAN. “The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models”. In : *Annals of Economic and Social Measurement* 5.4 (1976), p. 475–492 (cf. p. 103).

[98]R.R. HOCKING. “The analysis and selection of variables in linear regression”. In : *Biometrics* 32.1 (1976), p. 1–49 (cf. p. 87).

[99]Arthur E. HOERL et Robert W. KENNARD. “Ridge regression : biased estimation for nonorthogonal problems”. In : *Technometrics* 12.1 (1970), p. 55–67. arXiv : `9809069v1 [arXiv:gr-qc]` (cf. p. 87).

[100]J.W. HOGAN et N.M. LAIRD. “Mixture models for the joint distribution of repeated measures and event times”. In : *Statistics in Medecine* 16.1-3 (1997), p. 239–257 (cf. p. 103).

[101]S. HOLM. “A Simple Sequentially Rejective Multiple Test Procedure”. In : *Scandinavian Journal of Statistics* 6.2 (1979), p. 65–70 (cf. p. 44).

[102]J. HONAKER, G. KING et M. BLACKWELL. “Amelia II : a program for missing data”. In : *Journal of Statistical Software* 45 (2011), p. 7. arXiv : `arXiv:1501.0228` (cf. p. 79, 108, 109).

[103]H-C. HUANG, Y. NIU et L-X. QIN. “Differential expression analysis for RNA-Seq : an overview of statistical methods and computational software.” In : *Cancer Informatics* 14.Suppl 1 (2015), p. 57–67 (cf. p. 44).

[104]P.J. HUBERT et E.M. RONCHETTI. *Robust Statistics*. Wiley Series in Probability and Statistics. Hoboken, NJ, USA : Wiley, 2009 (cf. p. 87).

[105]M. HUISMAN. “Imputation of missing item responses : some simple techniques”. In : *Quality & Quantity* 34.4 (2000), p. 331–351 (cf. p. 83).

[106]A. ILIN et T. RAIKO. “Practical approaches to Principal Component Analysis in the presence of missing values”. In : *Journal of Machine Learning Research* 11 (2010), p. 1957–2000 (cf. p. 89, 91).

[107]A. IMBERT, A. VALSESIA, C. LE GALL et al. “Multiple hot-deck imputation for network inference from RNA sequencing data”. In : *Bioinformatics* 34.10 (2018), p. 1726–1732 (cf. p. 9, 58, 93, 100, 113).

[108]Alyssa IMBERT et Nathalie VILLA-VIALANEIX. *RNAseqNet : Log-Linear Poisson Graphical Model with Hot-Deck Multiple Imputation*. R package version 0.1.2. 2018 (cf. p. 9, 49).

[109]M. JAMSHIDIAN et S. JALAL. “Tests of homoscedasticity, normality, and missing completely at random for incomplete multivariate data”. In : *Psychometrika* 75.4 (2010), p. 649–674. arXiv : `NIHMS150003` (cf. p. 71).

[110]M. JAMSHIDIAN, S. JALAL et C. JANSEN. “MissMech : an R package for testing homoscedasticity, multivariate normality, and missing completely at random (MCAR)”. In : *Journal of Statistical Software* 56.6 (2014), p. 1–31 (cf. p. 71, 107).

[111]D.W. JOENSSEN et U. BANKHOFER. “Donor limited hot deck imputation : effect on parameter estimation”. In : *Journal of Theoretical and Applied Computer Science* 6.3 (2012), p. 58–70 (cf. p. 85).

[112]P. JÖNSSON et C. WOHLIN. “An evaluation of k-nearest neighbour imputation using lIkert data”. In : *Proceedings of the 10th International Symposium on Software Metrics*. (14–16 sept. 2004). Chicago, IL, USA : IEEE, 2004, p. 1530–1435 (cf. p. 83).

[113]J. JOSSE et F. HUSSON. “Handling missing values in exploratory multivariate data analysis methods”. In : *Journal de la Société Française de Statistique* 153.2 (2012), p. 79–99 (cf. p. 92).

[114]J. JOSSE et F. HUSSON. “missMDA : a package for handling missing values in multivariate data analysis”. In : *Journal of Statistical Software* 70.1 (2016), p. 1–31 (cf. p. 138).

[115]J. JOSSE, F. HUSSON et J. PAGÈS. “Gestion des données manquantes en Analyse en Composantes Principales”. In : *Journal de la Société Française de Statistique* 150.2 (2009), p. 28–51 (cf. p. 89, 91).

[116]J. JOSSE, M. CHAVENT, B. LIQUET et F. HUSSON. “Handling missing values with regularized iterative multiple correspondance analysis”. In : *Journal of Classification* 29.1 (2012), p. 91–116 (cf. p. 92, 100, 109).

[117]J. JOSSE, J. PAGÈS et F. HUSSON. “Multiple imputation in principal component analysis”. In : *Advances in Data Analysis and Classification* 5.3 (2011), p. 231–246 (cf. p. 99, 121, 138).

[118]J. KAISER. “Dealing with missing values in data”. In : *Journal of Systems Integration* 5.1 (2014), p. 42–51 (cf. p. 80).

[119]G. KALTON et D. KASPRZYK. “The treatment of missing survey data”. In : *Survey Methodology* 12.1 (1986), p. 1–16 (cf. p. 85).

[120]H.A.L. KIERS. “Weighted least squares fitting using ordinary least squares algorithms”. In : *Psychometrika* 62.2 (1997), p. 251–266 (cf. p. 91).

[121]Y. KIM, S. KWON et H. CHOI. “Consistent Model Selection Criteria on High Dimensions”. In : *Journal of Machine Learning Research* 13 (2012), p. 1037–1057 (cf. p. 51).

[122]R. KOHN et C. F. ANSLEY. “Estimation, prediction, and interpolation for ARIMA models with missing data”. In : *Journal of the American Statistical Association* 81.395 (1986), p. 751–761 (cf. p. 93).

[123]A. KOWARIK et M. TEMPL. “Imputation with the R Package VIM”. In : *Journal of Statistical Software* 74.7 (2016), p. 1–16 (cf. p. 65, 83, 107, 109).

[124]V. N. KRISTENSEN, O. C. LINGJÆRDE, H. G. RUSSNES et al. “Principles and methods of integrative genomic analyses in cancer”. In : *Nature Reviews Cancer* 14.5 (2014), p. 299–313 (cf. p. 134).

[125]T.M. LARSEN, S. DALSKOV, M. van BAAK et al. “The diet, obesity and genes (diogenes) dietary study in eight European countries - A comprehensive design for long-term intervention”. In : *Obesity Reviews* 11.1 (2010), p. 76–91 (cf. p. 31, 118, 134).

[126]S. LAURITZEN. *Graphical models*. T. 17. Oxford University Press, 1996 (cf. p. 45, 47).

[127]C. LAVIT, Y. ESCOUFIER, R. SABATIER et P. TRAISSAC. “The ACT (STATIS method)”. In : *Computational Statistics and Data Analysis* 18.1 (1994), p. 97–119 (cf. p. 100).

[128]C.W. LAW, Y. CHEN, W. SHI et G.K. SMYTH. “Voom : precision weights unlock linear model analysis tools for RNA-seq read counts”. In : *Genome Biology* 15.R29 (2014) (cf. p. 37, 43).

[129]K.A. LÊ CAO, I. GONZÁLEZ et S. DÉJEAN. “*****Omics : an R package to unravel relationships between two omics data sets”. In : *Bioinformatics* 25.21 (2009), p. 2855–2856 (cf. p. 92, 109).

[130]R. J. LITTLE. “A test of missing completely at random for multivariate data with missing values”. In : *Journal of the American Statistical Association* 83.404 (1988), p. 1198–1202 (cf. p. 70).

[131]R.J.A. LITTLE. “Modeling the drop-Out mechanism in repeated-measures studies”. In : *Journal of the American Statistical Association* 90.431 (1995), p. 1112–1121 (cf. p. 103, 105).

[132]R.J.A. LITTLE. “Pattern-mixture models for multivariate incomplete data”. In : *Journal of the American Statistical Association* 88.421 (1993), p. 125–134 (cf. p. 104).

[133]R.J.A. LITTLE et D.B. RUBIN. *Statistical Analysis with Missing Data*. Wiley, 2002, p. 408 (cf. p. 63–65, 68, 77, 85, 88, 89, 98, 99, 102, 115).

[134]Roderick J. a. LITTLE. “Regression with missing x’s : a review”. In : *Journal of the American Statistical Association* 87.420 (1992), p. 1227–1237 (cf. p. 74).

[135]H. LIU, K. ROEBER et L. WASSERMAN. “Stability approach to regularization selection (StARS) for high dimensional graphical models”. In : *Proceedings of Neural Information Processing Systems (NIPS 2010)*. (6–9 déc. 2010). T. 23. 292. Vancouver, Canada, 2010, p. 1432–1440 (cf. p. 49, 51, 53, 58, 118).

[136]Y. LIU, J. ZHOU et K.P. WHITE. “RNA-seq differential expression studies : more sequence or more replication ?” In : *Bioinformatics* 30.3 (2014), p. 301–304 (cf. p. 114).

[137]K.J. LIVAK et T.D. SCHMITTGEN. “Analysis of relative gene expression data using teal-time quantitative PCR and the 2-ΔΔCT method”. In : *Methods* 25.4 (2001), p. 402–408 (cf. p. 119).

[138]J. LONSDALE, J. THOMAS, M. SALVATORE et al. “The genotype-tissue expression (GTEx) project.” In : *Nature Genetics* 45 (2013), p. 580–585 (cf. p. 118).

[139]M. I. LOVE, W. HUBER et S. ANDERS. “Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2”. In : *Genome Biology* 15.12 (2014), p. 550. arXiv : `arXiv : 1303.3997v2` (cf. p. 37, 39, 43).

[140]S. LYSEN. “Permuted Inclusion Criterion : A Variable Selection Technique”. Thèse de doctorat. University of Pennsylvania, 2009 (cf. p. 52, 138).

[141]John MAINDONALD et W. John BRAUN. *Data Analysis and Graphics Using R*. 3e éd. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge, UK : Cambridge University Press, 2010 (cf. p. 35).

[142]J. C. MARIONI, C. E. MASON, S. M. MANE, M. STEPHENS et Y. GILAD. “RNA-seq : An assessment of technical reproducibility and comparison with gene expression arrays”. In : *Genome Research* 18.9 (2008), p. 1509–1517 (cf. p. 40).

[143]N. MEINSHAUSEN et P. BÜHLMANN. “High dimensional graphs and variable selection with the Lasso”. In : *Annals of Statistic* 34.3 (2006), p. 1436–1462 (cf. p. 48, 50, 117).

[144]N. MEINSHAUSEN et P. BÜHLMANN. “Stability selection”. In : *Journal of the Royal Statistical Society. Series B : Statistical Methodology* 72.4 (2010), p. 417–473. arXiv : `0809.2932` (cf. p. 51–53).

[145]M. MELÉ, P.G. FERREIRA, F. REVERTER et al. “The human transcriptome across tissues and individuals”. In : *Science* 348.6235 (2015), p. 660–665 (cf. p. 119).

[146]S.L. MENG et D.B. RUBIN. “Maximum likelihood estimation via the ECM algorithm : a general framework”. In : *Biometrika* 80.2 (1993), p. 267–278 (cf. p. 77).

[147]X.L. MENG et D.B. RUBIN. “Using EM to obtain asymptotic variance-covariance matrices : the SEM algorithm”. In : *Journal of the American Statistical Association* 86.416 (1991), p. 899–909 (cf. p. 101).

[148]M. MOEUR et A.R. STAGE. “Most similar neighbor : an improved sampling inference procedure for natural resources planning”. In : *Forest Science* 42.1 (1995), p. 337–359 (cf. p. 82).

[149]G. MOLENBERGHS, B. MICHIELS, M.G. KENWARD et P.J. DIGGLE. “Monotone missing data and pattern-mixture models”. In : *Statistica Neerlandica* 52.2 (1998), p. 153–161 (cf. p. 104).

[150]P. MOLL, M.l ANTE, A. SEITZ et T. REDA. “QuantSeq 3 ’ mRNA sequencing for RNA quantification”. In : *Nature Methods* 11.November (2014), p. 25 (cf. p. 27, 59, 133, 135).

[151]F.J. MOLNAR, B. HUTTON et D. FERGUSSON. “Does analysis using “last observation carried forward” introduce bias in dementia research ?” In : *Canadian Medical Association Journal* 179.8 (2008), p. 751–753 (cf. p. 81).

[152]E. MONTASTIER, N. VILLA-VIALANEIX, S. CASPAR-BAUGUIL et al. “System model network for adipose tissue signatures related to weight changes in response to calorie restriction and subsequent weight maintenance”. In : *PLoS Computational Biology* 11.1 (2015), e1004047 (cf. p. 113, 129, 134, 138, 153).

[153]S. MORITZ et T. BARTZ-BEIELSTEIN. “imputeTS : time series missing value imputation in R”. In : *The R Journal* 9.1 (2017), p. 207–218 (cf. p. 93, 109).

[154]S. MORITZ, A. SARDÁ, T. BARTZ-BEIELSTEIN, M. ZAEFFERER et J. STORK. “Comparison of different methods for univariate time series imputation in R”. Prepint arXiv 1510.03924. 2015 (cf. p. 94).

[155]A. MORTAZAVI, B. A. WILLIAMS, K. MCCUE, L. SCHAEFFER et B. WOLD. “Mapping and quantifying mammalian transcriptomes by RNA-Seq”. In : *Nature Methods* 5.7 (2008), p. 621–628. arXiv : 1111.6189v1 (cf. p. 40).

[156]K. B. MULLIS et F. A. FALOONA. “Specific Synthesis of DNA in Vitro via a Polymerase-Catalyzed Chain Reaction”. In : *Methods in Enzymology* 155.C (1987), p. 335–350 (cf. p. 23).

[157]M.E.J. NEWMAN et M. GIRVAN. “Finding and evaluating community structure in networks”. In : *Physical Review, E* 69 (2004), p. 026113 (cf. p. 126, 144).

[158]Janelle R. NOEL-MACDONNELL, Joseph USSET, Ellen L. GOODE et Brooke L. FRIDLEY. “Assessment of data transformations for model-based clustering of RNA-Seq data”. In : *PLoS ONE* 13.2 (2018) (cf. p. 36).

[159]A. OSHLACK et M. J. WAKEFIELD. “Transcript length bias in RNA-seq data confounds systems biology”. In : *Biology Direct* 4.1 (2009), p. 14 (cf. p. 41).

[160]E. PEBESMA. “spacetime : spatio-temporal data in R”. In : *Journal of Statistical Software* 51.7 (2012), p. 1–30 (cf. p. 94, 109).

[161]R. A. PETERSON. *bestNormalize : A suite of normalizing transformations*. R package version 3.4.1. 2017 (cf. p. 39).

[162]A.T. PETTERSSON, N. MEJHERT, M. JERNÅS et al. “Twist1 in human white adipose tissue and obesity.” In : *The Journal of Clinical Endocrinology and Metabolism* 96.1 (2011), p. 133–41 (cf. p. 128).

[163]V. PICHENY, J. VANDEL, M. VIGNES et N. VILLA-VIALANEIX. “Reconstruction quality of a biological network when its constituting elements are partially observed”. In : *AI & Statistics*. (22–25 avr. 2014). L014. Reykjavik, Iceland, 2014 (cf. p. 113).

[164]T.D. PIGOTT. “A review of methods for missing data”. In : *Educational Research and Evaluation* 7.4 (2001), p. 353–383 (cf. p. 73).

[165]J. PINHEIRO. “Linear Mixed-Effects Models : Basic Concepts and Examples”. In : *Mixed-Effects Models in Sand S-PLUS*. New York, NY : Springer New York, 2000, p. 3–56 (cf. p. 138).

[166]J. PINHEIRO, D. BATES, S. DEBROY, D. SARKAR et R CORE TEAM. “nlme : Linear and Nonlinear Mixed Effects Models”. In : 3 (jan. 2013). R package version 3.1-137, p. 1–113 (cf. p. 139).

[167]J.N.K. RAO et J. SHAO. “Jackknife variance estimation with survey data under hot deck imputation”. In : *Biometrika* 79.4 (1992), p. 811–822 (cf. p. 99).

[168]A. RAU et C. MAUGIS-RABUSSEAU. “Transformation and model choice for RNA-seq co-expression analysis”. In : *Briefings in Bioinformatics* 19.3 (2018), p. 425–436. arXiv : 1611.06654 (cf. p. 38).

[169]A. RAU, C. MAUGIS-RABUSSEAU, M. L. MARTIN-MAGNIETTE et G. CELEUX. “Co-expression analysis of high-throughput transcriptome sequencing data with Poisson mixture models”. In : *Bioinformatics* 31.9 (2015), p. 1420–1427 (cf. p. 37, 39).

[170]M. Reilly et M. Pepe. “The relationship between hot-deck multiple imputation and weighted likelihood”. In : *Statistics in Medecine* 16.1-3 (1997), p. 5–19 (cf. p. 72).

[171]M. E. Ritchie, D. Phipson B.and Wu, Y. Hu et al. “limma powers differential expression analyses for RNA-sequencing and microarray studies”. In : *Nucleic Acids Research* 43.7 (2015), e47 (cf. p. 39, 43).

[172]J.M. Robins et N. Wang. “Inference for imputation estimators”. In : *Biometrika* 87.1 (2000), p. 113–124 (cf. p. 72).

[173]J.M. Robins, A. Rotnitzky et L.P. Zhao. “Analysis of semiparametric regression models for repeated outcomes in the presence of missing data”. In : *Journal of the American Statistical Association* 90.429 (1995), p. 106–121 (cf. p. 103).

[174]M. Robinson, D. McCarthy et G. Smyth. “edgeR : a Bioconductor package for differential expression analysis of digital gene expression data”. In : *Bioinformatics* 26.1 (2010), p. 139–140 (cf. p. 137).

[175]M. D. Robinson et A. Oshlack. “A scaling normalization method for differential expression analysis of RNA-seq data”. In : *Genome Biology* 11.3 (2010), R25 (cf. p. 34, 41, 43, 136).

[176]M. D. Robinson et G. K. Smyth. “Small-sample estimation of negative binomial dispersion, with applications to SAGE data”. In : *Biostatistics* 9.2 (2008), p. 321–332 (cf. p. 43).

[177]F. Rohart, B. Gautier, A. Singh et K.-A. Lê Cao. “mixOmics : An R package for ‘omics feature selection and multiple data integration”. In : *PLOS Computational Biology* 13.11 (2017), e1005752. arXiv : `Rohart,Florian,2017,mixOmics` (cf. p. 134).

[178]Y. Rosseel. “lavaan : an R package for structural equation modeling”. In : *Journal of Statistical Software* 48.2 (2012) (cf. p. 79, 108).

[179]A. Rotnitzky, J.M. Robins et D.O. Scharfstein. “Semiparametric regression for repeated outcomes with nonignorable nonresponse”. In : *Journal of the American Statistical Association* 93.444 (1998), p. 1321–1339 (cf. p. 103).

[180]D.B. Rubin. “Formalizing subjective notions about the effect of nonrespondents in sample surveys”. In : *Journal of the American Statistical Association* 72.359 (1977), p. 538–543 (cf. p. 103).

[181]D.B. Rubin. “Inference and missing data”. In : *Biometrika* 63.3 (1976), p. 581–592 (cf. p. 71).

[182]D.B. Rubin. “Multiple imputation after 18+ years”. In : *Journal of the American Statistical Association* 91.434 (2012), p. 473–489 (cf. p. 97).

[183]D.B. Rubin. *Multlipe Imputation for Nonresponse in Surveys*. Wiley, 1987 (cf. p. 97–99).

[184]P.C. Sabeti, P. Varilly, B. Fry et al. “Genome-wide detection and characterization of positive selection in human populations”. In : *Nature* 449.7164 (2007), p. 913–918 (cf. p. 129).

[185]J. Schäfer et K. Strimmer. “A shrinkage approach to large-scale covariance matrix estimation and implication for functional genomics”. In : *Statistical Applications in Genetics and Molecular Biology* 4 (2005), p. 1–32 (cf. p. 47).

[186]J. Schäfer et K. Strimmer. “An empirical Bayes approach to inferring large-scale gene association networks”. In : *Bioinformatics* 21.6 (2005), p. 754–764 (cf. p. 47).

[187]J.L. Schafer. *Analysis of Incomplete Multivariate Data*. CRC Monographs on Statistics & Applied Probability. Boca Raton, FL, USA : Chapman et Hall/CRC, 1997 (cf. p. 63, 76).

[188]J.L. Schafer. “Multiple imputation : a primer”. In : *Statistical Methods in Medical Research* 8.1 (1999), p. 3–15 (cf. p. 97).

[189]J.L. Schafer et J.W. Graham. “Missing data : our view of the state of the art”. In : *Psychological Methods* 7.2 (2002), p. 147–177 (cf. p. 63, 80, 81, 86, 102).

[190]J.L. SCHAFER et M.K. OLSEN. “Multiple Imputation for multivariate missing-data problems : a data analyst’s perspective”. In : *Multivariate Behavioral Research* 33.4 (1998), p. 545–571 (cf. p. 79, 108).

[191]G. SCHWARZ. “Estimating the dimension of a model”. In : *The Annals of Statistics* 6.2 (1978), p. 461–464. arXiv : arXiv:1011.1669v3 (cf. p. 51).

[192]S.R. SEAMAN et I.R. WHITE. “Review of inverse probability weighting for dealing with missing data”. In : *Statistical Methods in Medical Research* 22.3 (2011), p. 278–295 (cf. p. 72).

[193]G.A. SIMON et J.S. SIMONOFF. “Diagnostic plots for missing data in least squares regression”. In : *Journal of the American Statistical Association* 81.394 (1986), p. 501–509 (cf. p. 95).

[194]S. A. SIMPSON, C. SHAW et R. MCNAMARA. “What is the most effective way to maintain weight loss in adults ?” In : *BMJ* 343.dec281 (2011), p. d8042 (cf. p. 133).

[195]R. de SMET et K. MARCHAL. “Advantages and limitations of current network inference methods”. In : *Nature Reviews Microbiology* 8 (2010), p. 717–729 (cf. p. 126).

[196]R. R. SOKAL et F. J. ROHLF. *Biometry : the principles and practice of statistics in biological research*. Sous la dir. de New YORK. T. 3. W.H. Freeman, 1969. arXiv : arXiv:1011.1669v3 (cf. p. 35).

[197]E. M. SOUTHERN. “An improved method for transferring nucleotides from electrophoresis strips to thin layers of ion-exchange cellulose”. In : *Analytical Biochemistry* 62.1 (1974), p. 317–318 (cf. p. 25).

[198]K. SPIEGEL, E. TASALI, R. LEPROULT et E. VAN CAUTER. “Effects of poor and short sleep on glucose metabolism and obesity risk”. In : *Nature Reviews Endocrinology* 5.5 (2009), p. 253–261. arXiv : 15334406 (cf. p. 31).

[199]A.P. ST-PIERRE, V. SHIKON et D. C. SCHNEIDER. “Count data in biology—Data transformation or model reformation ?” In : *Ecology and Evolution* 8.6 (2018), p. 3077–3085 (cf. p. 35).

[200]W. STACKLIES, H. REDESTIG, M. SCHOLZ, D. WALTHER et J. SELBIG. “pcaMethods – a bioconductor package providing PCA methods for incomplete data”. In : *Bioconductor* 23.9 (2007), p. 1164–1167 (cf. p. 92, 109).

[201]A.R. STAGE et N.L. CROOKSTON. “Partitioning error components for accuracy-assessment of near-neighbor methods of imputation”. In : *Forest Science* 53.1 (2007), p. 62–72 (cf. p. 96, 97, 102).

[202]D.J. STEKHOVEN et P. BÜHLMANN. “Missforest-non-parametric missing value imputation for mixed-type data”. In : *Bioinformatics* 28.1 (2012), p. 112–118. arXiv : 1105.0828 (cf. p. 87, 109).

[203]M STONE. “Cross-Validatory choice and assessment of statistical predictions”. In : *Journal of the Royal Statistical Society* 36.2 (1974), p. 111–147 (cf. p. 51).

[204]E.A. STUART, M. AZUR, C. FRANGAKIS et P. LEAF. “Multiple imputation with large data sets : a case study of the children’s mental health initiative”. In : *American Journal of Epidemiology* 169.9 (2009), p. 1133–1139 (cf. p. 95).

[205]Y.S. SU, A. GELMAN, J. HILL et M. YAJIMA. “Multiple imputation with diagnostics (mi) in R : opening windows into the black box”. In : *Journal of Statistical Software* 45 (2011), p. 2 (cf. p. 65, 99, 107, 109).

[206]N. SVANVIK, A. STÅHLBERG, U. SEHLSTEDT, R. SJÖBACK et M. KUBISTA. “Detection of PCR Products in Real Time Using Light-up Probes”. In : *Analytical Biochemistry* 287.1 (2000), p. 179–182 (cf. p. 24).

[207]M.A. TANNER et W. WONG. “The calculation of posterior distributions by data augmentation”. In : *Journal of the American Statistical Association* 82.398 (1987), p. 528–540 (cf. p. 78, 79).

[208]M. TEMPL, A. ALFONS et P. FILZMOSER. “Exploring Incomplete data using visualization techniques”. In : *Advances in Data Analysis and Classification* 6.1 (2012), p. 29–47 (cf. p. 65, 67, 107, 109).

[209]M. TENENHAUS. *La régression PLS Théorie et pratique.* 1998 (cf. p. 90).

[210]H. THIJS, G. MOLENBERGHS, B. MICHIELS, G. VERBEKE et D. CURRAN. “Strategies to fit pattern-mixture models”. In : *Biostatistics* 3.2 (2002), p. 245–265 (cf. p. 104, 106).

[211]R. TIBSHIRANI. “Regression shrinkage and selection via the Lasso”. In : *Journal of the Royal Statistical Society. Series B (Methodological)* 58.1 (1996), p. 267–288 (cf. p. 48, 87).

[212]N.J. TIERNEY, F.A. HARDEN, M.J.. HARDEN et K.L. MENGERSEN. “Using decision trees to understand structure in missing data”. In : *BMJ Open* 5.6 (2015), e007450 (cf. p. 65, 71).

[213]M.E. TIPPING et C.M. BISHOP. “Probabilistic principal component analysis”. In : *Journal of the Royal Statistical Association, Series B (Statistical Methodology)* 61 (1999), p. 611–622 (cf. p. 91).

[214]L. TORGO. *Data Mining with R : Learning with Case Studies*. CRC Data Mining and Knowledge Discovery Series. Boca Raton, Florida, USA : Chapman et Hall, 2010 (cf. p. 83, 109).

[215]O. TROYANSKAYA, M. CANTOR, G. SHERLOCK et al. “Missing value estimation methods for DNA microarrays”. In : *Bioinformatics* 17.6 (2001), p. 520–525 (cf. p. 83, 109).

[216]C. TSE et J. CAPEAU. “Quantification des acides nucléiques par PCR quantitative en temps réel”. In : *Annales de Biologie Clinique* 6.3 (2003), p. 279–293 (cf. p. 24).

[217]K. UNNEBRINK et J. WINDELER. “Intention-to-treat : methods for dealing with missing values in clinical trials of progressively deteriorating diseases”. In : *Statistics in Medecine* 20.24 (2001), p. 3931–3946 (cf. p. 81).

[218]W. N. VENABLES et B. D. RIPLEY. *Modern Applied Statistics with S*. Fourth. ISBN 0-387-95457-0. New York : Springer, 2002 (cf. p. 39).

[219]M. VERBANCK, J. JOSSE et F. HUSSON. “Regularised PCA to denoise and visualise data”. In : *Statistics and Computing* 25.2 (2015), p. 471–486 (cf. p. 91).

[220]G. VERBEKE, G. MOLENBERGHS, H. THIJS, E. LESAFFRE et M.G. KENWARD. “Sensitivity analysis for nonrandom dropout : a local influence approach”. In : *Biometrics* 57.1 (2001), p. 7–14 (cf. p. 105, 106).

[221]N. VERZELEN. “Minimax risks for sparse regressions : ultra-high-dimensional phenomenons”. In : *Electronic Journal of Statistics* 6 (2012), p. 38–90 (cf. p. 113, 137).

[222]N. VIGUERIE, E. MONTASTIER, J.J. MAORET et al. “Determinants of human adipose tissue gene expression : impact of diet, sex, metabolic status and cis genetic regulation”. In : *PLoS Genetics* 8.9 (2012), e1002959 (cf. p. 119, 129).

[223]N. VILLA-VIALANEIX, M. VIGNES, N. VIGUERIE et M. S. CRISTOBAL. “Inferring networks from multiple samples with consensus LASSO”. In : *Quality Technology and Quantitative Management* 11.1 (2014), p. 39–60 (cf. p. 152).

[224]N. VILLA-VIALANEIX, L. LIAUBET, T. LAURENT et al. “The structure of a gene co-expression network reveals biological functions underlying eQTLs”. In : *PLoS ONE* 8.4 (2013), e60045 (cf. p. 126).

[225]H.D. VINOD. “Canonical ridge and econometrics of joint production”. In : *Journal of Econometrics* 4.2 (1976), p. 147–166 (cf. p. 126).

[226]V. VOILLET, P. BESSE, L. LIAUBET, M. SAN CRISTOBAL et I. GONZÁLES. “Handling missing rows in multi-omics data integration : multiple imputation in multiple factor analysis framework”. In : *BMC Bioinformatics* 17.402 (2016). Forthcoming (cf. p. 93, 100, 116, 152).

[227]Willem M. van der WAL et Ronald B. GESKUS. “ipw : an R package for inverse probability weighting”. In : *Journal of Statistical Software* 43.13 (2011) (cf. p. 73, 108).

[228]Ying-Wooi WAN, Genevera I. ALLEN, Yulia BAKER et al. “XMRF : an R package to fit Markov Networks to high-throughput genetics data”. In : *BMC Systems Biology* 10.3 (2016), p. 69 (cf. p. 49).

[229]L. WASSERMAN et K. ROEDER. “High-dimensional variable selection”. In : *Annals of Statistics* 37.5A (2009), p. 2178–2201. arXiv : 0704.1139 (cf. p. 51).

[230]R. W.M. WEDDERBURN. “Quasi-likelihood functions, generalized linear models, and the gauss-newton method”. In : *Biometrika* 61.3 (1974), p. 439–447 (cf. p. 34).

[231]J. WHITTAKER. *Graphical Models in Applied Multivariate Statistics*. Sous la dir. de CHISCHESTER. Wiley Publishing, 2009 (cf. p. 45, 47).

[232]R. R. WING et S. PHELAN. “Long-term weight loss maintenance”. In : *The American journal of clinical nutrition* 82 (2005), 222S–225S (cf. p. 133).

[233]D. M. WITTEN. “Classification and clustering of sequencing data using a poisson model”. In : *Annals of Applied Statistics* 5.4 (2011), p. 2493–2518. arXiv : 1202.6201 (cf. p. 38).

[234]Herman WOLD. “Estimation of principal components and related models by iterative least squares”. In : *Multivariate Analysis*. Sous la dir. de KRISHNAIAH. New York, USA : Academic Press, 1966, p. 1391–1420. arXiv : arXiv:1011.1669v3 (cf. p. 90).

[235]H. WU, X. DENG et N. RAMAKRISHNAN. “Sparse estimation of multivariate Poisson log-normal models from count data”. In : *Statistical Analysis and Data Mining* 11.2 (2018), p. 66–77 (cf. p. 51).

[236]M.C. WU et R.J. CARROLL. “Estimation and comparison of changes in the presence of informative right censoring by modeling the censoring process”. In : *Biometrics* 44.1 (1988), p. 175–188 (cf. p. 104).

[237]Y. XIE, Y. LIU et W. VALDAR. “Joint estimation of multiple dependent Gaussian graphical models with applications to mouse genomics”. In : *Biometrika* 103.3 (2016), p. 493–511. arXiv : 1608.08659 (cf. p. 152).

[238]E. YANG, P. RAVIKUMAR, G.I. ALLEN et Z. LIU. “Graphical models via generalized linear models”. In : *Advances in Neural Information Processing Systems*. 2012, p. 1358–1366 (cf. p. 49).

[239]K. Y. YEUNG, C. FRALEY, A. MURUA, A. E. RAFTERY et W. L. RUZZO. “Model-based clustering and data transformations for gene expression data”. In : *Bioinformatics* 2001.17 (2001), p. 10 (cf. p. 37).

[240]M. YUAN et Y. LIN. “Model selection and estimation in the Gaussian graphical model”. In : *Biometrika* 94.1 (2007), p. 19–35 (cf. p. 48).

[241]A. ZEILEIS et G. GROTHENDIECK. “zoo : S3 infrastructure for regular and irregular time series”. In : *Journal of Statistical Software* 14.6 (2005), p. 1–27 (cf. p. 94, 109).

[242]L. ZHANG et B.K. MALLICK. “Inferring gene networks from discrete expression data”. In : *Biostatistics* 14.4 (2013), p. 708–722 (cf. p. 113).

[243]S. ZHANG. “Nearest neighbor selection for iterative kNN imputation”. In : *Journal of Systems and Software* 85.11 (2012), p. 2541–2552 (cf. p. 82).

[244]T. ZHAO, H. LIU, K. ROEDER, J. LAFFERTY et L. WASSERMAN. “The huge package for high-dimensional undirected graph estimation in R.” In : *Journal of machine learning research* 13 (2012), p. 1059–1062. arXiv : 15334406 (cf. p. 48, 138, 139).

[245]H. ZOU et T. HASTIE. “Regularization and variable selection via the elastic net”. In : *Journal of the Royal Statistical Society : Series B (Statistical Methodology)* 67.2 (2005), p. 301–320 (cf. p. 87).

[246]Isabella ZWIENER, Barbara FRISCH et Harald BINDER. “Transforming RNA-Seq data to improve the performance of prognostic gene signatures”. In : *PLoS ONE* 9.1 (2014), e85150. arXiv : NIHMS150003 (cf. p. 34, 36).