UNIVERSITÉ DU QUÉBEC À MONTRÉAL

ANALYSE DE DONNÉES DE GRANDE DIMENSION À L'AIDE DE MÉTHODES D'APPRENTISSAGE STATISTIQUE

MÉMOIRE

PRÉSENTÉ

COMME EXIGENCE PARTIELLE

DE LA MAÎTRISE EN MATHÉMATIQUES

PAR

MARIE-HÉLÈNE LAFOND

AVRIL 2017

REMERCIEMENTS

Je désire d'abord remercier Karim Oualkacha, mon directeur de recherche, pour m'avoir accueilli dans son équipe et pour m'avoir offert ce merveilleux projet. Je tiens à le remercier aussi pour sa passion, sa patience et sa confiance qu'il m'a démontrées tout au long de ma maîtrise et pour le soutien financier qu'il m'a accordé.

Je souhaite également remercier les professeurs qui m'ont enseigné lors de ma maîtrise à l'UQAM. J'ai appris énormément sur divers sujets de la statistique. Un gros merci également aux professeurs et aux chargés de cours qui m'ont offert des postes de démonstration. Je tiens aussi à remercier Gisèle Legault pour son aide au laboratoire informatique.

Je termine en remerciant ma famille pour leur soutien et leurs encouragements.

TABLE DES MATIÈRES

LISTE DES TABLEAUX

LISTE DES FIGURES

RÉSUMÉ

Les données massives sont une partie intégrante des nouvelles recherches. Nous recueillons des milliers de données dans le but de mieux comprendre certaines maladies complexes et la génétique sous-jacente à ces maladies. Lors des analyses statistiques, nos principaux objectifs étaient de vérifier l'association entre les gènes et la maladie (la variable réponse) et de prédire l'état de santé (malade ou non malade) des sujets. Nous avons dû faire face à deux grands défis. Tout d'abord, dans des jeux de données de grande dimension, plusieurs gènes ne sont pas informatifs. En effet, parmi tous les gènes recueillis, plusieurs n'ont pas de lien avec la maladie ni avec les autres gènes présents dans l'étude. Ce qui nous amène au deuxième défi qui est d'extraire l'information utile tout en tenant compte de la structure de dépendance dans les données.

Afin de réduire la dimension du jeu de données, d'extraire l'information utile et de prédire la maladie de façon appropriée, nous avons développé quelques méthodes. Nous avons entre autres combiné des méthodes de régularisation avec des méthodes classiques de classification telles que la régression logistique et l'analyse discriminante. Nous avons également combiné des méthodes non supervisées, par exemple la méthode des K-moyennes et la méthode « Weigthed Gene Co-expression Network Analysis » (WGCNA), avec des méthodes classiques et de régularisation. Par la suite, nous avons comparé toutes ces approches. Nous avons conclu que les approches constituées de la méthode WGCNA sont les plus performantes tant au niveau de l'association que de la prédiction. Nous avons donc analysé un jeu de données réelles sur la leucémie grâce à la méthode basée sur la WGCNA et nous avons conclu que la voie biologique reliée à la protéine p53 a un effet sur le gène RAS.

MOTS-CLÉS : Régression logistique, analyse discriminante, méthodes de régularisation, « Weighted Gene Co-expression Network Analysis », réseaux biologiques

INTRODUCTION

Les données volumineuses sont au coeur des problématiques émergentes en recherche, que ce soit dans le domaine de la biologie, de l'épidémiologie, de la physique ou de la finance. Dans ce mémoire, nous nous intéresserons principalement aux données reliées aux maladies complexes. Comme son nom l'indique, les maladies complexes sont difficiles à comprendre. Les chercheurs tentent donc de collecter un grand nombre de données sur chacun des sujets de l'étude afin de mieux saisir les mécanismes biologiques et génétiques entourant ces maladies complexes. Ainsi, nous nous retrouvons avec une grande quantité d'informations sur chacun des sujets, mais nous avons peu de sujets puisque de telles expériences engendrent des coûts élevés. De plus, grâce à des études antérieures, les chercheurs sont en mesure de dire que les maladies complexes sont causées par l'interaction de plusieurs gènes ou par l'interaction d'un gène et des facteurs environnementaux, et non par un seul gène. Par conséquent, il faut analyser les prédicteurs (les gènes) ensemble pour capter les structures des groupes de gènes qui sont responsables des maladies complexes. Nous ne voulons pas analyser les gènes un par un. Les gènes faisant partie d'un même groupe et interagissant ensemble forment ce qu'on appelle une « voie biologique ».

Dans nos études, nous avons un vecteur $\mathbf{y}$ de dimension $n \times 1$ contenant l'état de santé des sujets (malade ou non malade) et une matrice $\mathbf{X}$ de dimension $n \times p$ contenant les expressions génétiques, où le nombre de colonnes est beaucoup plus grand que le nombre de lignes ($p \gg n$). Les méthodes classiques de classification telles que la régression logistique et l'analyse discriminante linéaire ne peuvent plus être appliquées à ces jeux de données. En effet, lorsque $p \gg n$, la matrice $\mathbf{X}$

devient singulière et la classification devient difficile à cause du grand nombre de prédicteurs. Des méthodes statistiques, dites de régularisation, ont été développées dans le but de remédier à ces problèmes. Elles sont caractérisées par l'ajout d'un terme de pénalité à une fonction de perte. Parmi ces méthodes, nous retrouvons la régression Ridge (Hoerl et Kennard, 1970), la méthode LASSO (Tibshirani, 1996), la méthode groupe LASSO, la méthode Elastic Net (Zou et Hastie, 2005) et l'analyse discriminante pénalisée basée sur la fonction de Fisher (Witten et Tibshirani, 2011). En plus de rendre la matrice $\mathbf{X}^T\hat{\mathbf{W}}\mathbf{X}$ inversible, elles permettent d'éliminer les informations non pertinentes et de réduire le nombre de prédicteurs. Toutefois, les méthodes de régularisation ne résolvent pas tous les problèmes. En effet, elles sont destinées à des analyses de prédiction et non pas à des tests d'association. Nous ne pouvons donc pas trouver de valeur-p pour tester l'effet des gènes sur une maladie.

Une autre méthode destinée aux jeux de données de grande dimension a fait son apparition dans les années 2000 : la « Weighted Gene Co-expression Network Analysis » (WGCNA). Cette approche, développée par, entre autres, Steve Horvath, a été créée dans un cadre de statistique génétique. Son but est de regrouper les gènes en fonction de leur co-expression et d'identifier les ensembles de gènes qui ont un impact sur une maladie (Zhang et Horvath, 2005). Cette technique est basée sur la méthode de classification hiérarchique, sur des notions de topologie et sur la corrélation de Pearson.

Nous nous sommes donc questionnés quant à la performance des méthodes de régularisation et de la méthode WGCNA. Sont-elles aussi puissantes les unes des autres pour tester l'association entre les gènes et la maladie ? Sont-elles toutes aussi efficaces pour faire de la prédiction ? La méthode WGCNA garde-t-elle assez d'information en groupant les gènes sous forme de modules ? Pour répondre à ces questions, puisqu'il n'y a rien dans la littérature par rapport à cela, nous

avons proposé des tests d'association pour la méthode WGCNA, méthode vue comme une technique de réduction de la dimension, et nous avons suggéré des tests d'association pour les méthodes de régularisation. Dans un premier temps, nous avons testé l'effet global des gènes sur la maladie. Pour ce faire, nous avons combiné les méthodes classiques de classification avec la méthode WGCNA et les méthodes de régularisation, car ces dernières n'ont pas de test d'association. Ainsi, nous avons pu comparer les valeurs-p obtenues par chacune des approches à l'aide de simulations faites en R. Dans un deuxième temps, nous avons vérifié le pouvoir prédictif des méthodes de régularisation et de l'approche WGCNA-EN (WGCNA combinée avec l'Elastic Net). Ici, aucune méthode classique de classification n'a été utilisée, car les méthodes de régularisation sont destinées à la prédiction.

Enfin, nous avons analysé un jeu de données réelles. Nous avons tenté de découvrir les gènes qui ont un grand effet sur la leucémie. Nous avons déterminé les gènes qui ont un impact sur le fait d'avoir ou non le gène RAS chez de jeunes enfants atteints de la leucémie. De plus, nous avons utilisé la « Kyoto Encyclopedia of Genes and Genomes » (KEGG) pour détecter les voies biologiques qui sont endommagées par ce type de cancer.

Ce mémoire est organisé de la façon suivante. Dans le chapitre 1, nous étudions le fonctionnement de certaines voies biologiques. Dans le chapitre 2, nous explorons plusieurs méthodes d'apprentissage statistique : la régression logistique, l'analyse discriminante, la régression Ridge, la méthode LASSO, le groupe LASSO, l'Elastic Net, l'analyse discriminante pénalisée basée sur la fonction de Fisher, la méthode hiérarchique et la méthode des K-moyennes. Dans le chapitre 3, nous examinons la méthode « Weighted Gene Co-expression Network Analysis ». Dans le chapitre 4, nous faisons des simulations à l'aide du logiciel R pour comparer les diverses approches que nous avons développées. Enfin, dans le chapitre 5, nous analysons un vrai jeu de données portant sur la leucémie.

CHAPITRE I

BIOLOGIE

Le corps humain est constamment en interaction avec son environnement ce qui fait en sorte qu'il est confronté à une grande quantité de stimuli chaque jour. Un stimulus est un élément qui déclenche des phénomènes à l'intérieur de notre organisme et qui peut être interne ou externe selon de la provenance de la source de perturbation. Par exemple, la perturbation peut être reliée à une situation stressante, à une blessure ou à une infection. Dans tous les cas, les cellules du corps humain réagissent à ces stimuli et tentent de rétablir l'équilibre. Pour y arriver, des molécules présentes à l'intérieur des cellules produisent des actions menant à des changements cellulaires. L'ensemble de ces actions est appelé « voie biologique ». Il existe plusieurs types de voie biologique. Il y a entre autres les voies métabolique, de régulation, de transduction et d'interactions protéine-protéine (Tkačik et Bialek, 2009). Lorsque les voies biologiques interagissent, nous parlerons plutôt d'un réseau biologique.

Il arrive parfois que certaines voies biologiques ne fonctionnent pas comme elles le devraient. Par conséquent, des maladies telles que le cancer font leur apparition et empêchent le retour à l'équilibre. Des biologistes, des médecins et bien d'autres chercheurs se sont donc tournés vers l'analyse des voies biologiques pour mieux comprendre les dysfonctionnements reliés à certaines maladies et pour découvrir

les gènes qui y sont impliqués. Ils ont toutefois été confrontés à un nouveau problème : les données de grande dimension. Ils sont maintenant capables de recueillir des informations sur des milliers de gènes, de protéines, etc., par contre ces données sont difficilement interprétables dues à leur grand nombre. Des approches statistiques ont été développées dans le but de modéliser les voies et les réseaux et de donner un sens biologique à toutes ces informations.

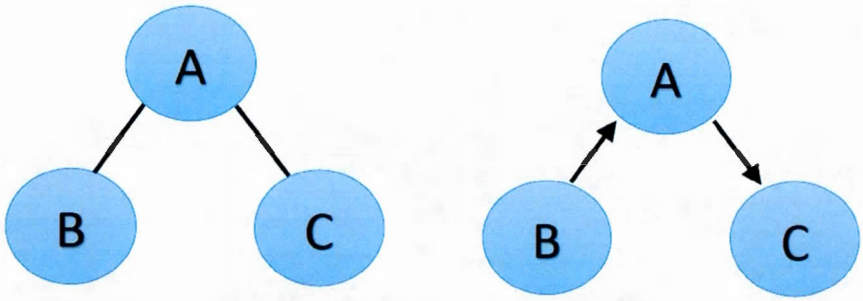
Avant d'entamer l'étude des modèles statistiques, nous verrons dans ce chapitre comment les voies biologiques peuvent être représentées graphiquement et comment les quatre principales voies fonctionnent en temps normal.

1.1 Graphe des voies biologiques

Nous pouvons illustrer les voies et les réseaux graphiquement à l'aide de lignes et de points. Un point, appelé noeud, représente un élément biologique et une ligne, appelée lien, représente l'existence d'une connexion entre deux éléments. Il existe deux types de graphes : orienté et non orienté.

Le graphe non orienté représente une situation dans laquelle les noeuds interagissent de manière symétrique. Supposons que nous avons comme graphe la figure (1.1a) où les points symbolisent des protéines et les lignes sont des interactions entre les protéines. Le lien entre les protéines A et B signifie que la protéine A peut s'attacher à la protéine B et que l'inverse est aussi possible (la protéine B peut s'attacher à la protéine A). Nous pouvons interpréter le lien entre les protéines A et C de la même façon.

Le graphe orienté représente la causalité entre les noeuds. Prenons par exemple la figure (1.1b). Le noeud B a un effet causal direct sur le noeud A qui à son tour a un effet causal direct sur C. Le noeud C dépend de B à travers A.



(a) Graphe non orienté (b) Graphe orienté

Figure 1.1: Deux types de graphes

1.2 Voies biologiques

1.2.1 Voie du métabolisme

Le métabolisme est l'ensemble des réactions biochimiques. Ces réactions peuvent être cataboliques si les molécules complexes sont transformées en molécules simples et s'il y a libération d'énergie ou ces réactions peuvent être anaboliques si les molécules simples sont transformées en molécules complexes et si de l'énergie est nécessaire pour arriver à ces transformations (Campbell et Reece, 2007). La voie métabolique est donc l'ensemble des réactions chimiques engendrées par les substrats, les enzymes et les cofacteurs. Une enzyme est généralement une protéine servant de catalyseur et elle agit sur un réactif appelé substrat. Pour catalyser le substrat, l'enzyme a parfois besoin de substances non protéiques, les cofacteurs (Campbell et Reece, 2007). Un exemple de voie métabolique est le processus par lequel la nourriture est décomposée en molécules d'énergie (National Human Genome Research Institute, 2015).

Dans un graphe d'une voie métabolique, les noeuds peuvent être les substrats et les liens peuvent être les liaisons entre les substrats lorsque ces derniers prennent part

à une même réaction (Tkačik et Bialek, 2009). Le graphe d'une voie métabolique est orienté (Albert, 2005).

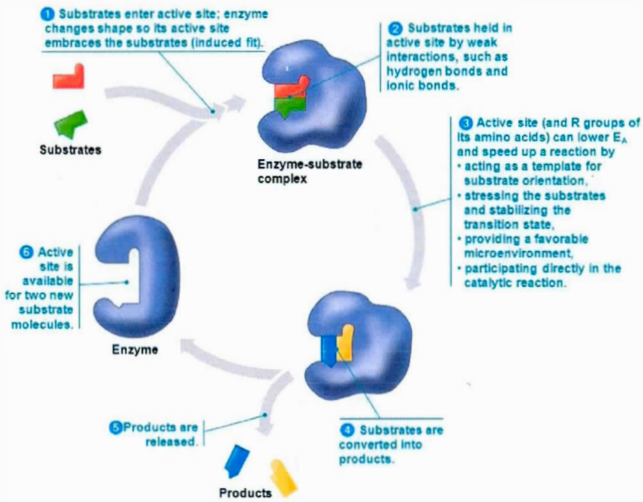


Figure 1.2: Exemple d'une voie métabolique (Campbell et Reece, 2005). Les substrats (molécule rouge et molécule verte) se lient à l'enzyme (protéine mauve) et ils sont maintenus en place par des liaisons faibles comme les liaisons d'hydrogène et ioniques. Par la suite, les substrats sont transformés en produits (molécule bleue et molécule jaune). C'est ce qu'on appelle une réaction catabolique.

1.2.2 Voie de transduction

La transduction d'un stimulus se fait par un ensemble de réactions menant à une réponse intracellulaire suite à un changement environnemental. Un stimulus externe provoque l'arrimage d'une molécule de communication à un récepteur de la cellule. Un récepteur est une protéine sensible à un stimulus particulier. Lorsque la molécule de communication est liée au récepteur, ce dernier active une protéine qui à son tour active une autre protéine et ainsi de suite. C'est ce qu'on appelle l'effet de cascade. L'information se déplace donc au travers des interactions entre protéines. D'autres molécules intermédiaires, appelées seconds messagers, inter-

viennent dans la transduction du stimulus. Ces molécules intermédiaires peuvent amplifier la réponse en transmettant le stimulus à plusieurs autres molécules. Ces réactions en chaîne se produisent jusqu'à ce que la protéine responsable de la réponse soit activée (Campbell et Reece, 2007). Un exemple de voie de transduction est l'ensemble des protéines kinase qui contrôlent plusieurs fonctions, dont celles de la prolifération cellulaire et de la différenciation (Tkačik et Bialek, 2009). Dans ce type de voie biologique, la direction des effets et l'échelle de temps sont importantes.

Dans un graphe de voie de transduction, les noeuds illustrent les protéines et les liens représentent l'interaction entre les protéines.

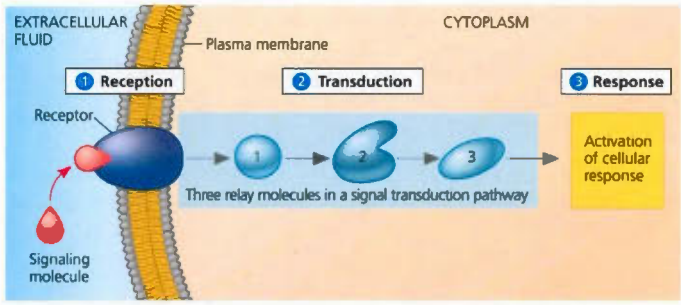


Figure 1.3: Voie de transduction (Campbell et Reece, 2005). La voie de transduction se divise entre trois grandes phases : la réception du stimulus par la membrane plasmique, la transduction du stimulus et la réponse de la cellule. Lors de la première phase, le récepteur (protéine mauve sur le dessin) reçoit une molécule de communication (molécule rouge). Un médiateur chimique détecte cette liaison. La deuxième phase consiste à transformer le stimulus en molécules intermédiaires qui se transforment à leur tour en d'autres molécules (molécules bleues 1, 2 et 3). Il peut y avoir plusieurs modifications successives. C'est ce qu'on appelle voie de transduction. Enfin, la troisième phase est l'activation de la réponse.

1.2.3 Voie d'interactions protéine-protéine

La voie d'interactions protéine-protéine est l'ensemble des interactions entre protéines. Il existe divers types d'interaction, par exemple, une protéine peut transporter de l'oxygène ou une protéine peut se lier à d'autres protéines pour construire ou activer un complexe protéique. Beaucoup de liens entre protéines ont été découverts grâce à diverses technologies (double hybride, spectrométrie de masse), par contre le type d'interaction demeure encore souvent inconnu (Tamassia, 2013). Toutefois, les experts s'entendent pour dire que les protéines présentes dans les voies d'interactions sont souvent impliquées dans les voies de transduction et de régulation.

Nous pouvons représenter une voie d'interactions protéine-protéine par un graphe non orienté dans lequel les protéines sont représentées par des points et les interactions par des lignes (Tkačik et Bialek, 2009; Albert, 2005).

1.2.4 Voie de régulation

La voie de régulation est responsable de la transcription des gènes. En effet, elle peut activer ou arrêter la transcription. Ceci a un grand impact sur la production des protéines, car c'est grâce à la transcription des gènes que les protéines peuvent être synthétisées. Donc, s'il n'y a pas de transcription, il n'y a pas de production de protéine. Pour plus de renseignements sur le fonctionnement de la transcription, veuillez vous référer à l'appendice A.

Dans ce type de voie, les noeuds représentent les gènes et les lignes indiquent l'activation ou la répression d'un gène (Tkačik et Bialek, 2009). Un graphe orienté doit être utilisé pour bien traduire l'enchaînement des actions provoquées par les gènes.

1.3 Lien entre les voies biologiques et la statistique

L'étude des voies biologiques est de plus en plus populaire. Elle nous permet de mieux comprendre le fonctionnement du corps humain et ainsi de trouver des remèdes contre des maladies. Dans ce présent mémoire, nous analyserons des données de leucémie et nous tenterons de déterminer les voies biologiques défectueuses qui sont responsables de la leucémie. Pour y arriver, nous mesurerons l'association entre les milliers de gènes sur lesquels nous détenons de l'information et la maladie. Habituellement, nous utiliserions des méthodes classiques telles que l'analyse discriminante ou la régression logistique pour mesurer l'association, mais ces méthodes ne sont plus adéquates lorsque nous analysons des jeux de données de grande dimension. Pour remédier à ce problème, nous nous sommes tournés vers des méthodes de régularisation. Bien que celles-ci soient applicables à des jeux de données de grande dimension, ces méthodes sont normalement utilisées pour faire de la prédiction et non pour mesurer l'association. Nous allons donc combiner les méthodes de régularisation et les méthodes classiques dans le but de sélectionner les gènes qui ont une influence sur la leucémie et de mesurer l'association. Nous pourrons par la suite trouver les voies biologiques qui jouent un rôle important dans la leucémie.

CHAPITRE II

APPRENTISSAGE STATISTIQUE

L'apprentissage statistique fait référence à l'ensemble des méthodes statistiques utilisées pour modéliser adéquatement une certaine caractéristique en se basant sur un ensemble fini de données indépendantes et identiquement distribuées. Cet apprentissage peut être soit supervisé, soit non supervisé.

L'apprentissage supervisé englobe les méthodes dont le but est de trouver la relation entre la variable réponse et les variables prédictives. En d'autres mots, nous connaissons les couples $(\mathbf{x}_i, y_i)$, $i = 1, \ldots, n$, pour les n individus et nous voulons trouver la fonction f telle que $Y = f(X) + \epsilon$, où ϵ est l'erreur.

L'apprentissage non supervisé quant à lui est composé des méthodes statistiques qui visent à regrouper les sujets en un certain nombre de classes homogènes en ne se basant que sur la structure de la matrice $\mathbf{X} = (\mathbf{x}_1^T, \ldots, \mathbf{x}_n^T)$ où $\mathbf{x}_i^T$ est un vecteur colonne de dimensions $p \times 1$. Elles ne tiennent pas compte de la variable réponse.

Parmi les méthodes d'apprentissage statistique supervisé et non supervisé, certaines sont réservées à des jeux de données de petite dimension $(p < n)$ tandis que d'autres sont destinées aux données de grande dimension $(p > n)$. Dans le cas des données de petite dimension, le nombre de variables explicatives est plus petit que

le nombre d'individus et nous pouvons appliquer les méthodes dites classiques. Par exemple, la régression linéaire multiple peut être utilisée si la variable réponse est continue et des analyses de discrimination (telles la régression logistique et l'analyse discriminante linéaire) peuvent être utilisées si la variable réponse est binaire. Nous verrons plus loin que ces méthodes ne peuvent pas être employées lorsque le nombre de variables explicatives est plus grand que le nombre d'individus. Nous utiliserons plutôt des méthodes de régularisation, c'est-à-dire des modèles de régression linéaire contraints à des pénalités.

Dans les sections 2.1 et 2.2, il sera question des méthodes d'apprentissage supervisé employées lorsque $p < n$ et $p > n$ respectivement. Dans la section 2.3, nous aborderons les méthodes d'apprentissage non supervisé.

2.1 Apprentissage supervisé pour des données de petite dimension

Nous explorerons deux méthodes d'apprentissage supervisé dans cette section : la régression logistique et l'analyse discriminante.

2.1.1 Régression logistique

Le modèle

La régression logistique est un modèle linéaire généralisé utilisé pour décrire la relation entre une variable réponse dichotomique et un ensemble de variables explicatives. Le vecteur $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})$, où $i = 1, \ldots, n$ représente l'individu, peut être composé de p variables qualitatives (nominales ou ordinales) et quantitatives (avec $p < n$). Par exemple, le vecteur $\mathbf{x}_i$ peut être composé de l'âge de l'individu, de son sexe et de sa taille. La variable réponse quant à elle peut être

codée par 0 et 1 et elle représente l'absence ou la présence d'une certaine caractéristique. Par exemple, $Y = 1$ signifie qu'une personne est atteinte du cancer de la prostate et $Y = 0$ signifie qu'elle n'a pas le cancer. Supposons que nous avons un échantillon de n individus et que pour chaque individu nous connaissons les observations $(\mathbf{x}_i, y_i)$, $i = 1, \ldots, n$. Si tous les individus ont des vecteurs $\mathbf{x}_i$ différents, alors $Y_i \sim Bernouilli(\pi_i)$ et nous sommes dans le cas de données non groupées. Au contraire, si plusieurs individus ont des vecteurs $\mathbf{x}_i$ égaux, $Y_i \sim binomiale(m_i, \pi_i)$ avec $i = 1, \ldots, k$ et $k < n$ et nous parlerons de données groupées.

Dans le cas d'une régression logistique, la fonction de lien la plus souvent utilisée est le logit, c'est-à-dire

$$\begin{aligned} g(\pi_i) &= \text{logit}(\pi_i) \\ &= \log\left(\frac{\pi_i}{1-\pi_i}\right) \\ &= \exp\{\beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}\}. \end{aligned}$$

La fonction $g(\pi_i)$ est alors définie entre $-\infty$ et $+\infty$ et la probabilité

$$\pi_i = P(Y_i = 1|x_i) = \frac{\exp\{\beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}\}}{1 + \exp\{\beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}\}}$$

est définie entre 0 et 1.

La vraisemblance et les estimateurs

Pour estimer les paramètres de la régression logistique, nous utilisons l'estimateur du maximum de vraisemblance. Dans le cas de données non groupées, nous pouvons écrire la vraisemblance de la manière suivante :

$$L(\beta_0, \beta_1, \ldots, \beta_p) = \prod_{i=1}^{n} \pi_i^{y_i} (1 - \pi_i)^{(1-y_i)}.$$

Cela revient à maximiser la log-vraisemblance $l(\boldsymbol{\beta})$:

$$\begin{aligned}
\log L(\boldsymbol{\beta}) &= \log \left\{ \prod_{i=1}^{n} \pi_i^{y_i} (1-\pi_i)^{(1-y_i)} \right\} \\
&= \sum_{i=1}^{n} \log \pi_i^{y_i} (1-\pi_i)^{(1-y_i)} \\
&= \sum_{i=1}^{n} [y_i \log \pi_i + (1-y_i) \log(1-\pi_i)] \\
&= \sum_{i=1}^{n} [y_i \log \pi_i - y_i \log(1-\pi_i) + \log(1-\pi_i)] \\
&= \sum_{i=1}^{n} \left[y_i \log \left(\frac{\pi_i}{1-\pi_i} \right) + \log(1-\pi_i) \right] \\
&= \sum_{i=1}^{n} \left[y_i \log \left(\exp\{\beta_0 + \boldsymbol{\beta}^T \mathbf{x}_i\} \right) + \log \left(1 - \frac{\exp\{\beta_0 + \boldsymbol{\beta}^T \mathbf{x}_i\}}{1 + \exp\{\beta_0 + \boldsymbol{\beta}^T \mathbf{x}_i\}} \right) \right] \\
&= \sum_{i=1}^{n} \left[y_i (\beta_0 + \boldsymbol{\beta}^T \mathbf{x}_i) + \log \left(\frac{1}{1 + \exp\{\beta_0 + \boldsymbol{\beta}^T \mathbf{x}_i\}} \right) \right] \\
&= \sum_{i=1}^{n} \left[y_i (\beta_0 + \boldsymbol{\beta}^T \mathbf{x}_i) - \log \left(1 + \exp\{\beta_0 + \boldsymbol{\beta}^T \mathbf{x}_i\} \right) \right] . \qquad (2.1)
\end{aligned}$$

Pour trouver les paramètres β_j, $j = 0, \ldots, p$, qui maximisent cette log-vraisemblance, il faut dériver l'équation (2.1) par rapport à chacun des paramètres et mettre les expressions obtenues égales à 0. Pour accommoder l'ordonnée à l'origine et pour simplifier l'écriture, nous incluons la constante 1 comme première entrée au vecteur $\mathbf{x}_i$. Les dérivées premières sont

$$\begin{aligned}
\frac{\partial l(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} &= \sum_{i=1}^{n} \left\{ y_i \mathbf{x}_i - \frac{\exp\{\boldsymbol{\beta}^T \mathbf{x}_i\} \mathbf{x}_i}{1 + \exp\{\boldsymbol{\beta}^T \mathbf{x}_i\}} \right\} \\
&= \sum_{i=1}^{n} [y_i \mathbf{x}_i - \pi_i \mathbf{x}_i] \\
&= \sum_{i=1}^{n} \mathbf{x}_i (y_i - \pi_i)
\end{aligned}$$

et les $p+1$ équations à résoudre sont

$$\sum_{i=1}^{n} \mathbf{x}_i(y_i - \pi_i) = \mathbf{0}.$$

Contrairement à la régression linéaire, nous obtenons des équations non linéaires qui ne peuvent être résolues que par des procédures numériques itératives telles que la méthode de Newton-Raphson. L'algorithme de Newton-Raphson peut s'écrire de la manière suivante (Hastie *et al.*, 2009) :

$$\boldsymbol{\beta}^{(m+1)} = \boldsymbol{\beta}^{(m)} - \left(\frac{\partial^2 l(\boldsymbol{\beta}^{(m)})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} \right)^{-1} \frac{\partial l(\boldsymbol{\beta}^{(m)})}{\partial \boldsymbol{\beta}}.$$

Ainsi, pour obtenir les estimateurs, nous avons également besoin des dérivées secondes

$$\begin{aligned}
\frac{\partial^2 l(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} &= \frac{\partial}{\partial \boldsymbol{\beta}} \sum_{i=1}^{n} \mathbf{x}_i(y_i - \pi_i) \\
&= \sum_{i=1}^{n} \frac{\partial}{\partial \boldsymbol{\beta}} \mathbf{x}_i(y_i - \pi_i) \\
&= -\sum_{i=1}^{n} \mathbf{x}_i \frac{\partial}{\partial \boldsymbol{\beta}} \pi_i \\
&= -\sum_{i=1}^{n} \mathbf{x}_i \frac{\partial}{\partial \boldsymbol{\beta}} \left(\frac{\exp\{\boldsymbol{\beta}^T \mathbf{x}_i\}}{1 + \exp\{\boldsymbol{\beta}^T \mathbf{x}_i\}} \right) \\
&= -\sum_{i=1}^{n} \mathbf{x}_i \frac{\exp\{\boldsymbol{\beta}^T \mathbf{x}_i\} \mathbf{x}_i^T (1 + \exp\{\boldsymbol{\beta}^T \mathbf{x}_i\}) - (\exp\{\boldsymbol{\beta}^T \mathbf{x}_i\})^2 \mathbf{x}_i^T}{(1 + \exp\{\boldsymbol{\beta}^T \mathbf{x}_i\})^2} \\
&= -\sum_{i=1}^{n} \mathbf{x}_i \frac{\exp\{\boldsymbol{\beta}^T \mathbf{x}_i\} \mathbf{x}_i^T (1 + \exp\{\boldsymbol{\beta}^T \mathbf{x}_i\} - \exp\{\boldsymbol{\beta}^T \mathbf{x}_i\})}{(1 + \exp\{\boldsymbol{\beta}^T \mathbf{x}_i\})^2} \\
&= -\sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^T \left(\frac{\exp\{\boldsymbol{\beta}^T \mathbf{x}_i\}}{1 + \exp\{\boldsymbol{\beta}^T \mathbf{x}_i\}} \right) \left(\frac{1}{1 + \exp\{\boldsymbol{\beta}^T \mathbf{x}_i\}} \right) \\
&= -\sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^T \pi_i (1 - \pi_i).
\end{aligned}$$

L'algorithme de Newton-Raphson est donc

$$\hat{\boldsymbol{\beta}}^{(m+1)} = \hat{\boldsymbol{\beta}}^{(m)} + \left(\sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^T \hat{\pi}_i(1-\hat{\pi}_i)\right)^{-1} \left(\sum_{i=1}^{n} \mathbf{x}_i (y_i - \hat{\pi}_i)\right), \qquad (2.2)$$

avec $\hat{\boldsymbol{\beta}}^{(m)}$ obtenu de l'itération précédente et utilisé dans le calcul de $\hat{\pi}_i$. Quand le critère de convergence est rencontré, par exemple $\frac{\left\|\hat{\boldsymbol{\beta}}^{(m+1)} - \hat{\boldsymbol{\beta}}^{(m)}\right\|_2}{\left\|\hat{\boldsymbol{\beta}}^{(m)}\right\|_2} \leq \varepsilon$, alors l'estimateur de $\boldsymbol{\beta}$ est $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}^{(m)}$.

Nous pouvons réécrire l'équation (2.2) sous sa forme matricielle

$$\hat{\boldsymbol{\beta}}^{(m+1)} = \hat{\boldsymbol{\beta}}^{(m)} + (\mathbf{X}^T \hat{\mathbf{W}} \mathbf{X})^{-1} \mathbf{X}^T (\mathbf{y} - \hat{\boldsymbol{\pi}}),$$

où $\mathbf{X}$ est la matrice $n \times (p+1)$, $\mathbf{y}$ est le vecteur $n \times 1$ contenant les y_i, $\hat{\pi}$ est le vecteur $n \times 1$ contenant les $\hat{\pi}_i$ estimés à l'itération m et $\hat{\mathbf{W}}$ est une matrice $n \times n$ diagonale avec $w_{ii} = \hat{\pi}_i(1-\hat{\pi}_i)$.

Les estimateurs des paramètres s'obtiennent de manière similaire pour des données groupées.

L'interprétation des coefficients

L'interprétation de la valeur des coefficients de la régression logistique est basée sur la notion de « cote ». Une cote, notée par Ω, est la chance d'avoir $Y = 1$ par rapport à $Y = 0$ et elle est définie mathématiquement comme ceci :

$$\Omega = \frac{\pi_i}{1-\pi_i}.$$

La plupart du temps, ce qui nous intéresse est de comparer la cote d'un groupe par rapport à celle d'un groupe de référence. Pour y arriver, nous calculons le rapport de cotes

$$\omega = \frac{\Omega_{\{Y=1\}}}{\Omega_{\{Y=0\}}}.$$

Ce rapport est en fait égal à $\exp\{\beta_j\}$. Par exemple, si nous voulons comparer le risque d'avoir le cancer chez un fumeur et un non-fumeur, nous calculons le rapport de cotes

$$\begin{aligned}\omega &= \frac{\frac{\pi_i}{1-\pi_i}\Big|x_{i1}=1}{\frac{\pi_h}{1-\pi_h}\Big|x_{h1}=0} \\ &= \frac{\exp\{\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+\ldots+\beta_p x_{ip}\}}{\exp\{\beta_0+\beta_1 x_{h1}+\beta_2 x_{h2}+\ldots+\beta_p x_{hp}\}} \\ &= \frac{\exp\{\beta_0+\beta_1\cdot 1+\beta_2 x_{i2}+\ldots+\beta_p x_{ip}\}}{\exp\{\beta_0+\beta_1\cdot 0+\beta_2 x_{i2}+\ldots+\beta_p x_{ip}\}} \\ &= \exp\{\beta_1\}.\end{aligned}$$

Le risque d'avoir le cancer chez la personne qui fume (individu i) est $\exp\{\beta_1\}$ fois celui chez la personne qui ne fume pas (individu h) lorsque toutes les autres variables demeurent inchangées ($x_{ij} = x_{hj}$ pour $j = 2, \ldots, p$).

Nous obtenons un rapport de cotes similaire pour un prédicteur continu

$$\begin{aligned}\omega &= \frac{\frac{\pi_i}{1-\pi_i}\Big|x_{i1}+1}{\frac{\pi_i}{1-\pi_i}\Big|x_{i1}} \\ &= \frac{\exp\{\beta_0+\beta_1(x_{i1}+1)+\beta_2 x_{i2}+\ldots+\beta_p x_{ip}\}}{\exp\{\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+\ldots+\beta_p x_{ip}\}} \\ &= \exp\{\beta_1\}.\end{aligned}$$

De manière générale, si β_j est supérieur à 0, alors le risque de succès augmente lorsque x_{ij} augmente d'une unité et que toutes les autres variables restent inchangées. Au contraire, si β_j est inférieur à 0, alors le risque de succès diminue. Finalement, si β_j est égal à 0, alors le risque ne change pas et la variable explicative n'a pas d'effet.

Nous verrons dans les prochaines sections comment faire pour tester la significativité de ces coefficients. Nous étudierons, entre autres, les tests d'hypothèses et la déviance, le test d'Hosmer-Lemeshow et le test de Wald pour mesurer l'association. Nous étudierons également la courbe ROC, courbe qui permet de mesurer

la force de prédiction de notre modèle. Nous avons accès à des tests d'association et de prédiction, car la régression logistique est une méthode statistique employée pour des jeux de données de petite dimension. Nous verrons plus loin dans ce chapitre que les méthodes destinées aux données massives ne sont pas associées à des tests d'association.

Les tests d'hypothèses et la déviance

Il est possible que les variables explicatives de notre modèle de départ ne soient pas toutes nécessaires pour avoir un bon ajustement. Un modèle réduit peut être plus approprié dans certains cas. Pour nous aider à choisir le meilleur modèle, c'est-à-dire celui qui s'ajuste le mieux aux données, nous faisons une analyse de la déviance. Considérons tout d'abord le modèle M1 composé des variables explicatives $k+1$ à p $(0 < k < p)$ et le modèle M2 composé des variables 1 à p. Le modèle M1, qui ne contient que quelques variables, est appelé « modèle réduit » et le modèle M2, qui est composé de toutes les variables, est appelé « modèle saturé ». Ces modèles sont dits des modèles emboîtés, car les paramètres du modèle réduit se retrouvent tous dans le modèle saturé. Le but est de tester si le bloc de variables 1 à k est significatif via les hypothèses suivantes :

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_k = 0$$

$$H_1 : \text{Au moins un } \beta_j \text{ est différent de 0, } j = 1, \ldots, k.$$

Pour comparer ces deux modèles, nous calculons la déviance résiduelle, une mesure basée sur les log-vraisemblances. Elle est définie comme ceci :

$$D = 2 \cdot [logL_{M2} - logL_{M1}].$$

Il est à noter que le modèle M2 peut être un modèle complet (saturé), mais aussi un modèle réduit emboîtant le modèle M1. Si M2 est saturé, alors nous parlerons

de déviance résiduelle et s'il est réduit, nous parlerons plutôt de déviance. Sous l'hypothèse nulle, la déviance suit asymptotiquement une loi Khi-Deux à ν degrés de liberté où ν est égal au nombre de paramètres dans le modèle M2 moins le nombre de paramètres dans le modèle M1. Nous pouvons ainsi calculer une valeur-p. Si la valeur-p est supérieure à α (par exemple, $\alpha = 0.05$), alors nous ne rejetons pas H_0 et le bloc de variables n'a pas d'effet. Le modèle réduit est donc adéquat. Pour ce qui est de la déviance résiduelle, elle suit aussi une loi Khi-Deux à ν degrés de liberté sous H_0 pour les données groupées. Toutefois, elle ne suit pas cette loi lorsque nous sommes en présence de données non groupées. Nous ne pouvons pas utiliser l'analyse de déviance dans cette situation. Nous pouvons cependant effectuer le test d'Hosmer-Lemeshow pour tester l'ajustement du modèle.

Le test d'Hosmer-Lemeshow

Le test d'Hosmer-Lemeshow peut être utilisé pour tester l'adéquation d'un modèle lorsque les données sont non groupées. L'idée générale de ce test est de partitionner l'intervalle [0,1] en g groupes, puis de classer les n individus dans les g sous-intervalles en se basant sur leur probabilité estimée $\hat{\pi}_i$ et finalement de comparer le nombre observé et le nombre espéré de cas dans chacun des groupes. La statistique d'Hosmer-Lemeshow est obtenue en calculant la statistique du Khi-Deux à partir de tableaux $g \times 2$ de fréquences observées et espérées (Hosmer et Lemeshow, 2000). Cette statistique est donnée par

$$\chi^2_{HL} = \sum_{l=1}^{g} \frac{(\tilde{Y}_l - m_l \bar{\pi}_l)^2}{m_l \bar{\pi}_l^2 (1 - \bar{\pi}_l)},$$

où $\tilde{Y}_l$ est le nombre d'événements dans le groupe l, m_l est le nombre de personnes dans le groupe l et $\bar{\pi}_l$ est la moyenne des probabilités estimées dans le groupe l. Il a été démontré grâce à des simulations que la statistique d'Hosmer-Lemeshow suit une loi Khi-Deux de $g - 2$ degrés de liberté (Hosmer et Lemeshow, 2000).

Le test de Wald

Il est également possible de tester la significativité des coefficients sans avoir à comparer deux modèles emboîtés. Nous pouvons tester

$$H_0 : \boldsymbol{\beta} = \mathbf{0},$$
$$H_1 : \beta_j \neq 0 \text{ pour au moins un j.}$$

La statistique de Wald est

$$\begin{aligned} W &= \hat{\boldsymbol{\beta}}^{\boldsymbol{T}} \left[\widehat{Var(\hat{\boldsymbol{\beta}})}\right]^{-1} \hat{\boldsymbol{\beta}} \\ &= \hat{\boldsymbol{\beta}}^{\boldsymbol{T}} [\mathbf{X^T \hat{V}^{-1} X}] \hat{\boldsymbol{\beta}}, \end{aligned}$$

$$\text{où } \mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1p} \\ 1 & x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{np} \end{bmatrix} \text{ et } \mathbf{V} = diag\{\pi_i(1-\pi_i)\}_{i=1}^{n}.$$

Cette statistique suit une loi Khi-Deux avec $p+1$ degrés de liberté. Si la valeur-p est inférieure à 0.05, alors nous rejetons H_0 et nous pouvons affirmer qu'il y a au moins une variable qui affecte la réponse lorsque les autres covariables sont présentes dans le modèle.

La courbe ROC

Une fois les variables explicatives choisies pour former le modèle final, nous pouvons tester son efficacité de prédiction à l'aide de la courbe ROC (« Receiver Operating Characteristic »). La courbe ROC est une mesure de performance d'un classificateur et elle est basée sur les notions de sensibilité et de spécificité. La

sensibilité est la probabilité de classer une personne dans le groupe des malades sachant qu'elle est malade ($Y = 1$) et la spécificité est la probabilité de classer une personne dans le groupe des non-malades sachant qu'elle n'est pas malade ($Y = 0$). La sensibilité est aussi appelée « taux de vrais positifs » et la quantité $1 -$ spécificité est aussi connue sous le nom de « taux de faux positifs ». Nous voulons que le modèle ait le plus grand taux de vrais positifs et le plus petit taux de faux positifs possible.

Le test fonctionne de la manière suivante. Soient p_0 le taux de faux positifs déterminé par le chercheur, $\hat{\pi}_i$ la valeur estimée de π_i et y_i^* la prédiction du groupe pour l'individu i. Si $\hat{\pi}_i \geq p_0$, alors la personne est diagnostiquée malade et $y_i^* = 1$. Si $\hat{\pi}_i < p_0$, alors la personne est déclarée non malade et $y_i^* = 0$. Pour chaque individu, nous vérifions sa valeur prédite et nous la comparons avec la valeur réelle y_i, puis nous calculons la sensibilité. La courbe ROC est la représentation des différentes valeurs de taux de faux positifs possibles sur l'axe des abscisses et de la sensibilité sur l'axe des ordonnées. Plus la courbe est éloignée de la diagonale et est au-dessus de cette diagonale, plus la prédiction est bonne et meilleur est le modèle.

Il est également possible de calculer l'aire sous la courbe ROC. Si l'aire est près de 0.5, alors notre modèle ne prédit pas mieux que le hasard et si elle est près de 1, notre modèle est un bon classificateur. L'aire sous la courbe de la figure (2.1) est égale à 0.9333 ce qui est très près de 1.

L'aire sous la courbe ROC nous permettra de comparer le pouvoir prédictif de diverses méthodes statistiques et par le fait même, de trouver la méthode qui prédit le mieux la présence du cancer chez les individus.

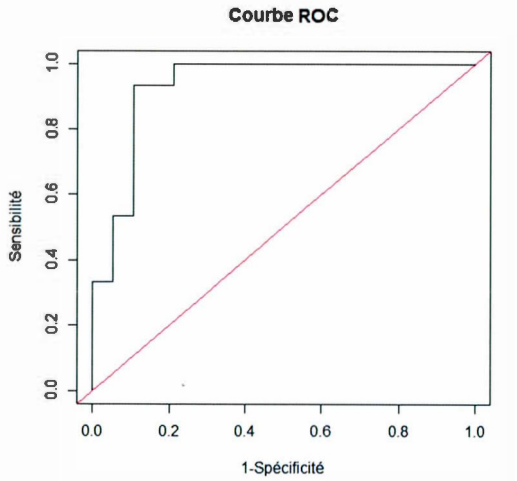


Figure 2.1: Exemple d'une courbe ROC.

2.1.2 Analyse discriminante

Tout comme la régression logistique, l'analyse discriminante (AD) est une technique statistique permettant de trouver les variables qui contribuent le plus à distinguer les groupes et de prédire l'appartenance d'un individu à un groupe. Bien que ces deux méthodes ont des buts semblables, elles n'opèrent pas de la même manière et elles ne supposent pas toutes les mêmes hypothèses.

Analyse multivariée de la variance

Avant de commencer l'analyse discriminante proprement dite, nous pouvons faire une analyse multivariée de la variance (MANOVA, « Multivariate Analysis of Variance ») pour tester l'égalité des vecteurs des moyennes. Cette analyse nous permet de vérifier si la moyenne de chacune des variables explicatives est la même dans chacun des groupes.

Soit $\mathbf{x}_{ij}$ un vecteur d'observations, où $i = 1, \ldots, k$ représente le groupe et $j = 1, \ldots, n_i$ représente l'individu. Nous supposons que les $\mathbf{x}_{ij}$ sont aléatoires et indépendants. Nous voulons tester si les vecteurs de moyennes $\boldsymbol{\mu}_i$, $i = 1, \ldots, k$, sont égaux via les hypothèses suivantes :

$$H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \ldots = \boldsymbol{\mu}_k,$$

H_1 : Au moins un vecteur $\boldsymbol{\mu}_i$ est différent des autres.

L'hypothèse nulle peut aussi être écrite comme suit :

$$H_0 : \begin{bmatrix} \mu_{11} \\ \mu_{12} \\ \vdots \\ \mu_{1p} \end{bmatrix} = \begin{bmatrix} \mu_{21} \\ \mu_{22} \\ \vdots \\ \mu_{2p} \end{bmatrix} = \ldots = \begin{bmatrix} \mu_{k1} \\ \mu_{k2} \\ \vdots \\ \mu_{kp} \end{bmatrix}.$$

Toutes les égalités doivent être vraies pour ne pas rejeter l'hypothèse nulle. Pour tester l'homogénéité des moyennes, quatre tests sont à notre disposition et ils sont tous basés sur les matrices des sommes des carrés inter-groupes et intra-groupes.

La matrice des sommes des carrés inter-groupes, notée $\mathbf{H}$, est donnée par

$$\mathbf{H} = \sum_{i=1}^{k} n_i (\bar{\mathbf{x}}_{i.} - \bar{\mathbf{x}}_{..})(\bar{\mathbf{x}}_{i.} - \bar{\mathbf{x}}_{..})^T$$

et la matrice des sommes des carrés intra-groupes, notée $\mathbf{E}$, est définie par

$$\mathbf{E} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (\mathbf{x}_{ij} - \bar{\mathbf{x}}_{i.})(\mathbf{x}_{ij} - \bar{\mathbf{x}}_{i.})^T,$$

où $\bar{\mathbf{x}}_{i.} = \sum_{j=1}^{n_i} \mathbf{x}_{ij}/n_i$, $\bar{\mathbf{x}}_{..} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \mathbf{x}_{ij}/N$ et $N = \sum_{i=1}^{k} n_i$.

Un des tests qui nous permet de vérifier si les vecteurs $\boldsymbol{\mu}_i$ sont significativement différents les uns des autres est celui de Wilks. Nous devons supposer pour ce

test que les $\mathbf{x}_{ij}$ sont de loi normale multivariée $N_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$. Le test de Wilks est équivalent à un test de rapport de vraisemblance et il a pour statistique

$$\Lambda = \frac{|\mathbf{E}|}{|\mathbf{E}+\mathbf{H}|}.$$

Si $\Lambda \leq \Lambda_{\alpha,p,\nu_H,\nu_E}$ (distribution bêta multidimensionnelle) où p est le nombre de variables explicatives, $\nu_H = k-1$ est le degré de liberté associé à $\mathbf{H}$ et $\nu_E = N-k$ le degré de liberté associé à $\mathbf{E}$, alors nous rejetons H_0 (Rencher, 2002). Autrement dit, s'il y a l'homogénéité à l'intérieur des groupes et l'hétérogénéité entre les groupes, les moyennes des groupes seront significativement différentes. La statistique de Wilks peut se transformer en une Fisher exacte ou en une approximation de Fisher selon les valeurs de p et de ν_H (Rencher, 2002). Il est également possible d'écrire cette statistique en fonction des valeurs propres de la manière suivante :

$$\Lambda = \prod_{i=1}^{s} \frac{1}{1+\lambda_i},$$

où $s = \min(p, \nu_H)$.

Un deuxième test possible est le test d'union-intersection de Roy aussi appelé « Roy's largest root test ». Il vise à trouver le vecteur qui discrimine le plus les groupes. Autrement dit, nous cherchons le vecteur $\mathbf{a}$ qui maximise

$$F = \frac{\dfrac{\mathbf{a}^T\mathbf{H}\mathbf{a}}{k-1}}{\dfrac{\mathbf{a}^T\mathbf{E}\mathbf{a}}{N-k}},$$

où F suit une loi de Fisher.

Ceci est maximisé par le vecteur propre $\mathbf{a}_1$:

$$\max_{\mathbf{a}} F = \frac{\dfrac{\mathbf{a}_1^T\mathbf{H}\mathbf{a}_1}{k-1}}{\dfrac{\mathbf{a}_1^T\mathbf{E}\mathbf{a}_1}{N-k}} = \frac{N-k}{k-1}\lambda_1. \tag{2.3}$$

Puisque nous maximisons sur a, l'équation (2.3) ne suit pas une loi de Fisher (Rencher, 2002). La statistique de Roy est donc donnée par

$$\theta = \frac{\lambda_1}{1+\lambda_1}.$$

Nous rejetons l'hypothèse nulle pour de grandes valeurs de θ.

Les deux derniers tests pour vérifier l'homogénéité des moyennes sont les tests de Pillai et de Lawley-Hotelling. La statistique de Pillai est

$$V^{(s)} = tr[(H+E)^{-1}H] = \sum_{i=1}^{s} \frac{\lambda_i}{1+\lambda_i}$$

et celle de Lawley-Hotelling est

$$U^{(s)} = tr[E^{-1}H] = \sum_{i=1}^{s} \lambda_i.$$

La région de rejet pour les tests de Pillai et de Lawley-Hotelling est pour de grandes valeurs de $V^{(s)}$ et de $U^{(s)}$ respectivement.

Il est possible de vérifier que si $\min(p, \nu_H) = 1$, alors les quatre tests donnent la même statistique F et la même valeur-p. Nous obtenons les mêmes conclusions, peu importe la méthode choisie.

Pour effectuer les quatre tests ci-haut, nous avons émis les hypothèses que les $\mathbf{x}_{ij} \sim N_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$. Une des suppositions qui a été faite est donc que les données proviennent d'une normale multivariée. Si ce n'est pas le cas, les tests sont quand même assez robustes à condition qu'il y ait beaucoup d'observations ou que la non-normalité soit due à l'asymétrie des observations et non à des données aberrantes (Tabachnick et Fidell, 2007). L'autre supposition est l'homogénéité des matrices de variance-covariance : $\Sigma_1 = \Sigma_2 = \ldots = \Sigma_k$. Si nous sommes en présence d'un modèle balancé, c'est-à-dire que tous les groupes ont la même taille, alors les quatre tests sont robustes à l'hétérogénéité des matrices. Au contraire, si le modèle est

non balancé, c'est-à-dire que les groupes sont de différentes tailles, alors les tests ne sont pas tous aussi performants. Le test le plus robuste à l'hétérogénéité des matrices de variance-covariance est celui de Pillai. Néanmoins, le test le plus utilisé est celui de Wilks (Tabachnick et Fidell, 2007). L'homogénéité des matrices peut être vérifiée grâce au test M de Box. Il faut cependant être prudent avec ce test, car il est sensible à la non-normalité des observations et aux groupes de différentes tailles. Dans le cas où toutes les hypothèses sont respectées et que les vecteurs des moyennes sont colinéaires, il est préférable d'utiliser le test de Roy, car il est le plus puissant (Rencher, 2002).

La fonction discriminante

Sachant que les vecteurs des moyennes sont significativement différents, nous pouvons trouver un sous-ensemble de variables qui séparent bien les individus en plusieurs groupes distincts et connus. Pour y arriver, nous devons premièrement trouver la ou les fonction(s) discriminante(s) qui maximise(nt) la distance des centroïdes. En d'autres mots, nous devons trouver la ou les combinaison(s) linéaire(s) des p variables qui permet(tent) de distinguer le plus les groupes.

Supposons que nous avons $k = 2$ groupes, que l'hypothèse d'égalité des deux matrices de variance-covariance n'est pas rejetée et que les vecteurs des moyennes sont significativement différents. Nous transformons les vecteurs $\mathbf{x}_{ij}$ en scalaire de la manière suivante :

$$\begin{aligned} z_{ij} &= \mathbf{a}^T \mathbf{x}_{ij} \\ &= a_1 x_{ij1} + a_2 x_{ij2} + \ldots + a_p x_{ijp}, \end{aligned}$$

où $i = \{1, 2\}$ est le groupe et $j = 1, \ldots, n_i$ l'individu. Par la suite, nous trouvons

la fonction discriminante $\mathbf{a}$ en maximisant le carré des différences standardisées

$$\frac{(\bar{z}_1 - \bar{z}_2)^2}{s_z^2} = \frac{[\mathbf{a}^T(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)]^2}{\mathbf{a}^T\mathbf{S}_{pl}\mathbf{a}} = \frac{\mathbf{a}^T(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^T\mathbf{a}}{\mathbf{a}^T\mathbf{S}_{pl}\mathbf{a}}, \tag{2.4}$$

où $\bar{z}_i = \mathbf{a}^T\bar{\mathbf{x}}_i$ est la moyenne transformée du groupe i et $\mathbf{S}_{pl} = \sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}}$ est l'écart-type pondéré.

Le maximum est atteint lorsque

$$\mathbf{a} = \mathbf{S}_{pl}^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2).$$

Il n'y a qu'une seule fonction discriminante puisque nous n'avons que deux groupes et celle-ci est parallèle à la ligne joignant $\bar{\mathbf{x}}_1$ et $\bar{\mathbf{x}}_2$ (Rencher, 2002).

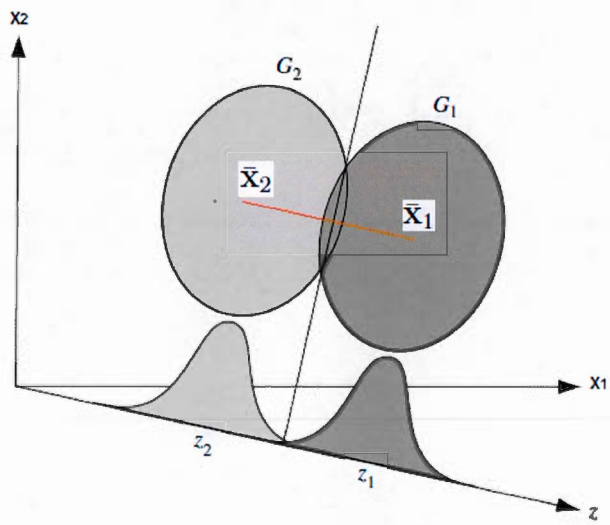


Figure 2.2: Exemple d'une fonction discriminante. La droite joingnant $\bar{\mathbf{x}}_1$ et $\bar{\mathbf{x}}_2$ est parallèle à la fonction discriminante $\mathbf{z}$. Note : La figure a été prise de Rencher (2002), mais elle a été légèrement modifiée.

Nous pouvons interpréter les coefficients standardisés $a_r, r = 1, \ldots, p$, comme suit : si la valeur absolue du coefficient est grande, alors la variable r distingue bien les deux groupes. Par exemple, si $z_{1j} = 5x_{1j1} - 7x_{1j2} + 0.1x_{1j3}$, alors les variables 1 et 2 discriminent bien les groupes. Comme le coefficient en valeur absolue de la variable 2 est plus grand que celui de la variable 1, nous pouvons ajouter que la variable 2 est plus efficace pour séparer les groupes.

Tout ce qui a été fait précédemment pour deux groupes peut être généralisé pour $k > 2$. Les fonctions discriminantes sont obtenues en remplaçant $(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^T$ par la matrice inter-groupes $\mathbf{H}$ et $\mathbf{S}_{pl}$ par la matrice intra-groupes $\mathbf{E}$ dans l'équation (2.4). Ainsi, nous avons

$$\lambda_j = \frac{\mathbf{a}_j^T \mathbf{H} \mathbf{a}_j}{\mathbf{a}_j^T \mathbf{E} \mathbf{a}_j},$$

où λ_j est la valeur propre associée au vecteur propre $\mathbf{a}_j$, $j = 1, \ldots, s = \min(k - 1, p)$. Nous pouvons interpréter les coefficients de manière similaire que pour deux groupes.

Tout comme la MANOVA, l'analyse discriminante est assez robuste à la non-normalité si elle provient d'une asymétrie et à l'hétérogénéité des matrices de variance-covariance si les échantillons sont de grande taille. De plus, les groupes de taille inégale ne semblent pas causer de problème pour cette analyse (Tabachnick et Fidell, 2007).

Règle de classification

Un autre aspect intéressant de l'analyse discriminante est la prédiction. Nous voulons classer un nouvel individu dans un des k groupes. Supposons qu'il y a $k = 2$ groupes (G1 et G2) et que les matrices de variance-covariance sont égales. Fisher a proposé une règle de classification linéaire qui minimise la probabilité

de mauvais classement. Cette méthode est basée sur la fonction discriminante. La procédure de classification de Fisher assigne l'individu ayant un vecteur $\mathbf{x}$ connu à G1 si $z = \mathbf{a}^T\mathbf{x}$ est près de la moyenne transformée $\bar{z}_1 = \mathbf{a}^T\bar{\mathbf{x}}_1 = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^T\mathbf{S}_{pl}^{-1}\bar{\mathbf{x}}_1$. De la même manière, la procédure de classification assigne l'individu à G2 si z est proche de $\bar{z}_2$. Mais qu'est-ce ça signifie être « proche » de la moyenne transformée ? Le score z est proche de $\bar{z}_1$ si $z > \frac{1}{2}(\bar{z}_1 + \bar{z}_2)$, sinon il est proche de $\bar{z}_2$. En résumé, la personne est classée dans le groupe G1 si

$$\mathbf{a}^T\mathbf{x} = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^T\mathbf{S}_{pl}^{-1}\mathbf{x} > \frac{1}{2}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^T\mathbf{S}_{pl}^{-1}(\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2),$$

ou dans le groupe G2 si

$$\mathbf{a}^T\mathbf{x} = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^T\mathbf{S}_{pl}^{-1}\mathbf{x} < \frac{1}{2}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^T\mathbf{S}_{pl}^{-1}(\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2).$$

La règle de classification peut être généralisée au cas $k > 2$. Pour classer un nouvel individu, la distance de Mahalanobis est de mise. Elle nous permet de connaître la distance qui sépare $\mathbf{x}$ de la moyenne de chaque groupe. Lorsqu'il y a homogénéité des matrices de variance-covariance, la distance de Mahalanobis se calcule via

$$D_i^2(\mathbf{x}) = (\mathbf{x} - \bar{\mathbf{x}}_i)^T\mathbf{S}_{pl}^{-1}(\mathbf{x} - \bar{\mathbf{x}}_i), \tag{2.5}$$

où

$$\mathbf{S}_{pl}^{-1} = \frac{1}{N-k}\sum_{i=1}^{k}(n_i - 1)\mathbf{S}_i = \frac{\mathbf{E}}{N-k}$$

est la matrice de variance-covariance pondérée, n_i est la taille du groupe i et $\mathbf{S}_i$ est la matrice de variance-covariance du groupe i.

Nous assignons l'individu $\mathbf{x}$ au groupe qui a la plus petite distance $D_i^2(\mathbf{x})$.

En général, si $\Sigma_1 = \Sigma_2 = \ldots = \Sigma_k$, nous avons

$$\begin{aligned} D_i^2(\mathbf{x}) &= (\mathbf{x} - \bar{\mathbf{x}}_i)^T \mathbf{S}_{pl}^{-1} (\mathbf{x} - \bar{\mathbf{x}}_i) \\ &= (\mathbf{x}^T - \bar{\mathbf{x}}_i^T)(\mathbf{S}_{pl}^{-1}\mathbf{x} - \mathbf{S}_{pl}^{-1}\bar{\mathbf{x}}_i) \\ &= \mathbf{x}^T \mathbf{S}_{pl}^{-1}\mathbf{x} - \bar{\mathbf{x}}_i^T \mathbf{S}_{pl}^{-1}\mathbf{x} - \mathbf{x}^T \mathbf{S}_{pl}^{-1}\bar{\mathbf{x}}_i + \bar{\mathbf{x}}_i^T \mathbf{S}_{pl}^{-1}\bar{\mathbf{x}}_i \\ &= \mathbf{x}^T \mathbf{S}_{pl}^{-1}\mathbf{x} - 2\bar{\mathbf{x}}_i^T \mathbf{S}_{pl}^{-1}\mathbf{x} + \bar{\mathbf{x}}_i^T \mathbf{S}_{pl}^{-1}\bar{\mathbf{x}}_i. \end{aligned}$$

Ainsi, la fonction discriminante

$$L_i(\mathbf{x}) = -2\bar{\mathbf{x}}_i^T \mathbf{S}_{pl}^{-1}\mathbf{x} + \bar{\mathbf{x}}_i^T \mathbf{S}_{pl}^{-1}\bar{\mathbf{x}}_i$$

est une fonction linéaire en $\mathbf{x}$.

Si nous avons l'hétérogénéité des matrices de variance-covariance, nous remplaçons $\mathbf{S}_{pl}$ par $\mathbf{S}_i$ dans l'équation (2.5) et nous obtenons

$$D_i^2(\mathbf{x}) = (\mathbf{x} - \bar{\mathbf{x}}_i)^T \mathbf{S}_i^{-1} (\mathbf{x} - \bar{\mathbf{x}}_i).$$

Dans ce cas, nous parlerons d'analyse discriminante quadratique, car

$$\begin{aligned} D_i^2(\mathbf{x}) &= (\mathbf{x} - \bar{\mathbf{x}}_i)^T \mathbf{S}_i^{-1} (\mathbf{x} - \bar{\mathbf{x}}_i) \\ &= (\mathbf{x}^T - \bar{\mathbf{x}}_i^T)(\mathbf{S}_i^{-1}\mathbf{x} - \mathbf{S}_i^{-1}\bar{\mathbf{x}}_i) \\ &= \mathbf{x}^T \mathbf{S}_i^{-1}\mathbf{x} - \bar{\mathbf{x}}_i^T \mathbf{S}_i^{-1}\mathbf{x} - \mathbf{x}^T \mathbf{S}_i^{-1}\bar{\mathbf{x}}_i + \bar{\mathbf{x}}_i^T \mathbf{S}_i^{-1}\bar{\mathbf{x}}_i \\ &= \mathbf{x}^T \mathbf{S}_i^{-1}\mathbf{x} - 2\bar{\mathbf{x}}_i^T \mathbf{S}_i^{-1}\mathbf{x} + \bar{\mathbf{x}}_i^T \mathbf{S}_i^{-1}\bar{\mathbf{x}}_i \end{aligned}$$

est une fonction quadratique en $\mathbf{x}$ (à cause du terme $\mathbf{x}^T \mathbf{S}_i^{-1}\mathbf{x}$).

L'analyse discriminante vs la régression logistique

Bien que la régression logistique et l'analyse discriminante linéaire (ADL) répondent à la même question, la régression logistique est une technique plus flexible. En effet, elle ne fait aucune supposition quant à la normalité des données (x_{ij}) et quant à l'homogénéité des matrices de variance-covariance. De plus, l'ADL a tendance à surestimer l'association entre la variable dépendante et les variables indépendantes si ces dernières sont dichotomiques (Tabachnick et Fidell, 2007). C'est pourquoi la régression logistique est souvent privilégiée à l'analyse discriminante. Par contre, si les suppositions faites dans l'ADL sont respectées, elle est plus précise que la logistique, car cette dernière ignore la distribution des x_{ij}. Cependant, ni la régression logistique ni l'ADL n'est robuste aux données extrêmes.

2.2 Apprentissage supervisé pour des données de grande dimension

Lorsque nous parlons de données de grande dimension, nous faisons référence à des jeux de données qui sont composés d'un très grand nombre de variables explicatives et d'un faible nombre d'individus. Autrement dit, la matrice $\mathbf{X}$ a plus de colonnes que de lignes ($p \gg n$). Ce type de données pose problème lors de l'estimation des coefficients β_j puisque la matrice $\mathbf{X}^T\hat{\mathbf{W}}\mathbf{X}$ n'est pas inversible. Nous verrons plus en détail dans la section 2.2.1 en quoi la singularité de la matrice $\mathbf{X}$ nous empêche d'appliquer les méthodes classiques vues précédemment et comment nous pouvons remédier à ce problème.

Un autre obstacle auquel nous sommes confrontés avec les grandes masses de données est la présence de plusieurs informations non pertinentes. En effet, plusieurs données parmi celles recueillies ne sont que du bruit. Nous devons faire une « fouille de données » (« data mining ») et éliminer toutes les variables non pertinentes pou-

vant amener à une mauvaise classification. L'objectif général est donc de tester l'influence des variables et de sélectionner le meilleur sous-ensemble possible de variables associées à une certaine caractéristique (une maladie, par exemple).

Pour résoudre ces deux principaux problèmes, une multitude de méthodes statistiques ont été développées. Nous verrons en premier lieu la régression Ridge dans la section 2.2.2. Puis, nous étudierons la méthode LASSO et le groupe LASSO dans les sections 2.2.3 et 2.2.4. Par la suite, nous examinerons la méthode Elastic Net dans la section 2.2.5. Finalement, il sera question de l'analyse discriminante pénalisée basée sur la fonction de Fisher dans la section 2.2.6. Il est à noter que ces méthodes sont efficaces pour faire de la prédiction, par contre elles ne résolvent pas tous les problèmes. En effet, ces techniques de réduction sont bonnes pour éliminer le bruit et ainsi prédire une variable réponse, mais elles ne sont associées à aucun test d'association. Il faudra donc revenir aux méthodes standard (classiques) après avoir fait la réduction pour tester l'association.

2.2.1 Problématique

Pour un modèle de régression logistique avec $p < n$, nous avons vu précédemment que nous pouvions estimer le vecteur $\boldsymbol{\beta}$ comme ceci :

$$\hat{\boldsymbol{\beta}}^{(m+1)} = \hat{\boldsymbol{\beta}}^{(m)} + (\mathbf{X}^T\hat{\mathbf{W}}\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{y} - \hat{\boldsymbol{\pi}})$$

où $\mathbf{X}$ est la matrice $n \times (p+1)$, $\mathbf{y}$ est le vecteur contenant les y_i, $\hat{\pi}$ est le vecteur contenant les $\hat{\pi}_i$ estimés à l'itération m et $\hat{\mathbf{W}}$ est une matrice diagonale avec $w_{ii} = \hat{\pi}_i(1 - \hat{\pi}_i)$. Les estimateurs $\hat{\beta}_j$, $j = 0, \ldots, p$, obtenus sont uniques, sans biais et de petite variance.

Par contre, lorsque $p \gg n$, la matrice $\mathbf{X}^T\hat{\mathbf{W}}\mathbf{X}$ devient de rang au plus n, et par le fait même elle devient singulière, ce qui implique que $\mathbf{X}^T\hat{\mathbf{W}}\mathbf{X}$ n'est plus inversible.

Les solutions pour β, bien qu'elles demeurent sans biais, ne sont plus uniques (il y a une infinité de solutions possibles) et les variances tendent vers l'infini.

Pour remédier à ce problème, plusieurs auteurs se sont tournés vers des méthodes dites de régularisation. Ces méthodes consistent à ajouter un terme, appelé pénalité, à une fonction de perte. En effet, soit $L(\mathbf{X}, \mathbf{y})$ une fonction de perte et $P(\beta)$ une pénalité qui dépend de $\beta \in \Re^p$. Les méthodes de régularisation permettent d'estimer (β_0, β) en minimisant

$$\min_{(\beta_0,\beta)} \{L(\mathbf{X}, \mathbf{y}) + \lambda \cdot P(\beta)\},$$

où $\lambda \geq 0$ est le paramètre de régularisation.

Dans le cas où $\mathbf{y}$ est binaire, la fonction de perte est égale à *moins* la log-vraisemblance du modèle logistique.

La majorité des méthodes de régularisation permettent non seulement de résoudre le problème de singularité, mais aussi celui relié aux bruits (variables non pertinentes). Le paramètre λ contrôle la force de réduction, c'est-à-dire le nombre de variables qui auront un coefficient égal à 0. Si λ est grand, plusieurs coefficients sont égaux à 0 et au contraire, si λ est petit, peu de coefficients sont égaux à 0. Ceci permet de réduire le nombre de variables explicatives pour ne garder que celles significatives.

Les méthodes de régularisation introduisent un biais dans l'estimé des paramètres du modèle, toutefois la variance est inférieure à celle du modèle non pénalisé. Il faut donc trouver le juste milieu entre le biais et la variance pour bien refléter les données et pour faire de bonnes prédictions.

2.2.2 Régression Ridge

La régression Ridge, développée par Hoerl et Kennard (1970), est une méthode de régularisation ayant pour pénalité la norme l_2, définie comme suit

$$P(\boldsymbol{\beta}) = ||\boldsymbol{\beta}||_2^2 = \sum_{j=1}^{p} \beta_j^2 = \boldsymbol{\beta}^T \boldsymbol{\beta} = l_2.$$

Son estimateur est défini par

$$\begin{aligned}\left(\hat{\beta}_0, \hat{\boldsymbol{\beta}}\right) &= \underset{(\beta_0, \boldsymbol{\beta})}{\operatorname{argmin}} \left\{L(\mathbf{X}, \mathbf{y}) + \lambda l_2\right\} \\ &= \underset{(\beta_0, \boldsymbol{\beta})}{\operatorname{argmin}} \left\{L(\mathbf{X}, \mathbf{y}) + \lambda \boldsymbol{\beta}^T \boldsymbol{\beta}\right\}.\end{aligned}$$

Puisque nous considérons le cas où $\mathbf{y}$ est dichotomique, nous pouvons réécrire l'estimateur Ridge avec *moins* la log-vraisemblance comme fonction de perte :

$$\left(\hat{\beta}_0, \hat{\boldsymbol{\beta}}\right) = \underset{(\beta_0, \boldsymbol{\beta})}{\operatorname{argmin}} \left\{ -\sum_{i=1}^{n} \left[y_i \log \pi_i + (1 - y_i) \log(1 - \pi_i)\right] + \lambda \boldsymbol{\beta}^T \boldsymbol{\beta} \right\}. \tag{2.6}$$

L'estimateur de $(\beta_0, \boldsymbol{\beta})$ peut être obtenue à l'aide de l'algorithme de Newton-Raphson. Lorsqu'il y a convergence, l'estimateur des paramètres est

$$\hat{\boldsymbol{\beta}}^{\lambda} = \{\Omega(\boldsymbol{\beta}_0) + 2\lambda \mathbf{I}\}^{-1} \Omega(\boldsymbol{\beta}_0) \hat{\boldsymbol{\beta}},$$

où $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}_0 + \Omega^{-1}(\boldsymbol{\beta}_0) U(\boldsymbol{\beta}_0)$, avec $\Omega^{-1}(\boldsymbol{\beta}_0)$ la matrice de covariance et $U(\boldsymbol{\beta}_0)$ la dérivée de la log-vraisemblance non restreinte. Veuillez vous référer à l'appendice B pour plus de détails et à l'article de Le Cessie et Van Houwelingen (1992).

Le biais asymptotique, qui est introduit lorsque nous utilisons l'algorithme avec une pénalité, est

$$\mathbb{E}\left(\hat{\boldsymbol{\beta}}^{\lambda} - \boldsymbol{\beta}_0\right) = -2\lambda \{\Omega(\boldsymbol{\beta}_0) + 2\lambda \mathbf{I}\}^{-1} \boldsymbol{\beta}_0$$

et la variance asymptotique est

$$\widehat{\text{Var}\left(\hat{\boldsymbol{\beta}}^{\lambda}\right)} = \{\Omega(\boldsymbol{\beta}_0) + 2\lambda\mathbf{I}\}^{-1}\Omega(\boldsymbol{\beta}_0)\{\Omega(\boldsymbol{\beta}_0) + 2\lambda\mathbf{I}\}^{-1}.$$

Ainsi, le fait d'ajouter la pénalité Ridge à notre modèle nous a permis d'obtenir une matrice inversible et de diminuer grandement la variance tout en gardant un biais raisonnable.

L'ajout de la pénalité l_2 dans l'équation (2.6) permet également de rétrécir les coefficients des variables qui sont fortement corrélées vers une même valeur. Elle est donc très efficace lorsque les colonnes de la matrice $\mathbf{X}$ sont corrélées. Ainsi, les gènes faisant partie d'une même voie biologique auront le même coefficient. Cependant, cette technique ne fait que tendre les coefficients des variables non significatives vers 0, elle ne permet pas d'avoir des 0 exacts. Il n'y a donc pas de sélection de variables et les informations non pertinentes demeurent présentes dans notre jeu de données. Les gènes n'ayant aucune relation avec la maladie ne sont pas éliminés avec la régression Ridge.

2.2.3 Méthode LASSO

La méthode de régularisation LASSO (Least Absolute Shrinkage and Selection Operator), proposée par Tibshirani (1996), estime les coefficients de régression en imposant une norme l_1 comme pénalité plutôt que la norme l_2 :

$$P(\boldsymbol{\beta}) = ||\boldsymbol{\beta}||_1 = \sum_{j=1}^{p} |\beta_j| = l_1.$$

L'estimateur des paramètres de la régression logistique pénalisée utilisant LASSO est :

$$\left(\hat{\beta}_0, \hat{\boldsymbol{\beta}}\right) = \underset{(\beta_0,\boldsymbol{\beta})}{\text{argmin}} \left\{ -\sum_{i=1}^{n} [y_i \log \pi_i + (1 - y_i) \log(1 - \pi_i)] + \lambda \sum_{j=1}^{p} |\beta_j| \right\}.$$

Puisque la pénalité LASSO contient une valeur absolue, nous ne pouvons pas dériver la log-vraisemblance pénalisée, ni utiliser l'algorithme de Newton-Raphson pour estimer les paramètres. Nous utilisons plutôt l'algorithme de la descente par coordonnée (Tseng, 2001).

Descente par coordonnée

La descente par coordonnée est un algorithme permettant d'estimer un paramètre à la fois. Considérons le cas où $\mathbf{y}$ est une variable binaire, $\mathbf{x}_i = (\mathbf{x}_{i1}, \ldots, \mathbf{x}_{ip})$ est un vecteur $1 \times p$ et $\boldsymbol{\beta}$ est un vecteur $p \times 1$. Supposons maintenant que les éléments x_{ij} sont standardisés, c'est-à-dire

$$\sum_{i=1}^{n} x_{ij} = 0 \text{ et } \frac{1}{n}\sum_{i=1}^{n} x_{ij}^2 = 1, \quad j = 1, \ldots, p.$$

L'algorithme de la descente par coordonnée tel que développé par Friedman *et al.* (2010) est basé sur l'approximation quadratique (série de Taylor) de la log-vraisemblance

$$l_Q(\beta_0, \boldsymbol{\beta}) = \frac{-1}{2n}\sum_{i=1}^{n} \tilde{w}_i(\tilde{z}_i - \beta_0 - \mathbf{x}_i\boldsymbol{\beta})^2 + C(\tilde{\beta}_0, \tilde{\boldsymbol{\beta}})^2,$$

où

$$\tilde{z}_i = \tilde{\beta}_0 + \mathbf{x}_i\tilde{\boldsymbol{\beta}} + \frac{y_i - \tilde{\pi}_i}{\tilde{\pi}_i(1 - \tilde{\pi}_i)},$$
$$\tilde{w}_i = \tilde{\pi}_i(1 - \tilde{\pi}_i),$$

et $\tilde{\pi}_i$ est évalué avec les paramètres estimés $(\tilde{\beta}_0, \tilde{\boldsymbol{\beta}})$ de l'itération courante. Le terme $C(\tilde{\beta}_0, \tilde{\boldsymbol{\beta}})^2$ est une constante.

L'algorithme de la descente par coordonnée permet de résoudre le problème suivant :

$$\min_{(\beta_0,\boldsymbol{\beta})} \{-l_Q(\beta_0,\boldsymbol{\beta}) + \lambda P(\boldsymbol{\beta})\}.$$

Puisque nous sommes dans le cas d'une régression LASSO et que $P(\boldsymbol{\beta}) = \sum_{j=1}^{p} |\beta_j|$, nous commençons par considérer le cas où β_j est positif. Soit

$$R_{pos}(\beta_j) = -l_Q(\beta_0,\boldsymbol{\beta}) + \lambda\beta_j.$$

La dérivée de $R_{pos}(\beta_j)$ par rapport à β_j est

$$\left.\frac{\partial R_{pos}(\beta_j)}{\partial \beta_j}\right|_{(\beta_0,\boldsymbol{\beta})=(\tilde{\beta}_0,\tilde{\boldsymbol{\beta}})} = \frac{1}{n}\sum_{i=1}^{n} \tilde{w}_i(\tilde{z}_i - \tilde{\beta}_0 - \mathbf{x}_i^T\tilde{\boldsymbol{\beta}})x_{ij} + \lambda. \tag{2.7}$$

Posons $\tilde{y}_i^{(j)} = \tilde{\beta}_0 + \sum_{l\neq j} x_{il}\tilde{\boldsymbol{\beta}}_l$ la valeur prédite sans la contribution de x_{ij} et égalisons l'équation (2.7) à 0 :

$$\frac{1}{n}\sum_{i=1}^{n} \tilde{w}_i x_{ij}(\tilde{z}_i - \tilde{y}_i^{(j)} - x_{ij}\beta_j) + \lambda = 0.$$

Nous obtenons

$$\beta_j = \frac{\frac{1}{n}\sum_{i=1}^{n} \tilde{w}_i x_{ij}(\tilde{z}_i - \tilde{y}_i^{(j)}) - \lambda}{\frac{1}{n}\sum_{i=1}^{n} \tilde{w}_i x_{ij}^2}.$$

Considérons maintenant le cas où β_j est négatif. Nous avons

$$R_{neg}(\beta_j) = -l_Q(\beta_0,\beta) - \lambda\beta_j.$$

En égalisant la dérivée de R_{neg} par rapport à β_j à 0, nous trouvons que

$$\beta_j = \frac{\frac{1}{n}\sum_{i=1}^{n} \tilde{w}_i x_{ij}(\tilde{z}_i - \tilde{y}_i^{(j)}) + \lambda}{\frac{1}{n}\sum_{i=1}^{n} \tilde{w}_i x_{ij}^2}.$$

De manière générale, pour β_j positif ou négatif, nous pouvons écrire

$$\begin{aligned}\beta_j &= \frac{\frac{1}{n}\sum_{i=1}^{n}\tilde{w}_i x_{ij}(\tilde{z}_i - \tilde{y}_i^{(j)}) + \lambda\ \text{sign}(\beta_j)}{\frac{1}{n}\sum_{i=1}^{n}\tilde{w}_i x_{ij}^2} \\ &= \frac{\text{sign}\left(\frac{1}{n}\sum_{i=1}^{n}\tilde{w}_i x_{ij}(\tilde{z}_i - \tilde{y}_i^{(j)})\right)\left(\left|\frac{1}{n}\tilde{w}_i x_{ij}(\tilde{z}_i - \tilde{y}_i^{(j)})\right| - \lambda\right)_+}{\frac{1}{n}\sum_{i=1}^{n}\tilde{w}_i x_{ij}^2} \\ &= \frac{S\left(\frac{1}{n}\sum_{i=1}^{n}\tilde{w}_i x_{ij}(\tilde{z}_i - \tilde{y}_i^{(j)}), \lambda\right)}{\sum_{i=1}^{n}\tilde{w}_i x_{ij}^2},\end{aligned}$$

où $S(\delta, \gamma)$ est un opérateur avec la valeur

$$S(\delta,\gamma) = \text{sign}(\delta)(|\delta| - \gamma)_+ = \begin{cases} \delta - \gamma & \text{si } \delta > 0 \text{ et } \gamma < |\delta| \\ \delta + \gamma & \text{si } \delta < 0 \text{ et } \gamma < |\delta| \\ 0 & \text{si } \gamma \geq |\delta|. \end{cases}$$

Avantages et désavantages du LASSO

La pénalité l_1 a pour effet la création d'une matrice creuse (ou parcimonieuse), c'est-à-dire la création d'une solution composée de coefficients égaux à 0 exactement. Les variables explicatives dont leur coefficient est égal à 0 ne sont pas associées à la variable réponse. Par exemple, si le coefficient d'un gène est égal à 0, alors ce gène n'a pas de relation avec la maladie. Au contraire, un gène qui a un coefficient différent de 0 est lié à la maladie. Le principal avantage du LASSO est donc qu'il permet la sélection de variables.

L'inconvénient du LASSO par rapport à la régression Ridge est que les variables corrélées ne tendent pas nécessairement vers une même valeur. Le LASSO ne tient pas en considération la structure des variables. Par exemple, deux variables faisant partie du même chemin biologique n'auront pas nécessairement le même coefficient. Le LASSO a tendance à ne choisir qu'une seule variable explicative

parmi celles corrélées et à ignorer toutes les autres (Friedman *et al.*, 2010). De plus, le LASSO sélectionne au plus n variables lorsque $p > n$ (Zou et Hastie, 2005) et le nombre de coefficients égaux à 0 est très grand comparativement à ceux différents de 0 (Friedman *et al.*, 2010).

2.2.4 Méthode groupe LASSO

Le groupe LASSO est une méthode de réduction qui a été développée dans le but de remédier au problème de la corrélation dans la régression LASSO. Le groupe LASSO tient compte des groupes de prédicteurs a priori connus afin d'attribuer des coefficients égaux à toutes les covariables d'un groupe. Par exemple, si 25 gènes font partie du même chemin biologique (même groupe), alors ils auront les mêmes coefficients. Ils seront tous différents de 0 ou exactement égaux à 0 selon s'ils prédisent bien ou non la réponse (le fait d'être malade ou non).

L'estimateur des paramètres obtenu à la suite d'une régression logistique pénalisée basée sur le groupe LASSO est

$$\left(\hat{\beta}_0, \hat{\boldsymbol{\beta}}\right) = \underset{(\beta_0,\boldsymbol{\beta})}{\operatorname{argmin}} \left\{ -\sum_{i=1}^{n} [y_i \log \pi_i + (1 - y_i) \log(1 - \pi_i)] + \lambda \sum_{l=1}^{L} \sqrt{p_l} ||\boldsymbol{\beta}_l||_2 \right\}$$

où L est le nombre de groupes de variables explicatives, p_l est le nombre de covariables dans le groupe l, $\mathbf{X}_l$ est la matrice des prédicteurs dans le groupe l, $\boldsymbol{\beta}_l$ représente le vecteur des coefficients du groupe l et $||.||_2$ est la norme euclidienne (Hastie *et al.*, 2009).

La pénalité est donc sous la forme

$$P(\boldsymbol{\beta}) = \sum_{l=1}^{L} \sqrt{p_l} ||\boldsymbol{\beta}_l||_2.$$

Le désavantage de la méthode groupe LASSO est qu'elle nécessite de connaître

les groupes avant même de l'appliquer. Si nous ne connaissons pas les ensembles de prédicteurs préalablement, nous ne pouvons pas utiliser cette approche.

2.2.5 Méthode Elastic Net

L'Elastic Net, une méthode de régularisation introduite par Zou et Hastie (2005), est un mélange du LASSO et de la régression Ridge. Cette technique permet d'allouer un 0 exact aux coefficients et de donner des valeurs similaires aux prédicteurs corrélés, et ce, sans connaître les groupes de prédicteurs a priori. La pénalité de l'Elastic Net est une combinaison des deux pénalités vues précédemment (la norme l_1 et la norme l_2) :

$$\begin{aligned} P(\boldsymbol{\beta}) &= (1-\alpha)\frac{1}{2}||\boldsymbol{\beta}||_2^2 + \alpha||\boldsymbol{\beta}||_1 \\ &= \sum_{j=1}^{p}\left[\frac{1}{2}(1-\alpha)\beta_j^2 + \alpha|\beta_j|\right]. \end{aligned}$$

L'Elastic Net nous permet de résoudre le problème suivant :

$$\min_{(\beta_0,\boldsymbol{\beta})} \{L(\mathbf{X},\mathbf{y}) + \lambda P(\boldsymbol{\beta})\}.$$

Pour estimer les paramètres $(\beta_0, \boldsymbol{\beta})$, Jerome Friedman utilise la méthode de la descente par coordonnée (Friedman *et al.*, 2010). Ainsi, nous obtenons

$$\beta_j = \frac{S\left(\frac{1}{n}\sum_{i=1}^{n}\tilde{w}_i x_{ij}(\tilde{z}_i - \tilde{y}_i^{(j)}), \lambda\alpha\right)}{\sum_{i=1}^{n}\tilde{w}_i x_{ij}^2 + \lambda(1-\alpha)} \tag{2.8}$$

en utilisant l'approximation quadratique de la log-vraisemblance comme fonction de perte.

Nous remarquons que l'équation (2.8) est équivalente à celle du LASSO lorsque $\alpha = 1$ et à celle du Ridge lorsque $\alpha = 0$. Nous avons montré comment estimer les

paramètres pour la régression Ridge à partir de l'algorithme de Newton-Raphson, mais il est également possible de les estimer par la méthode de la descente par coordonnée. En fait, il est préférable d'utiliser la méthode de la descente par coordonnée pour estimer les paramètres, car inverser la matrice $(\mathbf{X}^T\hat{\mathbf{W}}\mathbf{X}+2\lambda\mathbf{I})^{-1}$ peut s'avérer très coûteux en temps de calcul.

Un autre avantage de l'Elastic Net est qu'elle permet d'avoir plus de n coefficients différents de 0 lorsque $p > n$. Ceci nous permet donc de trouver au-delà de n gènes significatifs.

Peu importe la méthode choisie entre la régression Ridge, le LASSO, le groupe LASSO ou l'Elastic Net, les valeurs des paramètres de réglage λ et α sont obtenues à l'aide de validations croisées (Zou et Hastie, 2005). Nous retenons celles associées à la plus petite erreur de prédiction.

2.2.6 Analyse discriminante pénalisée basée sur la fonction de Fisher

La méthode « Penalized LDA-L1 » est une autre méthode permettant de résoudre le problème de la singularité de la matrice $\mathbf{X}$ lorsque $p > n$. Cette méthode, développée par Daniela M. Witten et Robert Tibshirani, vise à projeter les observations dans une plus petite dimension en maximisant la variance inter-classe par rapport à la variance intra-classe en se basant sur la fonction discriminante de Fisher (Witten et Tibshirani, 2011). Cette approche permet d'obtenir un problème biconvexe qui peut être optimisé grâce à un algorithme de minimisation-maximisation.

Witten et Tibshirani (2011) définissent la méthode « Penalized LDA-L1 » comme ceci :

$$\max_{\boldsymbol{\beta_k}} \left\{ \boldsymbol{\beta_k}^T \hat{\Sigma}_b^k \boldsymbol{\beta_k} - \lambda_k \sum_{j=1}^{p} |\hat{\sigma}_j \boldsymbol{\beta_{kj}}| \right\} \text{ sous la contrainte } \boldsymbol{\beta_k} \tilde{\Sigma}_w \boldsymbol{\beta_k} \leq 1, \qquad (2.9)$$

où

$$\hat{\Sigma}_b^k = \frac{1}{n}\mathbf{X}^T\mathbf{Y}(\mathbf{Y}^T\mathbf{Y})^{-1/2}\mathbf{P}_k^{\perp}(\mathbf{Y}^T\mathbf{Y})^{-1/2}\mathbf{Y}^T\mathbf{X}.$$

$\hat{\Sigma}_b^k$ est l'estimateur de la matrice inter-classe avec $k = 1, \ldots, K-1$ et K est le nombre de groupes d'individus, $\hat{\sigma}_j$ est l'écart-type intra-classe du prédicteur j, $\tilde{\Sigma}_w$ est la matrice estimée définie positive de Σ_w et $\mathbf{P}_k^{\perp}$ est la matrice de projection orthogonale.

Le vecteur $\hat{\boldsymbol{\beta}}_k$ obtenu est parcimonieux et facilement interprétable. Par exemple, si $K = 2$, nous n'avons qu'un seul vecteur composé de valeurs égales à 0 ou différentes de 0. Si les valeurs sont égales à 0, alors les prédicteurs associés ne discriminent pas bien les deux groupes et ils ne sont pas associés à la variable réponse.

2.3 Apprentissage non supervisé

Les méthodes vues jusqu'à présent nécessitent de connaître le groupe de chacun des individus pour vérifier l'association entre la matrice $\mathbf{X}$ et le vecteur $\mathbf{y}$ et pour faire de la prédiction. Nous allons maintenant voir deux méthodes dites non supervisées qui ne regardent que la structure de $\mathbf{X}$. Elles ne tiennent pas compte de $\mathbf{y}$.

2.3.1 Méthode hiérarchique

La méthode hiérarchique, aussi connue sous le nom d'« analyse en cluster », permet de réunir des variables en un nombre inconnu de groupes homogènes en se basant sur la structure de $\mathbf{X}$. Selon ce que nous désirons étudier, la méthode hiérarchique regroupe les lignes de la matrice $\mathbf{X}$ (les individus) ou les colonnes de $\mathbf{X}$ (les

prédicteurs). Cette méthode est utile pour l'étude de maladies complexes, car elle permet de classer les gènes en fonction de la structure de la matrice $\mathbf{X}$ sans connaître le nombre de groupes.

La méthode hiérarchique est une technique itérative qui implique une fusion de variables ou de groupes de variables à chaque étape. Au départ, nous partons avec p groupes (objets, gènes). Nous fusionnons deux groupes qui sont très similaires pour ne former qu'un seul groupe. Par la suite, nous fusionnons deux autres groupes, puis encore deux autres groupes, jusqu'à ce que nous ne nous retrouvons qu'avec un seul groupe contenant tout le jeu de données.

Pour quantifier la similarité des groupes, plusieurs mesures de distance sont à notre disposition. Une parmi elles est la distance « moyenne » donnée par

$$D(A,B) = \frac{1}{n_A n_B} \sum_{i=1}^{n_A} \sum_{j=1}^{n_B} d(\mathbf{x}_i, \mathbf{x}_j), \tag{2.10}$$

où A et B sont deux classes distinctes, n_A et n_B indiquent le nombre d'objets dans les classes A et B respectivement et $d(\mathbf{x}_i, \mathbf{x}_j)$ est une distance (par exemple, euclidienne ou de Mahalanobis) (Rencher, 2002). Les classes qui ont la plus petite distance sont groupées ensemble. L'avantage de cette mesure est qu'elle n'est pas sensible aux valeurs aberrantes.

Il est à noter que la méthode hiérarchique ne permet pas à elle seule de réduire le nombre de covariables. Son but est seulement de regrouper les prédicteurs selon leur similarité.

Dendogramme

La fusion des groupes peut être représentée par un graphique appelé « dendogramme ». Chacune des fusions est illustrée à l'aide d'une ligne horizontale. La

figure (2.3) est un exemple de dendogramme obtenu à la suite de l'application de la méthode hiérarchique sur 500 gènes. En dessous du diagramme, une bande horizontale de couleurs indique le groupe auquel chacun des gènes appartient. Dans cet exemple, les 500 gènes sont répartis en 6 groupes : gris, jaune, brun, bleu, turquoise et vert.

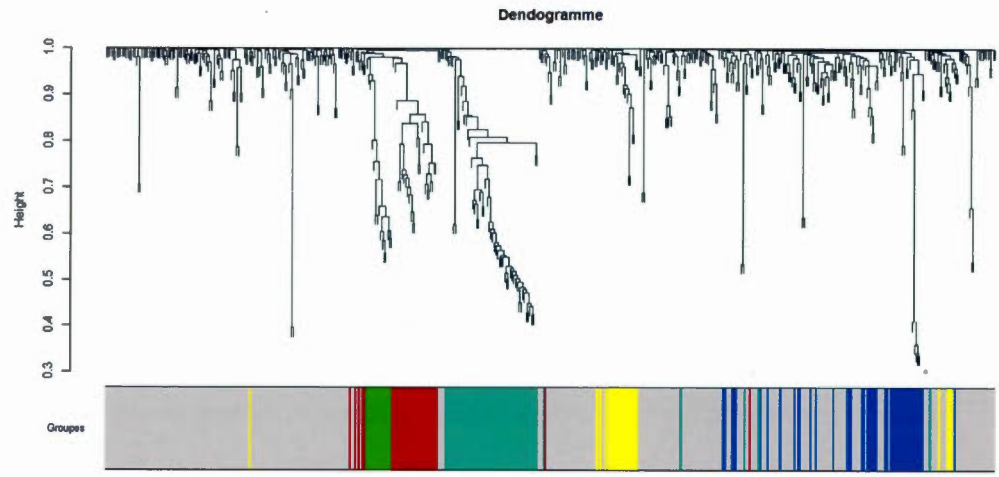


Figure 2.3: Exemple d'un dendogramme

2.3.2 K-moyennes

La méthode des « K-moyennes » est une méthode d'apprentissage non supervisé permettant de partitionner p objets dans k groupes distincts (Wu, 2012). Chaque groupe est représenté par son centroïde, c'est-à-dire par la moyenne des objets présents à l'intérieur du groupe . L'algorithme des K-moyennes fonctionne comme suit :

1. Déterminer le nombre de groupes désirés ;
2. Choisir au hasard K objets (parmi les p objets) qui serviront de centroïdes initiaux ;

3. Assigner chacun des p objets au centroïde le plus près (en utilisant la distance euclidienne, par exemple) ;
4. Recalculer le centroïde de chacun des groupes créés ;
5. Refaire les étapes 3 et 4 jusqu'à ce que les objets ne changent plus de groupe.

Cette méthode comporte deux inconvénients majeurs pour le partitionnement des gènes dans des groupes distincts. Premièrement, cette méthode nécessite de connaître préalablement le nombre de groupes. Deuxièmement, tout comme la méthode hiérarchique, la méthode des K-moyennes ne permet pas de faire de la réduction de variables.

2.4 Solution possible aux problèmes rencontrés

Dans ce chapitre, nous avons d'abord étudié la régression logistique et l'analyse discriminante. Nous avons vu que ces deux méthodes nous permettent de faire de la prédiction et de tester l'association entre des variables. Cependant, elles ne sont applicables qu'à des jeux de données de petite dimension. Nous avons ensuite vu les méthodes de régularisation telles la régression Ridge, la méthode LASSO, la méthode groupe LASSO, la méthode Elastic Net et l'analyse discriminante pénalisée basée sur la fonction de Fisher. Celles-ci sont utiles pour faire de la prédiction et de la sélection de variables, toutefois elles ne sont pas faites pour mesurer l'association entre des variables.

Steve Horvath (2011) a proposé une méthode, appelée « Weighted Gene Co-Expression Network Analysis », qui permet de regrouper les gènes en fonction de leur interconnectivité. Autrement dit, elle permet d'analyser les prédicteurs ensemble afin de capter les structures de groupe. Cette approche, jumelée avec des méthodes classiques ou de régularisation, nous permettra de trouver les groupes de gènes qui sont associés à une maladie complexe et de faire de la prédiction.

CHAPITRE III

APPROCHE « WEIGHTED GENE CO-EXPRESSION NETWORK ANALYSIS »

La méthode « Weighted Gene Co-expression Network Analysis » (WGCNA) est une approche permettant de capter la co-expression entre les gènes. En d'autres mots, elle permet de décrire la relation entre les gènes en se basant sur leur profil d'expression. Cette méthode nous permet de regrouper les gènes en fonction de leurs corrélations avec les autres gènes. Elle est également utile pour trouver les gènes qui ont une grande influence sur les autres gènes, c'est-à-dire ceux qui sont très connectés aux autres. Ces gènes à grande connectivité sont communément appelés « hub ».

3.1 Définitions de quelques matrices

Pour trouver les modules (groupes) importants et les gènes à grande connectivité, nous devons d'abord définir quelques matrices : matrice de similarité, matrice d'adjacence et matrice de chevauchement topologique.

3.1.1 Mesure de similarité

La mesure de similarité est définie comme étant le degré de concordance entre les profils d'expression des gènes pour toutes les paires de gènes possibles (Zhang et Horvath, 2005). Soit $\mathbf{x}_i$ le profil d'expression du gène i et $\mathbf{x}_j$ le profil d'expression du gène j. Ces vecteurs sont de dimension n x 1, où n représente le nombre d'individus dans l'étude. Les éléments de la matrice de similarité s'obtiennent en prenant la valeur absolue de la corrélation de Pearson entre les profils d'expression des gènes i et j :

$$s_{ij} = |cor(\mathbf{x}_i, \mathbf{x}_j)|.$$

Si nous voulons tenir compte du signe de la corrélation, nous utiliserons plutôt la mesure de similarité avec signe

$$s_{ij} = \frac{1 + cor(\mathbf{x}_i, \mathbf{x}_j)}{2}.$$

La matrice de similarité $S = [s_{ij}]$ est de dimension $p \times p$, où p représente le nombre de gènes.

La matrice de similarité telle que définie ci-haut n'est applicable que pour les graphes non orientés puisque la matrice est symétrique. Si nous sommes en présence d'un réseau de voies biologiques orienté, alors il est préférable d'utiliser une approche basée sur un modèle à équations structurelles (Horvath, 2011).

3.1.2 Matrice d'adjacence

La matrice d'adjacence $\mathbf{A} = [a_{ij}]$ s'obtient à partir de la matrice de similarité à la suite d'une transformation. Pour choisir la bonne transformation, nous devons premièrement déterminer le type de réseau que nous voulons : pondéré ou non

pondéré.

Dans un réseau non pondéré, la matrice d'adjacence peut s'obtenir de la manière suivante :

$$a_{ij} = \begin{cases} 1 & \text{si } s_{ij} \geqslant \tau \\ 0 & \text{si } s_{ij} < \tau. \end{cases}$$

Lorsque la corrélation est supérieure ou égale à un certain τ, l'élément a_{ij} est égal à 1 et il y a un lien entre les gènes i et j. Ceci signifie qu'il y a une connexion (un chemin) entre ces deux gènes. Par contre, lorsque la corrélation est inférieure à ce τ, l'élément a_{ij} est égal à 0 et il n'y a aucune connexion entre les deux gènes. Le passage de la matrice de similarité à la matrice d'adjacence est une transformation par seuillage dur, car il n'y a que deux possibilités : soit il y a une connexion, soit il n'y en a pas. Peu importe que l'élément s_{ij} soit très près du paramètre τ mais inférieur à celui-ci ou qu'il soit égal à 0, l'élément a_{ij} est égal à 0. C'est donc un gros désavantage, car il y a une perte d'information.

Dans un réseau pondéré, nous pouvons utiliser la fonction suivante :

$$a_{ij} = s_{ij}^{\beta},$$

avec $\beta \geq 1$. Ainsi, les éléments a_{ij} prennent des valeurs entre 0 et 1 et ceux sur la diagonale prennent des valeurs égales à 1 (Horvath, 2011). Le passage de la matrice de similarité à la matrice d'adjacence dans le cas pondéré est une transformation par seuillage doux puisque si deux gènes sont très peu corrélés, ils sont quand même connectés ensemble par un chemin. Le désavantage avec le seuillage doux est que les bruits associés à la collecte des données restent présents dans notre réseau de gènes. Par contre, en calculant s_{ij} puissance β, les bruits ont des a_{ij} qui tendent vers 0. Un grand avantage du réseau pondéré par rapport au réseau non pondéré est que le réseau pondéré garde la continuité des profils d'expression des gènes. Nous prioriserons donc un réseau pondéré. La matrice d'adjacence pour un

réseau pondéré représente la force des connexions. La valeur de β est déterminée par le critère du « réseau invariant d'échelle ». Nous verrons comment l'obtenir dans la section 3.2.3.

3.1.3 Matrice de chevauchement topologique

La matrice de chevauchement (Topological Overlap Matrix), notée TOM, est une matrice contenant de l'information sur l'interconnectivité des gènes. Elle est définie comme suit :

$$w_{ij} = \frac{l_{ij} + a_{ij}}{\min(k_i, k_j) + 1 - a_{ij}},$$

où $l_{ij} = \sum_{u \neq i,j} a_{iu} a_{uj}$ et $k_i = \sum_u a_{iu}$ (Zhang et Horvath, 2005).

Lorsque nous sommes dans le cas d'un réseau non pondéré, k_i représente le nombre de connexions au gène i et l_{ij} le nombre de gènes qui sont à la fois connectés au gène i et au gène j. On appelle « voisin » un gène qui est connecté au gène dont nous faisons l'analyse. Ainsi, si $w_{ij} = 0$, cela signifie que les gènes i et j ne sont pas connectés et qu'ils n'ont pas de voisin en commun. Au contraire, si $w_{ij} = 1$, alors les gènes i et j sont connectés et le gène qui a le plus de voisins entre i et j est connecté à tous les voisins de l'autre gène.

Pour un réseau pondéré, k_i représente plutôt la force de connectivité. Nous pouvons interpréter les éléments w_{ij} de manière similaire que dans le cas non pondéré. Par exemple, si w_{ij} est égal à 0, alors les gènes i et j ne sont pas connectés et ils n'ont pas de voisin en commun (ils ne sont pas connectés aux mêmes gènes). De plus, w_{ij} prend des valeurs entre 0 et 1.

Il est à noter que la corrélation (s_{ij}) et le chevauchement topologique (w_{ij}) sont deux quantités bien différentes. La corrélation permet de mesurer l'association entre deux gènes isolés et le chevauchement topologique permet de mesurer l'as-

sociation entre deux gènes en tenant compte des autres gènes présents dans le réseau.

3.2 De la théorie des graphes aléatoires aux réseaux invariants d'échelle

3.2.1 Graphes aléatoires

Paul Erdos et Alfréd Rényi ont introduit la théorie des graphes aléatoires pour expliquer les réseaux topologiques complexes à la fin des années 1950. Leur modèle suppose que les noeuds sont connectés aléatoirement avec probabilités égales. Pour construire un graphe aléatoire, ils commencent avec N noeuds et ils les connectent avec probabilité p. Un noeud est de degré k s'il est relié à k autres des $N-1$ noeuds du graphe et cela arrive avec une probabilité

$$P(k) = \binom{N-1}{k} p^k (1-p)^{N-1-k} \underset{N\to\infty}{\sim} \frac{e^{-\lambda}\lambda^k}{k!}.$$

La distribution des degrés $P(k)$ suit une loi de Poisson de paramètre $\lambda \approx Np$ lorsque $N \to \infty$. La majorité des noeuds ont un degré k près de la moyenne des degrés (Albert, 2005). Le modèle d'Erdos et de Rényi suppose que la probabilité de trouver un noeud avec beaucoup de connexions décroît exponentiellement avec k et que les noeuds avec une connectivité élevée sont rares (Barabási et Albert, 1999).

3.2.2 Réseau invariant d'échelle

Le problème du modèle de Paul Erdos et d'Alfréd Rényi est qu'il ne représente pas bien la réalité des années 2000 avec les jeux de données de grande dimension et la complexité des structures des réseaux. Les réseaux complexes (biologiques ou autres) sont caractérisés par un degré d'auto-organisation élevé qui est traduit

par une loi de puissance :

$$P(k) \sim k^{-\gamma},$$

où $P(k)$ est la proportion de noeuds qui ont k connexions.

Le modèle suivant une loi de puissance, appelé modèle invariant d'échelle, est basé sur deux mécanismes (Barabási et Albert, 1999) :

(i) la croissance continuelle du réseau ;

(ii) l'« attachement préférentiel » des noeuds.

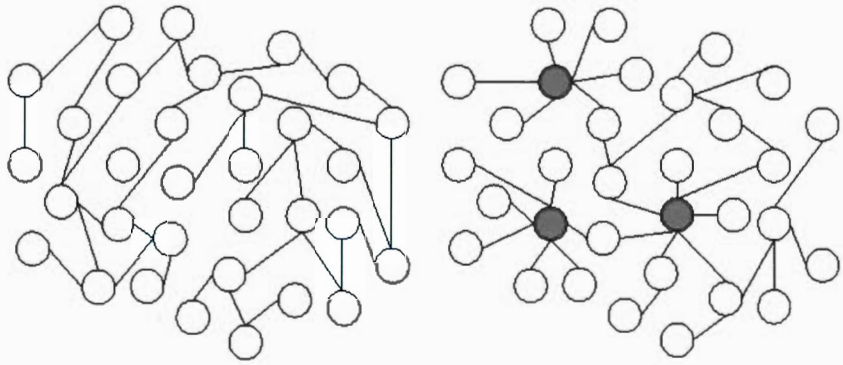
Le point (i) signifie qu'un réseau complexe est continuellement en expansion. De nouveaux noeuds et de nouveaux liens s'ajoutent aux autres déjà en place tout au long de la vie du réseau. Ceci s'oppose à l'idée d'un graphe aléatoire qui commence avec N noeuds et qui n'augmente jamais ce nombre. Le point (ii) suggère que la probabilité de connexion des nouveaux noeuds n'est pas uniforme. Un noeud qui est très connecté a tendance à se lier aux nouveaux noeuds tandis qu'un noeud avec une faible connectivité n'a qu'une petite probabilité de se lier aux nouveaux noeuds. Encore une fois, ce concept va à l'encontre de celui d'un réseau aléatoire qui suppose des connexions aléatoires et uniformes.

Un réseau suivant une loi de puissance est appelé « réseau invariant d'échelle », car sa fonction de puissance indique qu'il y a une diversité élevée des degrés des noeuds et qu'aucun noeud ne peut être utilisé comme référence pour les autres noeuds (Albert, 2005). Ainsi, s'il n'y a aucun degré typique ou aucune échelle caractéristique, nous disons qu'il y a invariance d'échelle.

Un tel modèle a ses avantages et ses désavantages. D'un côté, le réseau est plus stable et plus robuste face à un noeud défectueux. Puisqu'il y a beaucoup de noeuds à faible connectivité, il y a peu de chance qu'un « hub » soit endommagé. Ainsi, la perturbation ne cause pas une grosse perte de connectivité. D'un autre côté, si un noeud à haute connectivité (« hub ») est endommagé, alors le réseau

sera détruit en plusieurs sous-graphes isolés.

La figure (3.1) illustre un réseau aléatoire et un réseau invariant d'échelle. Dans la figure de droite, les points noirs représentent les hubs, c'est-à-dire les noeuds à connectivité élevée.



(a) Réseau aléatoire (b) Réseau invariant d'échelle

Figure 3.1: Deux types de réseaux

3.2.3 Choix de bêta

Nous avons dit précédemment qu'un réseau invariant d'échelle est un réseau dont les degrés de distribution suivent une loi de puissance. Nous pouvons donc écrire $P(k) \sim k^{-\gamma}$ sous la forme

$$P(k) = c \cdot k^{-\gamma}, \tag{3.1}$$

où c est une constante. Dans notre cas, puisque le nombre de connexions (k) dépend de la valeur de β, nous allons réécrire l'équation (3.1) comme ceci :

$$P(k(\beta)) = c \cdot k(\beta)^{-\gamma}. \tag{3.2}$$

En traçant le graphique de $\log[P(k(\beta))]$ en fonction de $\log[k(\beta)]$, nous pouvons vérifier si notre réseau suit une loi de puissance. Si c'est le cas, nous devrions voir

des points alignés le long d'une droite, puisque

$$\log[P(k(\beta))] = \log(c) - \gamma \log[k(\beta)].$$

Pour le choix de β, nous gardons celui associé aux données qui s'ajustent le mieux à une droite.

Selon Zhang et Horvath (2005), γ devrait être environ égal à 1 et, selon Albert (2005), γ devrait être entre 2 et 3. Ces valeurs ont été trouvées à la suite d'expérimentations, mais l'idée générale est que le paramètre γ contrôle le nombre de liens entre les gènes. En effet, si γ diminue, le nombre de liens augmente et si γ augmente, le nombre de liens diminue. Par conséquent, une valeur γ élevée entraîne un réseau constitué de petits groupes isolés et un graphe déconnecté. Au contraire, une valeur très petite de γ implique un très grand groupe et donc un graphe complètement connecté (Albert et Barabasi, 2002).

Zhang et Horvath (2005) proposent un autre critère topologique pour vérifier l'invariance d'échelle et pour choisir la valeur de β. Ils suggèrent de calculer le R^2 provenant d'un modèle de régression linéaire :

$$R^2 = \text{cor}\left\{\log[P(k(\beta))], \log[k(\beta)]\right\}^2.$$

Lorsque R^2 est environ égal à 1, le réseau est invariant d'échelle. Nous voulons donc le β qui permet d'avoir un R^2 près de 1. Il faut cependant faire attention à ceci : plus le R^2 est élevé, moins il y a de connexions entre les noeuds. Par conséquent, il faut choisir une valeur de β pour laquelle le R^2 associé est assez grand et pour laquelle la connectivité moyenne est aussi assez grande.

Il arrive parfois que la pente de la régression soit positive ce qui signifie qu'il y a davantage de « hub » que de gènes à faible connectivité. Au point de vue biologique, cela n'a pas de sens, alors Zhang et Horvath ont proposé un critère

topologique avec signe :

$$R_s^2 = \begin{cases} R^2 & \text{si } \gamma > 0 \\ -R^2 & \text{si } \gamma < 0. \end{cases}$$

Nous choisissons le β qui nous permet d'avoir un R_s^2 supérieur à 0.8.

Certains réseaux n'ont qu'approximativement les propriétés d'invariance d'échelle. Dans ce cas, un modèle avec une loi de puissance tronquée exponentiellement peut être plus adéquat :

$$P(k(\beta)) \sim k(\beta)^{-\gamma} \exp\{-\alpha k(\beta)\}.$$

Toutefois, Zhang et Horvath (2005) suggèrent quand même d'utiliser le critère avec signe (R_s^2) plutôt que celui du R^2 tronqué puisque la loi de puissance tronquée exponentiellement est trop flexible et peut mener à un mauvais choix de β.

3.3 Création des modules

Le but est de regrouper les gènes qui sont interconnectés dans un même module (groupe) pour ultimement trouver le ou les modules qui sont associés à la maladie (ou à une certaine caractéristique). Pour y arriver, nous appliquons la méthode hiérarchique telle que décrite dans le chapitre 2. Toutefois, la distance $d(\mathbf{x}_i, \mathbf{x}_j)$ de l'équation (2.10) est remplacée par $d_{ij} = 1 - w_{ij}$, la matrice de dissimilarité.

Pour chacun des modules créés, nous calculons la première composante principale qui est en fait une sorte de moyenne pondérée des profils d'expression (Langfelder et Horvath, 2008b). Ces vecteurs ainsi obtenus seront utilisés pour définir les modules qui sont significatifs.

Une fois les composantes principales trouvées, Steve Horvath propose de fusionner les modules dont les composantes principales ont une corrélation supérieure à 0.75

puisque les gènes qui s'y retrouvent ont une grande interconnectivité.

Pour vérifier quels sont les modules significatifs parmi ceux nouvellement créés, Peter Langfelder et Steve Horvath mesurent la corrélation de Pearson entre la variable $\mathbf{y}$ et chacune des composantes principales et ils calculent les valeurs-p du test du Khi-Deux de Pearson. Ceci permet d'éliminer tous les gènes faisant partie des groupes non significatifs et par le fait même de réduire le nombre de covariables.

3.4 Concepts fondamentaux en WGCNA

3.4.1 Importance d'un gène

Une fois les modules significatifs sélectionnés, nous pouvons calculer la relation entre un profil d'expression $\mathbf{x}_i$ d'un gène d'un module significatif et la caractéristique y (la maladie). Cette mesure est appelée « importance d'un gène » (« gene significance ») et elle est définie par :

$$GS_i = |\text{cor}(\mathbf{x}_i, y)|.$$

Ainsi, si la corrélation est forte, le gène i a un effet sur la maladie. Inversement, si la corrélation est nulle ou très faible, le gène i n'a aucun lien avec la maladie.

3.4.2 Connectivité standardisée

Il est également intéressant de calculer la connectivité standardisée d'un gène afin de savoir s'il est fortement connecté dans le réseau. La connectivité standardisée du gène i est définie par

$$K_i = \frac{k_i}{k_{max}},$$

où $k_i = \sum_{j=1}^{p} a_{ij}$ est la connectivité du gène i et k_{max} est la connectivité maximale parmi les p gènes (Horvath, 2011). Si K_i est près de 1, alors le gène i a une connectivité élevée relativement aux autres gènes.

3.5 Recommandation

Peter Langfelder et Steve Horvath recommandent de ne pas utiliser la « Weighted Gene Co-expression Network Analysis » si le nombre d'individus est inférieur à 15, car la corrélation contiendrait trop de bruit et les résultats ne seraient pas biologiquement parlant significatifs. Il est donc préférable que n soit relativement grand pour avoir une meilleure interprétation biologique et une meilleure robustesse.

3.6 Avantages et désavantages du WGCNA

Le principal avantage de l'approche WGCNA est que nous n'avons pas besoin de connaître le nombre et la taille des groupes avant d'effectuer notre analyse.

La méthode WGCNA comporte cependant quelques désavantages. En effet, le nombre de groupes est influencé par plusieurs composantes :

- le choix d'un seuillage doux ou d'un seuillage dur pour la construction de la matrice d'adjacence ;
- la valeur de β lors de la construction de la matrice d'adjacence ;
- la hauteur de découpage du dendogramme lors de la formation des groupes.

Ceci nous amène donc à nous demander quelle méthode parmi celles vues aux chapitres 2 et 3 est la meilleure pour trouver les groupes de gènes et/ou les gènes qui ont le plus d'impact sur la maladie et quelle méthode donne les meilleures prédictions. C'est ce que nous testerons au chapitre 4 grâce à des simulations.

CHAPITRE IV

COMPARAISONS DE DIVERSES MÉTHODES D'APPRENTISSAGE STATISTIQUE À L'AIDE DE SIMULATIONS

Jusqu'à présent, nous avons vu des méthodes standard et de régularisation au chapitre 2 et la méthode WGCNA au chapitre 3. Nous allons maintenant comparer ces approches à l'aide de simulations. En premier lieu, nous comparerons des méthodes en termes de puissance de tests. Nous testerons l'association globale entre les groupes de gènes et la variable réponse. Puisqu'aucune méthode de régularisation ne possède de test d'associativité, nous proposerons des approches en deux étapes qui combinent des méthodes de classification classiques (régression logistique et analyse discriminante linéaire) et de régularisation. Nous allons également jumeler des méthodes non supervisées (K-moyennes et WGCNA) avec des méthodes classiques. Pour toutes les approches proposées, la méthode utilisée à la première étape est vue comme une technique de réduction du nombre de variables et la méthode appliquée à la deuxième étape permet de mesurer et de tester l'association. En deuxième lieu, nous déterminerons la méthode avec le meilleur pouvoir prédictif. Puisque les méthodes de régularisation sont à la base destinées à faire de la prédiction, aucune méthode classique ne sera utilisée. Rappelons-nous que le but est de comparer les méthodes de régularisation qui exploitent la structure des groupes de gènes en se basant sur la corrélation de Pearson avec la méthode WGCNA qui exploite la structure des groupes de gènes en se basant

sur la connectivité. Nous allons donc comparer la méthode LASSO, l'Elastic Net, l'analyse discriminante pénalisée basée sur la fonction de Fisher et une version modifiée d'une technique proposée par Chou *et al.* (2014) basée sur la méthode WGCNA.

Dans la section (4.1), nous présenterons les différentes combinaisons de méthodes qui seront comparées. Par la suite, dans la section (4.2), nous décrirons les deux simulations qui ont été faites. Enfin, dans les sections (4.3) et (4.4), nous analyserons les résultats obtenus à la suite des tests d'association et de prédiction.

4.1 Combinaisons de méthodes

4.1.1 Approches pour l'associativité

Plusieurs approches ont été comparées dans le but de déterminer celle qui a la plus grande puissance. Nous avons d'abord jumelé la méthode WGCNA avec la régression logistique (approche WGCNA-Rlog). Ici, la méthode WGCNA est vue comme une technique de réduction de la dimension qui nous permet d'obtenir des scores (combinaisons linéaires de variables) et la régression logistique est une technique qui nous permet de tester l'associativité. La méthode WGCNA a d'abord été appliquée sur la matrice $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n]^T$, de dimension $n \times p$, pour créer des groupes de gènes à partir de leur profil d'expression. Ceci nous a permis d'obtenir une matrice $\mathbf{X}^* = [\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, \ldots, \mathbf{X}^{(k)}]$, avec $\mathbf{X}^{(j)}$ une matrice de dimension $n \times p_j$ contenant les expressions génétiques des p_j gènes faisant partie du groupe j et où $p = \sum_{j=1}^{k} p_j$. Par la suite, le premier vecteur propre $\mathbf{v}_j$ de la matrice $\mathbf{X}^{(j)T}\mathbf{X}^{(j)}$ et la première composante principale $\tilde{\mathbf{x}}_j = \mathbf{X}^{(j)}\mathbf{v}_j$ ont été calculés pour chacun des k groupes. Les scores $\tilde{\mathbf{x}}_j$ ainsi obtenus ont été utilisés dans la régression logistique. La régression logistique a été utilisée pour tester l'association entre les scores et la variable réponse à l'aide du test de rapport de vraisemblance. L'approche

WGCNA-RLog permet donc de réduire la dimension et d'obtenir une valeur-p qui pourra être comparée à celle des autres approches.

Nous avons fait de même pour la combinaison des méthodes K-moyennes et régression logistique (approche K-moyennes-Rlog). Nous avons commencé par appliquer la méthode des K-moyennes sur la matrice $\mathbf{X}$ pour créer des groupes de gènes. Nous avons alors obtenu une matrice $\mathbf{X}^* = [\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, \ldots, \mathbf{X}^{(k)}]$. À la différence avec la méthode WGCNA, le nombre k de groupes est déjà connu. En effet, le chercheur doit préalablement déterminer la valeur de k. Les vecteurs propres $\mathbf{v}_j$ et les premières composantes principales $\tilde{\mathbf{x}}_j$, pour $j = 1, \ldots, k$, ont ensuite été calculés. Enfin, une régression logistique a été effectuée sur les composantes principales et un test du rapport de vraisemblance a permis de vérifier la relation entre les composantes principales et la variable réponse.

Une autre approche que nous avons proposée est une combinaison de la méthode WGCNA avec l'analyse discriminante linéaire (approche WGCNA-ADL). Les gènes ont d'abord été regroupés selon leur profil d'expression grâce à la méthode WGCNA et une analyse discriminante a été appliquée sur les composantes principales $\tilde{\mathbf{x}}_j$. Par la suite, l'égalité des moyennes des composantes principales chez les cas et les témoins a été vérifiée via le test de Wilks. Les hypothèses étaient les suivantes :

$$H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \ldots = \boldsymbol{\mu}_k$$

$$H_1 : \text{Au moins un vecteur } \boldsymbol{\mu}_i \text{ est différent des autres,}$$

où i représente le module et k représente le nombre total de modules.

La méthode des K-moyennes et l'analyse discriminante linéaire ont également été combinées (approche K-moyennes-ADL). Nous avons procédé de la même manière que pour la méthode WGCNA, mais nous avons dû déterminer le nombre de groupes avant d'appliquer les techniques.

Un autre type de combinaison que nous avons essayé est celui composé d'une méthode non supervisée (WGCNA ou K-moyennes) et de la corrélation de Pearson (approches WGCNA-Cor et K-moyennes-Cor). Pour trouver les modules significatifs obtenus par la WGCNA, nous avons mesuré la corrélation entre la variable réponse y et chacune des composantes principales, $cor(\mathrm{y}, \tilde{\mathrm{x}}_j)$, et nous avons calculé les valeurs-p issues du test de Pearson. Cette manière de procéder avec la WGCNA a également été employée avec la méthode des K-moyennes. Puisqu'une valeur-p est associée à chacun des groupes, seule la valeur minimale des valeurs-p corrigées avec la méthode de Bonferroni a été utilisée pour vérifier l'existence d'une association globale :

$$\text{valeur-p}^* = k \cdot \min(\text{valeur-p}_j),$$

où valeur-p_j est la valeur-p du test de Pearson associée au groupe j, k est le nombre de modules et valeur-p^* est la valeur-p minimale corrigée avec la méthode de Bonferroni.

La prochaine approche que nous avons proposée est celle composée de la méthode LASSO et de la régression logistique (approche LASSO-Rlog). Nous avons commencé par trouver la valeur optimale de λ, c'est-à-dire la valeur de λ qui donne le meilleur modèle selon la validation croisée. Pour cette approche, nous avons appliqué la méthode LASSO directement sur les gènes et non sur des groupes de gènes. Ainsi, le modèle sélectionné est composé des profils d'expression des gènes retenus par la méthode LASSO. Nous avons ensuite effectué une régression logistique sur le modèle sélectionné et nous avons testé l'association entre les gènes retenus par la méthode LASSO (c'est-à-dire sur les gènes ayant des coefficients β différents de zéro dans la méthode LASSO) et la variable y. Pour ne pas utiliser les mêmes données deux fois, nous avons divisé les données en deux sous-échantillons : un pour sélectionner les gènes et un pour tester l'association. Lorsque le nombre de gènes retenus lors de la sélection du modèle était supérieur au nombre d'individus

utilisés dans le test d'association, nous avons dû réduire le nombre de gènes à $n_c - 1$, où n_c est le nombre de sujets utilisés pour la régression logistique. Pour choisir les gènes qui feraient partie du test d'associativité, nous avons classé les gènes selon leur coefficient et nous n'avons gardé que les $n_c - 1$ gènes avec les plus grands coefficients en valeur absolue. Le but de cette approche était de comparer la puissance des approches effectuées directement sur les gènes avec celles formant des groupes de gènes.

Enfin, nous avons jumelé la méthode Elastic Net avec la régression logistique (approche EN-Rlog). Une validation croisée a d'abord été faite sur le paramètre α de la pénalité et une autre sur le paramètre λ. Ensuite, la méthode Elastic Net a été appliquée directement sur les gènes. Une régression logistique a été effectuée sur les gènes ayant eu des coefficients différents de 0 et un test d'association avec la variable réponse a suivi. Encore une fois, les données ont été divisées en deux sous-échantillons pour contrôler le biais de l'erreur de type I : un groupe pour la sélection du modèle avec l'Elastic Net et un autre groupe pour le test d'associativité. Le nombre de gènes a également été réduit à $n_c - 1$ lorsque cela était nécessaire.

En résumé, nous avons comparé 8 approches :

1. WGCNA-Rlog
2. WGCNA-ADL
3. WGCNA-Cor
4. K-moyennes-Rlog
5. K-moyennes-ADL
6. K-moyennes-Cor
7. LASSO-Rlog
8. EN-Rlog.

4.1.2 Approches pour la sélection des variables et la prédiction

Un autre aspect auquel nous nous sommes intéressés est la prédiction. Puisque les méthodes de régularisation ont été élaborées dans le but de faire de la prédiction pour des jeux de grande dimension, nous les avons utilisées directement sans les combiner à des méthodes classiques. Nous avons comparé l'aire sous la courbe ROC des trois méthodes de régularisation suivantes : LASSO, Elastic Net et analyse discriminante pénalisée basée sur la fonction de Fisher.

Un autre objectif était d'évaluer la performance des trois méthodes mentionnées ci-haut avec des données simulées selon un modèle de régression standard $\mathbf{y} = \mathbf{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon}$ et avec des données simulées selon le modèle WGCNA qui crée des groupes de prédicteurs en se basant sur la notion de connectivité et de réseau. Afin d'évaluer la performance, nous avons comparé le nombre de faux positifs (FP), où

$$FP = \sum_{i=1}^{p} FP_i,$$

avec

$$FP_i = \begin{cases} 1 \text{ si } \hat{\beta}_i \neq 0 \text{ et } \beta_i = 0 \\ 0 \text{ sinon }, \end{cases}$$

le nombre de faux négatifs (FN), où

$$FN = \sum_{i=1}^{p} FN_i,$$

avec

$$FN_i = \begin{cases} 1 \text{ si } \hat{\beta}_i = 0 \text{ et } \beta_i \neq 0 \\ 0 \text{ sinon }, \end{cases}$$

et le taux de bonne parcimonie (BP), où

$$BP = \frac{1}{p} \sum_{i=1}^{p} A_i,$$

avec

$$A_i = \begin{cases} 1 & \text{si } \hat{\beta}_i = \beta_i = 0 \\ 1 & \text{si } \hat{\beta}_i \neq 0, \beta_i \neq 0 \\ 0 & \text{sinon,} \end{cases}$$

et p est le nombre de prédicteurs.

Les méthodes de régularisation énumérées ci-haut permettent de trouver les gènes qui sont utilisés pour faire de la prédiction, cependant elles ne permettent pas d'identifier les gènes à haute connectivité. C'est pourquoi nous nous sommes tournés vers une technique développée par Chou qui utilise à la fois la méthode WGCNA et l'Elastic Net pour sélectionner les « hubs » (Chou *et al.*, 2014). Nous avons toutefois légèrement modifié leur approche. Nous commençons par former des groupes de gènes avec la méthode WGCNA et par tester l'association entre chacun des groupes de gènes et la variable réponse grâce à la corrélation de Pearson. Puis, à l'intérieur de chacun des groupes de gènes significatifs à 5 %, nous calculons l'importance de chacun des gènes (GS, voir la section 3.4.1), la connectivité standardisée (K, voir la section 3.4.2) et la fréquence de prédiction (f). La fréquence de prédiction est obtenue via l'Elastic Net. Elle est définie comme suit : nous appliquons 1000 fois l'Elastic Net à chacun des modules significatifs et f_i est égal au nombre de fois parmi 1000 que $\hat{\beta}_i, i = 1, \ldots, p$, est différent de 0. Ainsi, nous pouvons écrire la fréquence de prédiction du gène i de la manière suivante :

$$f_i = \sum_{l=1}^{1000} c_l,$$

avec

$$c_l = \begin{cases} 1 & \text{si } \hat{\beta}_i \neq 0 \\ 0 & \text{sinon.} \end{cases}$$

Nous considérons des gènes comme étant des gènes à connectivité élevée, lorsque $K_i > 0.20$, $GS_i > 0.20$ et $f_i > 750$ simultanément. Nous allons comparer le

pouvoir prédicteur de cette approche avec celui de la méthode LASSO, de l'Elastic Net et de l'analyse discriminante basée sur la fonction de Fisher.

4.2 Scénarios

4.2.1 Scénario 1

Nous avons généré une matrice $\mathbf{X_{init}}$ de dimension $n_{init} \times p$ où $n_{init} = 2000$ individus et $p = 1000$ gènes. Les lignes de la matrice $\mathbf{X_{init}}$ ont été simulées de manière à suivre une loi normale multivariée $N_p(\mathbf{0}, \mathbf{\Sigma})$. La matrice $\mathbf{\Sigma}$ est une matrice par bloc avec des 1 sur sa diagonale, des valeurs entre 0.1 et 0.4 à l'intérieur de chaque bloc et des 0 pour tous les autres éléments. Les éléments des lignes et des colonnes 3 et 19 ont des valeurs entre 0.5 et 0.7 à l'intérieur de leur bloc, car les gènes 3 et 19 sont des gènes à haute connectivité. De plus, les dix premiers gènes forment un groupe, les dix suivants forment un second groupe et tous les autres gènes restants forment un troisième groupe. Le vecteur colonne β est de la forme :

$$\beta = \left(\underbrace{0.4, \ldots, 0.4}_{10}, \underbrace{-0.4, \ldots, -0.4}_{10}, \underbrace{0, \ldots, 0}_{980} \right)^T .$$

Nous avons ensuite simulé la variable dépendante $\mathbf{y_{init}}$ de telle sorte que $\mathbf{y_{init}} = \mathbf{X_{init}} * \beta + \epsilon$ où les éléments de ϵ proviennent d'une N(0,1). Nous voulions une variable réponse dichotomique, alors nous l'avons simulée à l'aide d'une binomiale de probabilité $1/(1 + exp(-\mathbf{y_{init}}))$. Finalement, nous avons sélectionné au hasard 75 individus avec $y_{init} = 1$ et 75 autres avec $y_{init} = 0$ pour former une étude de 150 individus. Nous avons donc le vecteur $\mathbf{y}$ composé des réponses des 150 individus. Nous avons également pris les lignes correspondantes dans la matrice $\mathbf{X_{init}}$ pour former la matrice $\mathbf{X}$ de dimension $n \times p$ où $n = 150$ et $p = 1000$.

4.2.2 Scénario 2

Pour ce scénario, nous avons simulé les données telles que décrites dans le tutoriel de Peter Langfelder et de Steve Horvath (Langfelder et Horvath, 2008a). Nous avons déterminé le nombre d'individus que nous voulions, soit 150 dont 75 sont malades (y=1) et 75 ne le sont pas (y=0). Nous avons ensuite déterminé le nombre de modules que nous voulions. Nous avons choisi qu'il y ait 6 modules en tout. Chacun des modules est représenté par une couleur : vert, brun, jaune, turquoise, bleu et gris. Nous avons commencé par simuler un vecteur vert $\mathbf{x}_{vert}$ de dimension $n \times 1$ de loi normale $N_n(\mathbf{0}, \mathbf{1})$ pour représenter la première composante principale du module vert. Par la suite, une variable réponse continue $\mathbf{y}_{cont}$ a ensuite été créée de manière à avoir une corrélation de $\alpha_{vert} = 0.25$ avec la composante principale du module vert :

$$\mathbf{y}_{cont} = \alpha_{vert}\mathbf{x}_{vert} + \boldsymbol{\varepsilon}\sqrt{1 - \alpha_{vert}^2},$$

où $\boldsymbol{\varepsilon} \sim N_n(\mathbf{0}, \mathbf{1})$. Ensuite, le vecteur $\mathbf{y}_{cont}$ a été transformé en une variable réponse dichotomique :

$$y_i = \begin{cases} 0 & \text{si } y_{cont,i} > m \\ 1 & \text{sinon,} \end{cases}$$

où m est la médiane des éléments constituant le vecteur $\mathbf{y}_{cont}$.

Par la suite, la première composante principale du module brun a été simulée :

$$\mathbf{x}_{brun} = \alpha_{brun}\mathbf{y} + \boldsymbol{\varepsilon}\sqrt{1 - \alpha_{brun}^2},$$

où $\alpha_{brun} = -0.25$ et $\boldsymbol{\varepsilon} \sim N_n(\mathbf{0}, \mathbf{1})$.

Les composantes principales des modules bleu, turquoise et jaune ont été simulées indépendamment de la variable réponse, mais les modules bleu et turquoise sont corrélés à 0.6 :

$$\mathbf{x}_{bleu} = 0.6 \cdot \mathbf{x}_{turquoise} + \boldsymbol{\varepsilon}\sqrt{1 - 0.6^2}.$$

Une fois les cinq composantes principales simulées, nous avons utilisé la fonction *simulateDatExpr5Module* en R pour générer les profils d'expression de chacun des gènes. Cette fonction permet également de générer les profils d'expression des gènes faisant partie du module gris, module composé de tous les gènes n'ayant aucun lien avec les autres.

La construction des prédicteurs est incluse dans la fonction *simulateDatExpr5Module* en R ; pour plus de détails, on doit se référer au tutoriel de Langfelder et d'Horvath (2008a).

4.3 Résultats et analyse des tests d'association

Nous présenterons les résultats des tests d'association du premier scénario dans la section (4.3.1) et ceux du deuxième scénario dans la section (4.3.2) et nous analyserons les résultats des deux scénarios ensemble dans la section (4.3.3).

4.3.1 Résultats du scénario 1

Avant de regarder et d'interpréter les résultats des tests d'association, nous avons tracé le réseau des 20 premiers gènes et nous avons vérifié que les gènes 3 et 19 sont bel et bien des gènes à haute connectivité. Dans la figure (4.1), nous n'avons présenté que les liens associés à des corrélations supérieures à 0.35 afin de mieux visualiser le réseau.

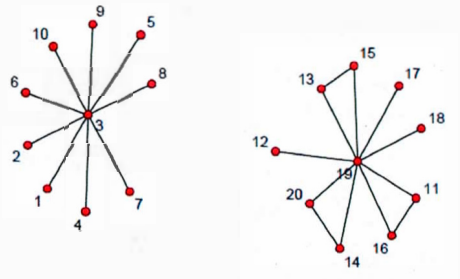


Figure 4.1: Réseau des 20 premiers gènes

Nous pouvons voir dans la figure (4.2) que la plupart des approches contrôlent bien l'erreur de type I. Toutefois l'approche composée de la méthode des K-moyennes et de la corrélation de Pearson est relativement conservatrice.

Afin de tester la performance de chacune des approches, nous avons calculé leur puissance pour différentes valeurs de niveau de significativité α. Vous trouverez les résultats dans le tableau (4.1). Nous remarquons que les puissances des approches comprenant la méthode WGCNA sont toutes égales à 1.000. La WGCNA est très puissante pour des données simulées selon un modèle de régression standard. De telles puissances ne sont cependant pas observées pour toutes les approches. En effet, les approches composées de la méthode des K-moyennes ont des puissances très faibles. Selon le niveau de α, les puissances varient entre 0.059 et 0.274. Nous remarquons également que les puissances sont encore plus faibles lorsque la méthode des K-moyennes est jumelée avec la corrélation de Pearson. Enfin, les puissances des approches comprenant les méthodes de régularisation se situent entre celles des approches avec la WGCNA et celles des approches avec les K-moyennes. Les puissances sont toutefois plus élevées avec la méthode LASSO qu'avec la méthode Elastic Net.

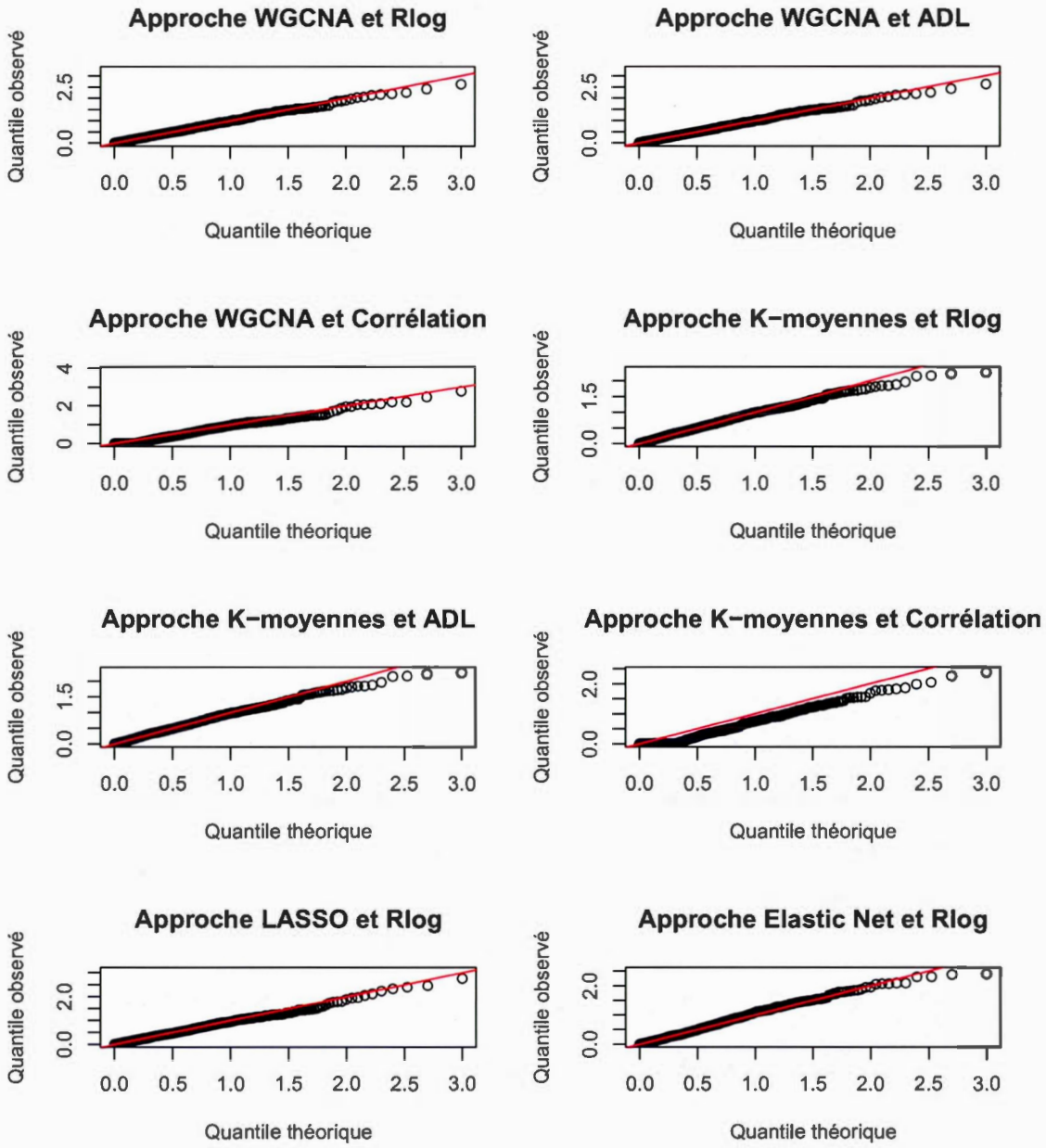


Figure 4.2: QQplot des huit approches pour le premier scénario.

Tableau 4.1: Scénario 1 : Puissance des approches pour différentes valeurs de α. La puissance est la proportion des 1000 réplications qui ont un seuil observé plus grand que le seuil théorique α.

Approches	<0.1	<0.05	<0.025	<0.01	<0.005
WGCNA-Rlog	1.000	1.000	1.000	1.000	1.000
WGCNA-ADL	1.000	1.000	1.000	1.000	1.000
WGCNA-Cor	1.000	1.000	1.000	1.000	1.000
K-moyennes-Rlog	0.194	0.132	0.094	0.066	0.059
K-moyennes-ADL	0.194	0.132	0.094	0.066	0.059
K-moyennes-Cor	0.106	0.076	0.057	0.046	0.042
LASSO-Rlog	0.926	0.881	0.824	0.731	0.672
EN-Rlog	0.641	0.559	0.502	0.429	0.386

4.3.2 Résultats du scénario 2

Sur la figure (4.3), nous pouvons voir un léger biais pour les approches WGCNA-Rlog et WGCNA-ADL. En effet, à partir du quantile théorique 1.5, les quantiles observés sont légèrement au-dessus de ce qu'ils devraient être sous l'hypothèse nulle. Par contre, les approches composées de la corrélation de Pearson contrôlent bien l'erreur de type I pour ce scénario.

Pour ce qui est des puissances, nous observons une moins grande différence entre celles des approches comprenant la méthode WGCNA et celles composées des K-moyennes. En effet, les puissances des approches avec WGCNA sont plus élevées d'environ 0.04. Nous tenterons de donner une explication à cette observation dans la section (4.3.3). De plus, nous remarquons que les puissances sont les mêmes lorsque nous utilisons la régression logistique et l'analyse discriminante. Enfin, dans ce deuxième scénario, ce sont les approches avec les méthodes de régula-

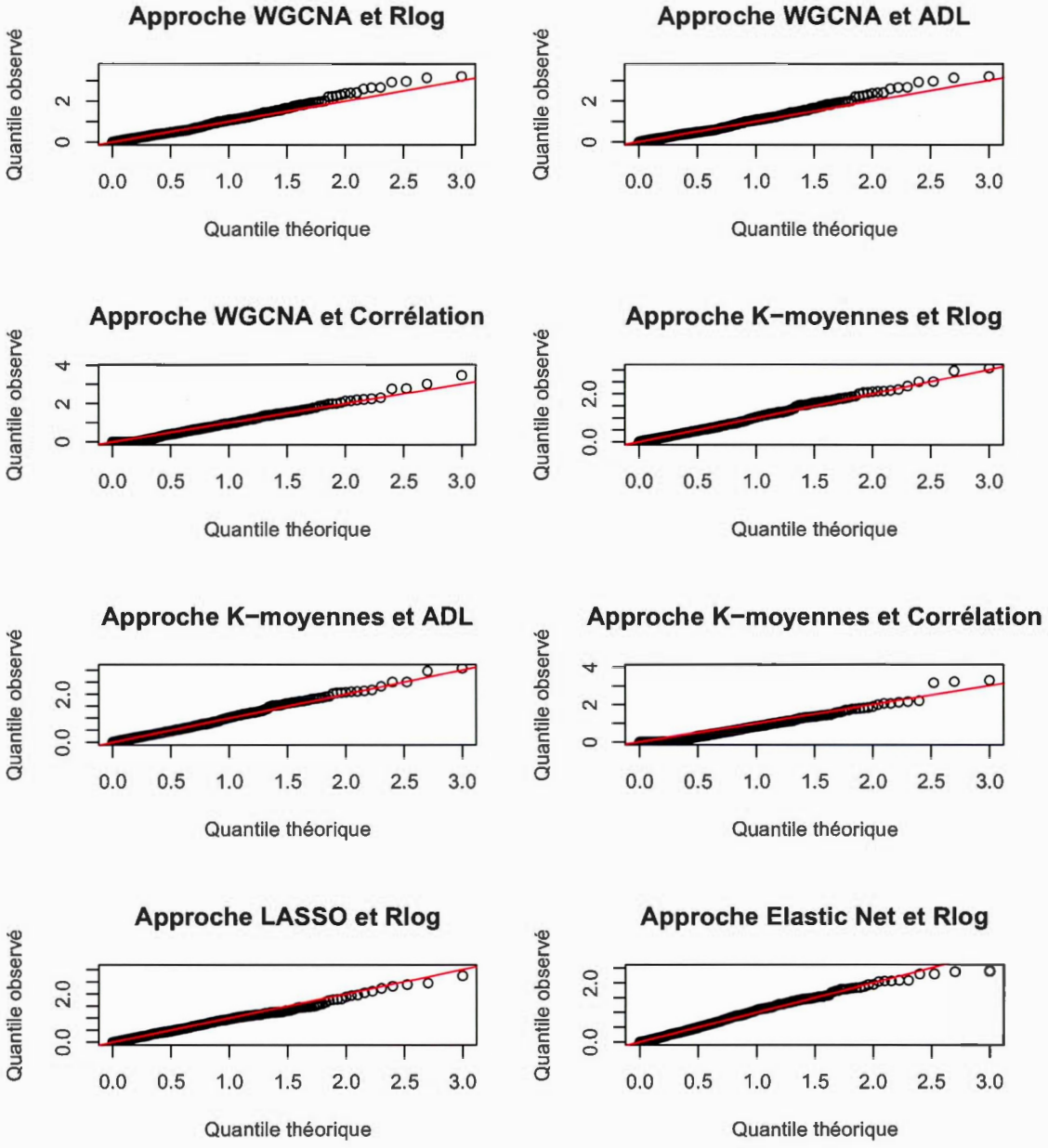


Figure 4.3: QQplot des huit approches pour le deuxième scénario.

risation qui sont les moins puissantes. Toutefois, la méthode LASSO reste légèrement plus performante que l'Elastic Net. Les résultats sont présentés dans le tableau (4.2).

Tableau 4.2: Scénario 2 : Puissance des approches pour différentes valeurs de α. La puissance est la proportion des 1000 réplications qui ont un seuil observé plus grand que le seuil théorique α.

Approches	<0.1	<0.05	<0.025	<0.01	<0.005
WGCNA-Rlog	0.683	0.564	0.454	0.333	0.269
WGCNA-ADL	0.683	0.564	0.454	0.333	0.269
WGCNA-Cor	0.686	0.574	0.474	0.344	0.272
K-moyennes-Rlog	0.640	0.522	0.431	0.304	0.235
K-moyennes-ADL	0.640	0.522	0.431	0.304	0.235
K-moyennes-Cor	0.619	0.529	0.449	0.327	0.269
LASSO-Rlog	0.121	0.078	0.044	0.020	0.013
EN-Rlog	0.115	0.054	0.032	0.013	0.009

4.3.3 Analyse des résultats

Tel que mentionné dans les résultats ci-haut, la majorité des approches contrôlent le biais de l'erreur de type I sous l'hypothèse nulle, mais certaines font exception telles les approches avec la corrélation de Pearson. Ceci s'explique par le fait que nous corrigeons les valeurs-p avec la méthode de Bonferroni et qu'en plus nous prenons la valeur-p minimale parmi toutes les valeurs-p corrigées. Ce test est donc conservateur.

Pour ce qui est de la performance des tests, nous pouvons discerner quelques points

en commun pour les deux scénarios. Tout d'abord, les approches comprenant la méthode WGCNA sont les plus puissantes. Peu importe le scénario et la valeur de α, les puissances reliées à la méthode WGCNA sont supérieures aux autres. Ceci nous amène à croire que la méthode WGCNA semble plus appropriée que les méthodes de régularisation et la méthode des K-moyennes.

De manière générale, avec la méthode LASSO, nous pouvons sélectionner jusqu'à n gènes (où n est le nombre de sujets et $n < p$) et avec la méthode Elastic Net, nous pouvons en retenir jusqu'à p gènes. Par contre, pour pouvoir utiliser la régression logistique et l'analyse discriminante, nous avons dû réduire le nombre de gènes obtenus par les méthodes de régularisation à $n_c - 1$ lorsque ces dernières en retenaient davantage, où n_c est le nombre de sujets utilisés pour les méthodes classiques (Rlog et ADL). De plus, comme nous avons eu recours à une méthode de régularisation et à une méthode standard, nous avons dû séparer notre jeu de données en deux groupes pour éviter un biais dans l'estimation de la distribution de la statistique du test sous l'hypothèse nulle. Nous avons pris 100 sujets pour la méthode de régularisation (i.e. pour faire la sélection des variables) et 50 sujets pour la méthode standard (i.e. pour tester l'association). Le nombre de sujets pour le test d'association est petit ce qui explique que les approches avec les méthodes de régularisation sont moins performantes. Par conséquent, les approches composées des méthodes de régularisation et de la régression logistique telles que décrites ne sont pas adéquates pour tester l'association.

Un autre point en commun observé lors des simulations est que les puissances obtenues avec la régression logistique et l'analyse discriminante sont similaires. Ce résultat n'est pas surprenant, car les données proviennent d'une loi normale multivariée. Dans une telle situation, la régression logistique et l'analyse discriminante sont deux approches équivalentes et donc donnent les mêmes résultats (Hastie *et al.*, 2009, p.127) .

Dans le premier scénario, les puissances des approches avec la méthode des K-moyennes sont beaucoup plus faibles que celles observées avec la méthode WGCNA. La méthode des K-moyennes ne capte pas bien la structure des groupes qui sont associés avec la variable réponse. Dans le deuxième scénario, les puissances de ces deux types d'approches sont très semblables. Cette différence entre les deux scénarios peut s'expliquer par le fait que le scénario 1 est très parcimonieux tandis que le scénario 2 l'est beaucoup moins. En effet, dans le scénario 1, nous avions 20 gènes associés à la variable réponse et dans le scénario 2, nous avions 120 gènes associés à la variable réponse. De plus, les approches avec la méthode WGCNA sont plus intéressantes que celles avec les K-moyennes, car elles n'ont pas besoin de connaître le nombre de groupes préalablement.

En résumé, la méthode WGCNA est plus performante que les autres méthodes. Elle capte la structure des groupes des prédicteurs qui sont connectés entre eux. Toutes les informations sur les gènes sont concentrées sous forme de module. La méthode (régression logistique, analyse discriminante ou corrélation de Pearson) pour déterminer les modules significatifs ne semble cependant pas influencer les résultats lorsque nous formons les groupes avec la méthode WGCNA.

4.4 Résultats et analyse de la sélection des variables et des tests de prédiction

Jusqu'à présent, nous avons comparé diverses approches en termes de puissance pour vérifier celle qui donne la meilleure associativité. Nous allons maintenant évaluer la performance de trois méthodes de régularisation (LASSO, Elastic Net et analyse discriminante pénalisée basée sur la fonction de Fisher) selon deux modèles de simulation. Nous vérifierons si la manière de simuler les données a un impact sur la performance des méthodes. Nous comparerons aussi le pouvoir prédictif des trois méthodes énumérées ci-haut avec celui de l'approche modifiée

basée sur les gènes à haute connectivité.

Nous présenterons les résultats du premier et du deuxième scénario dans les sections (4.4.1) et (4.4.2) respectivement. Dans la section (4.4.3), nous analyserons les résultats des deux scénarios ensemble.

4.4.1 Résultats du scénario 1

Avant de regarder les résultats, rappelons-nous que pour le premier scénario nous avons simulé $p = 1000$ gènes dont 20 étaient reliés à la maladie. Par conséquent, le nombre de faux positifs est au maximum 980 et le nombre de faux négatifs est au maximum 20 gènes.

Puisque l'approche modifiée de Chou *et al.* (2014) sert à trouver les gènes à haute connectivité et non les 20 gènes reliés à la maladie, nous n'avons pas calculé le taux de bonne parcimonie, le nombre de faux positifs et le nombre de faux négatifs pour cette méthode.

Vous trouverez le taux moyen de bonne parcimonie, le nombre moyen de faux positifs et le nombre moyen de faux négatifs pour les trois méthodes de régularisation dans le tableau (4.3) et les diagrammes à moustaches à la figure (4.4). Nous remarquons que le taux moyen de bonne parcimonie et le nombre moyen de faux négatifs associés à l'analyse discriminante pénalisée basée sur la fonction de Fisher sont inférieurs à ceux des deux autres méthodes. Par contre, c'est la méthode LASSO qui a le plus faible nombre moyen de faux positifs.

Tableau 4.3: Résultats du scénario 1 pour la sélection des variables à la suite de 100 réplications. Les colonnes 2 à 4 désignent le taux moyen de bonne parcimonie, le nombre moyen de faux positifs et le nombre moyen de faux négatifs. Les chiffres entre parenthèses représentent les écarts-types.

Méthodes de régularisation	Taux moyen de bonne parcimonie	Nombre moyen de faux positifs	Nombre moyen de faux négatifs
Elastic Net (EN)	0.92 (0.14)	68.59 (142.36)	12.00 (4.34)
LASSO	0.97 (0.01)	11.41 (7.86)	16.46 (1.41)
Fisher	0.80 (0.26)	195.00 (261.10)	9.04 (7.86)

Pour ce qui est de l'aire moyenne sous la courbe, celle de l'approche modifiée « WGCNA + EN » est significativement supérieure à celle des trois autres méthodes (voir tableau (4.4)). Ainsi, lorsque nous ne prenons que les gènes à haute connectivité, nous avons une meilleure prédiction. Les diagrammes à moustaches des aires sous la courbe pour les quatre approches sont présentés à la figure (4.5). Nous remarquons que l'approche WGCNA+EN est moins volatile que les trois autres.

Tableau 4.4: Résultats du scénario 1 pour la prédiction à la suite de 100 réplications. La colonne 2 désigne l'aire moyenne sous la courbe ROC. Les chiffres entre parenthèses représentent les écarts-types.

Approche	Aire moyenne sous la courbe
Elastic Net	0.80 (0.08)
LASSO	0.81 (0.09)
Fisher	0.61 (0.12)
WGCNA + EN	0.91 (0.04)

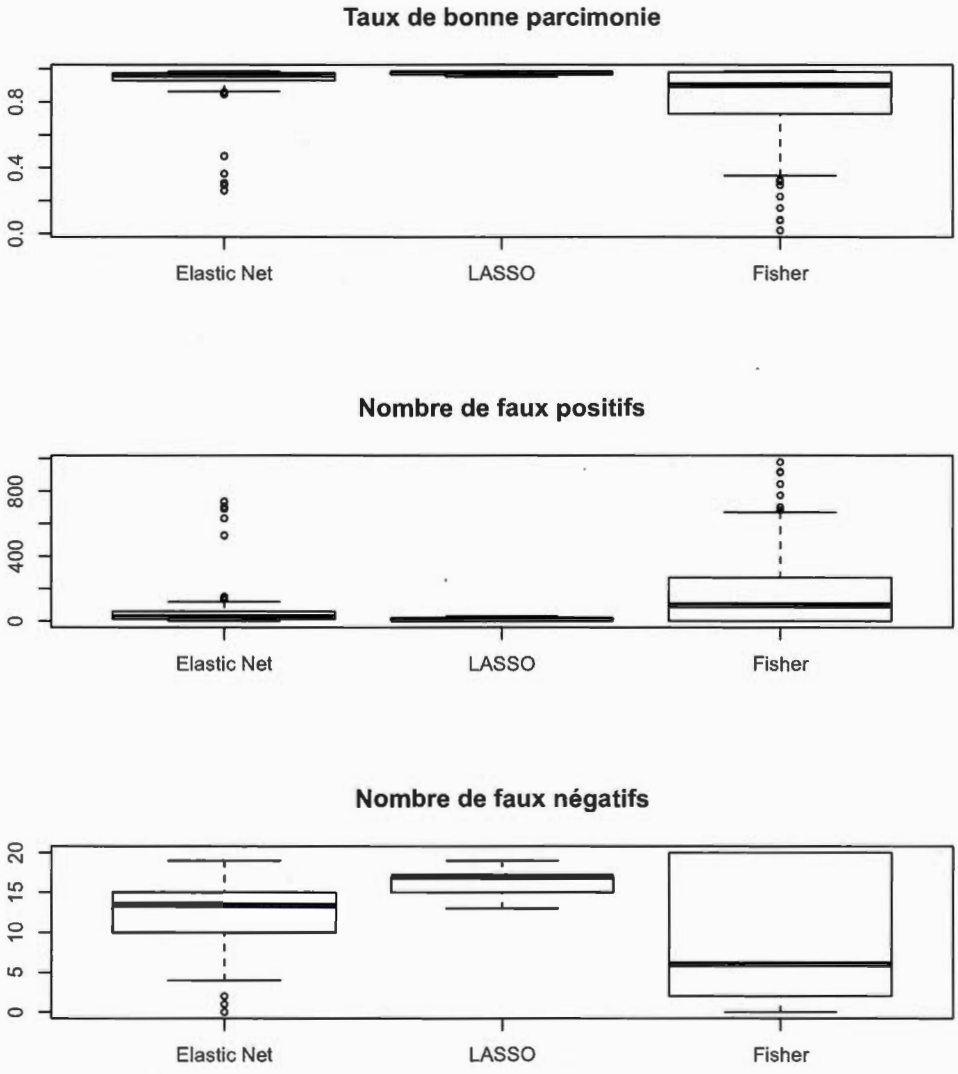


Figure 4.4: Diagrammes à moustaches pour le scénario 1.

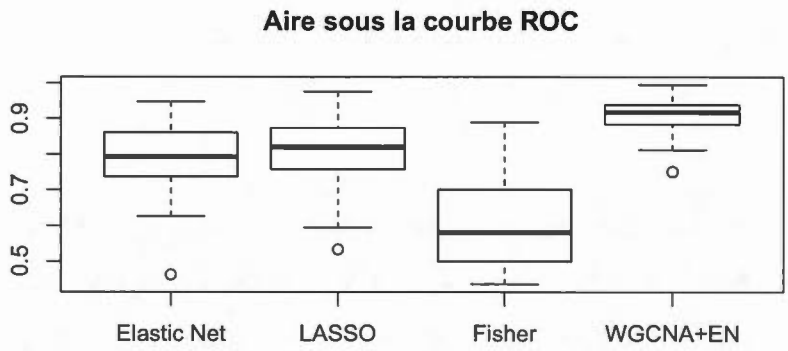


Figure 4.5: Diagrammes à moustaches des aires sous la courbe pour le scénario 1. 100 réplications ont été faites.

4.4.2 Résultats du scénario 2

Pour le scénario 2, nous avons simulé 1000 gènes dont 120 sont reliés à la maladie. Le nombre de faux positifs ne peut donc pas être supérieur à 880 et le nombre de faux négatifs ne dépassera jamais 120.

Encore une fois, le taux de bonne parcimonie, le nombre de faux positifs et le nombre de faux négatifs ne sont calculés que pour les méthodes de régularisation (pas pour celle WGCNA + EN). De plus, ne connaissant pas les coefficients β_j pour chacun des gènes, $j = 1, \ldots, p$, nous supposons que les coefficients associés aux gènes des deux groupes corrélés à la variable réponse sont différents de 0 et que les coefficients de tous les autres gènes sont égaux à 0. L'information concernant les coefficients n'est pas disponible lorsque nous simulons les données selon le modèle proposé dans le tutoriel d'Horvath et de Langfelder.

Nous remarquons dans le tableau (4.5) ou sur la figure (4.6) que la méthode LASSO a le meilleur taux moyen de bonne parcimonie, soit 0.88, et qu'elle a le

plus faible nombre moyen de faux positifs, soit 6.09. L'Elastic Net et l'analyse discriminante basée sur la fonction de Fisher ont cependant un nombre moyen de faux négatifs inférieur à celui de la méthode LASSO.

Tableau 4.5: Résultats du scénario 2 pour la sélection des variables à la suite de 100 réplications. Les colonnes 2 à 4 représentent le taux moyen de bonne parcimonie, le nombre moyen de faux positifs et le nombre moyen de faux négatifs pour les trois méthodes de régularisation (LASSO, Elastic Net et analyse discriminante pénalisée basée sur la fonction de Fisher). Les chiffres entre parenthèses sont les écarts-types.

Méthodes de régularisation	Taux moyen de bonne parcimonie	Nombre moyen de faux positifs	Nombre moyen de faux négatifs
Elastic Net	0.77 (0.16)	132.24 (188.69)	97.32 (29.15)
LASSO	0.88 (0.01)	6.09 (7.72)	118.37 (1.94)
Fisher	0.62 (0.28)	301.73 (321.35)	74.24 (45.90)

Finalement, l'aire sous la courbe de l'approche modifiée WGCNA + EN est significativement supérieure aux trois méthodes de régularisation (voir tableau (4.6)). Ainsi, l'approche WGCNA+EN donne la meilleure prédiction. Vous trouverez également les diagrammes à moustache des aires sous la courbe des quatre approches à la figure (4.7).

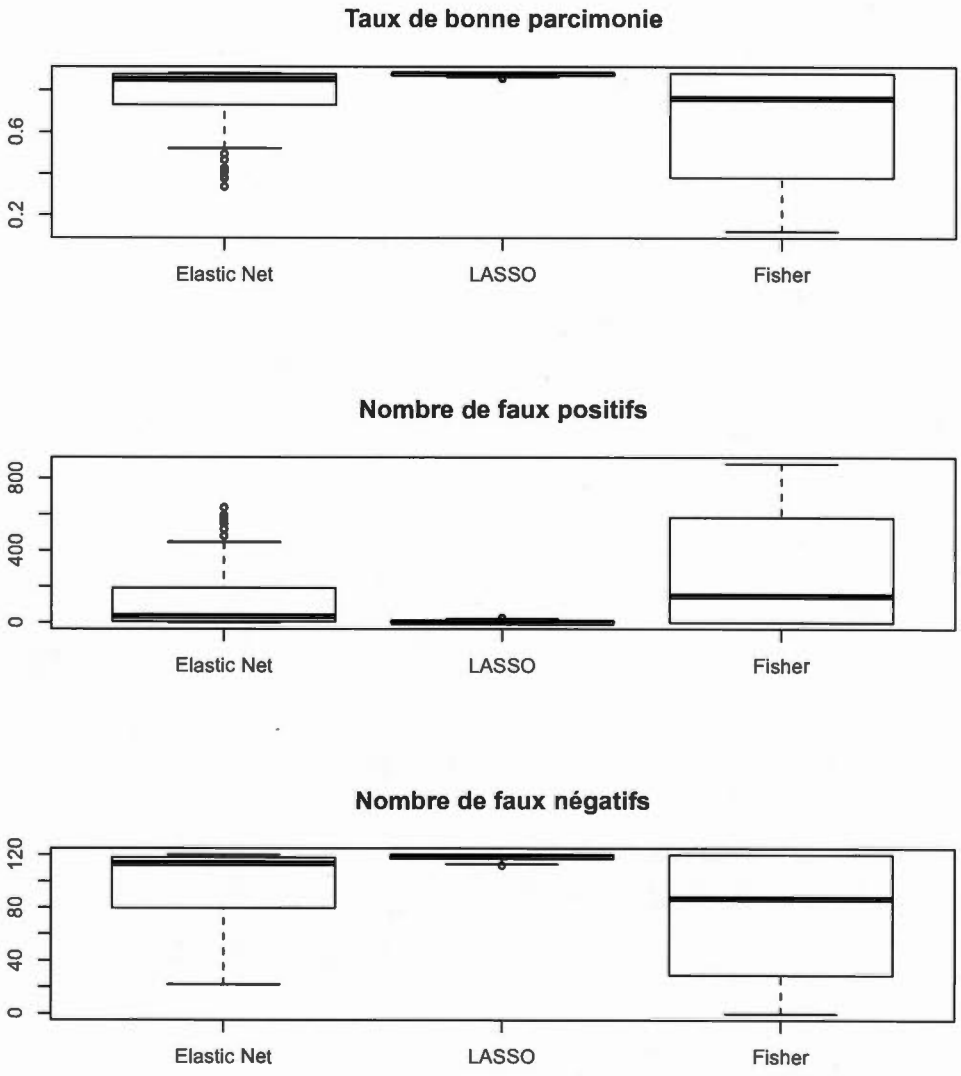


Figure 4.6: Diagrammes à moustaches pour le scénario 2.

Tableau 4.6: Résultats du scénario 2 pour la prédiction à la suite de 100 réplications. La colonne 2 désigne l'aire moyenne sous la courbe ROC. Les chiffres entre parenthèses représentent les écarts-types.

Approche	Aire sous la courbe
Elastic Net	0.54 (0.10)
LASSO	0.53 (0.07)
Fisher	0.52 (0.05)
WGCNA + EN	0.63 (0.08)

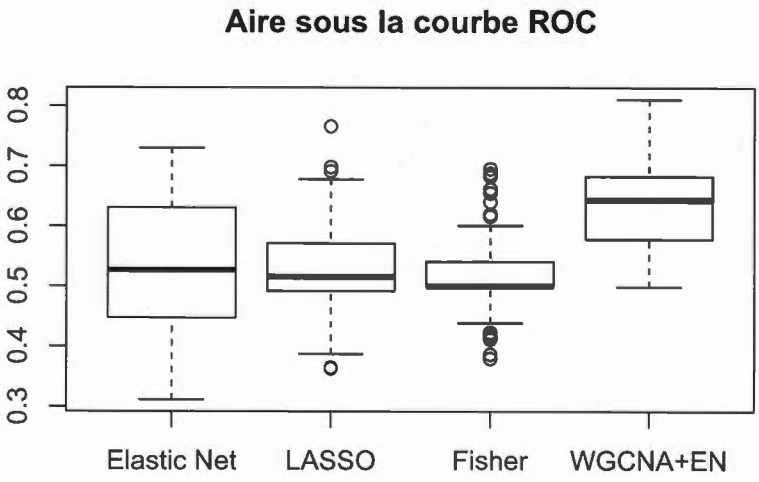


Figure 4.7: Diagrammes à moustaches des aires sous la courbe pour le scénario 2. 100 réplications ont été faites.

4.4.3 Analyse des résultats

Un des objectifs était de vérifier la performance des méthodes de régularisation avec des données simulées selon un modèle de régression standard et avec des données simulées selon le modèle WGCNA. Peu importe la manière dont les données sont simulées, la méthode LASSO est celle qui a le plus faible nombre moyen de faux positifs et l'analyse discriminante pénalisée basée sur la fonction de Fisher est celle qui admet le plus petit nombre moyen de faux négatifs.

Nous avons remarqué que la méthode Elastic Net donne des résultats avec de plus grands écarts-types que la méthode LASSO. Ceci peut être dû à la validation croisée qui a été faite sur le paramètre α, paramètre prenant place dans la pénalité de l'Elastic Net. Nous n'avons pris que quatre valeurs différentes de α à cause du temps d'exécution.

Un autre but de cette analyse était de comparer le pouvoir prédictif des modèles de régularisation avec celui de la méthode WGCNA+EN qui permet de trouver les gènes à haute connectivité. L'approche composée de la WGCNA et de l'Elastic Net est celle avec la plus grande aire sous la courbe, peu importe le scénario de simulation. Cette dernière permet non seulement de faire une bonne prédiction, mais elle permet aussi de trouver les gènes à haute connectivité. Elle a donc un grand avantage sur les autres. Elle est capable de réduire l'information en modules et de capter les gènes qui ont le plus d'influence sur la maladie et qui sont connectés à plusieurs autres gènes.

CHAPITRE V

ANALYSE D'UN JEU DE DONNÉES RÉELLES

Grâce à une collaboration avec une équipe du CHU Sainte-Justine, nous avons eu accès à des données sur la leucémie. Selon la Société canadienne du cancer, la leucémie « est un cancer qui prend naissance dans les cellules souches du sang. Les cellules souches sont des cellules de base qui se transforment en différents types de cellules qui ont des fonctions distinctes. » (Société canadienne du cancer, 2016) Les cellules souches du sang peuvent devenir des cellules souches lymphoïdes, puis se transformer en lymphocytes (types de globule blanc). Il existe trois types de lymphocytes : lymphocytes B, lymphocytes T et cellules tueuses naturelles (NK). Ces divers lymphocytes ont pour but de combattre les infections en fabriquant des anticorps. Il arrive malheureusement que ces globules blancs ne se développent pas complètement ou qu'ils ne se développement pas normalement. Ces cellules défectueuses, appelées blastes, se propagent dans le sang et envahissent les cellules normales. Ainsi, l'organisme n'est plus capable de produire les anticorps et il devient vulnérable face aux infections.

Le jeu de données que nous possédons contient de l'information sur 64 enfants atteints de la leucémie. Pour chacun des sujets, nous connaissons le sous-type de la lignée cellulaire atteinte, c'est-à-dire le type de cellules où la leucémie a pris naissance. Les deux sous-types de lignées cellulaires possibles dans nos études

sont : pré-B ou pré-T. Le préfixe « pré » signifie que le cancer s'est développé dans une cellule qui n'a pas atteint sa maturité maximale et la lettre suivant le mot « pré » fait référence aux lymphocytes B ou T. Nous avons également des renseignements pour chacun des patients sur leurs anomalies chromosomiques et génétiques : translocation des chromosomes 12 et 21 (t(12,21)), translocation des chromosomes 9 et 22 (t(9,22)), hyperdiploïdie ou autre. Une hyperdiploïdie est le fait d'avoir plus de 50 chromosomes et une translocation est une mutation génétique. Nous savons aussi si les enfants atteints de la leucémie ont fait une rechute, s'ils ont une mutation du gène RAS et s'ils ont le chromosome de Philadelphie (Ph-like). Finalement, nous avons leurs expressions génétiques pour 40763 gènes.

Analyse du gène RAS

Le but de cette analyse est de trouver les gènes qui ont une connectivité élevée et qui prédisent la présence ou non de la mutation du gène RAS. Pour ce faire, nous considérons le gène RAS comme une variable réponse dichotomique

$$y = \begin{cases} 1 & \text{s'il y a une mutation du gène RAS chez l'individu} \\ 0 & \text{s'il n'y a pas de mutation du gène RAS chez l'individu} \end{cases}$$

et les autres gènes comme des variables explicatives continues dont leurs expressions génétiques forment la matrice $\mathbf{X}$.

Pour une raison de temps d'exécution, nous diminuons le nombre de variables explicatives à 4384 gènes. La sélection des gènes est basée sur leur variation génétique. Nous ne gardons que ceux qui sont les plus volatiles et qui ont un écart-type supérieur à 1. De plus, seuls les patients de la lignée pré-B qui n'ont pas de translocation des chromosomes 12 et 21 sont considérés dans l'étude. Nous excluons les sujets avec t(12,21), car, selon l'équipe du CHU Sainte-Justine, l'association entre

cette translocation et la variable RAS est déjà connue. Ainsi, la matrice $\mathbf{X}$ est de dimension 34×4384.

Lors de l'analyse WGCNA, nous avons utilisé la matrice de similarité de base. Il est adéquat d'utiliser une matrice symétrique dans ce cas-ci puisque nous ne connaissons pas les liens de causalité entre les gènes. Nous avons ensuite déterminé la valeur de β qui nous permet d'avoir un modèle invariant d'échelle (la méthode pour choisir la valeur optimale de β est décrite dans la section (3.2.3)). Bien que $\beta = 3$ nous permet d'avoir un R_s^2 supérieur à 0.8, nous choisissons plutôt $\beta = 4$ pour deux raisons. La première raison est que plus le R_s^2 est élevé, plus notre modèle est invariant d'échelle. Avec $\beta = 3$, le R_s^2 est égal à 0.848 et avec $\beta = 4$, il est égal à 0.917 (voir table (5.1)). De plus, la connectivité moyenne est de 111 gènes avec $\beta = 4$ ce qui demeure élevé. La deuxième raison qui nous incite à choisir $\beta = 4$ est qu'il y a une saturation à partir de ce point. En effet, les valeurs de R_s^2 pour les β suivants sont autour de 0.917 (voir figure (5.1)).

Tableau 5.1: Choix de β.

Bêta	R_s^2	pente	Connectivité moyenne
1	0.210	1.469	1028.199
2	0.370	-0.696	390.613
3	0.848	-1.082	192.229
4	0.917	-1.150	111.373
5	0.914	-1.129	71.869
6	0.911	-1.118	49.894
7	0.926	-1.091	36.449
8	0.935	-1.072	27.623
9	0.947	-1.063	21.515
10	0.957	-1.060	17.113

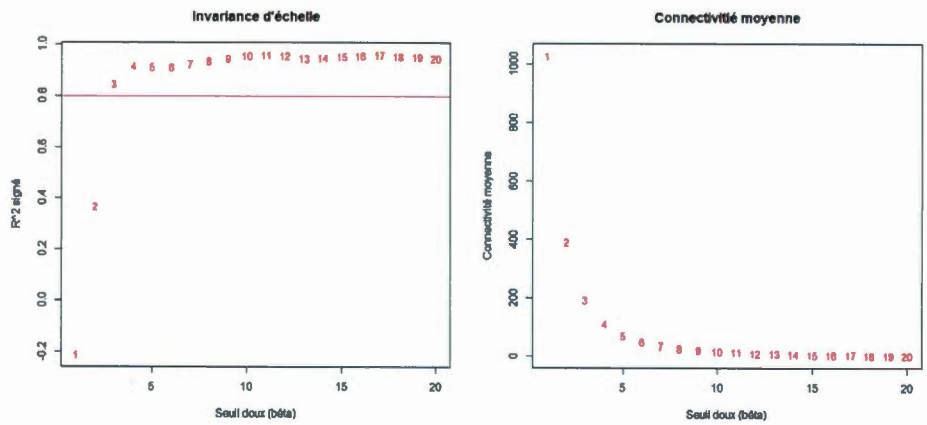


Figure 5.1: Choix de bêta.

Suite au choix de β, nous avons créé les modules tel que décrit dans la section 3.3. Nous avons obtenu 11 modules après la fusion de ceux corrélés (>0.75). Nous pouvons voir sur la figure (5.2) que les modules bleu et turquoise se sont fusionnés pour ne former qu'un seul module de couleur bleue. De même, les gènes des modules jaune et brun et ceux des modules mauve et saumon forment les modules brun et mauve respectivement.

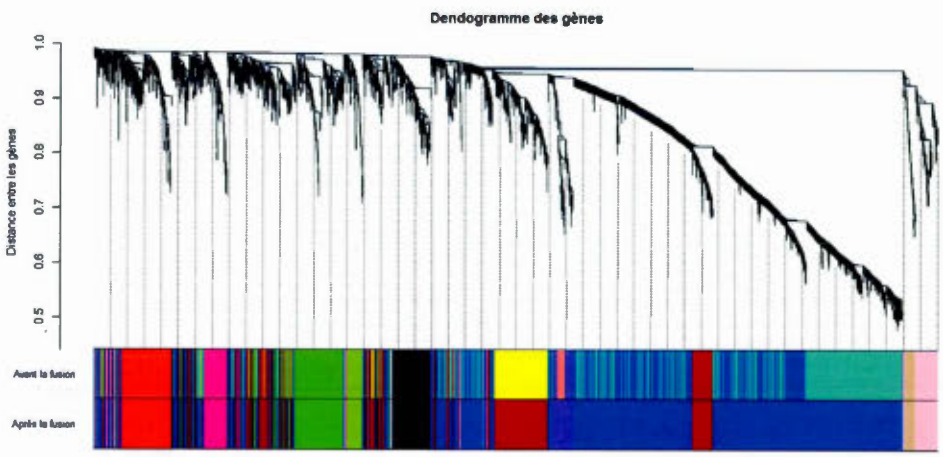


Figure 5.2: Dendogramme des gènes.

La distribution du nombre de gènes par module est présentée dans le tableau (5.2). Nous remarquons que le module bleu est le plus grand avec 2264 gènes et le module gris est le plus petit avec seulement 3 gènes. Le module gris est le module contenant tous les gènes n'ayant pas de connectivité avec les autres gènes.

Tableau 5.2: Nombre de gènes par module.

Module	Nombre de gènes
noir	225
bleu	2264
brun	735
vert	294
vert-jaune	86
gris	3
magenta	124
rose	163
mauve	150
rouge	271
beige foncé	69

Afin de mieux visualiser les distances entre les modules, nous avons tracé un graphique de positionnement multidimensionnel (« classical multidimensional scaling »). De manière générale, nous pouvons voir sur la figure (5.3) que les gènes d'un même module sont assez proches les uns des autres et que les modules sont éloignés les uns par rapport aux autres.

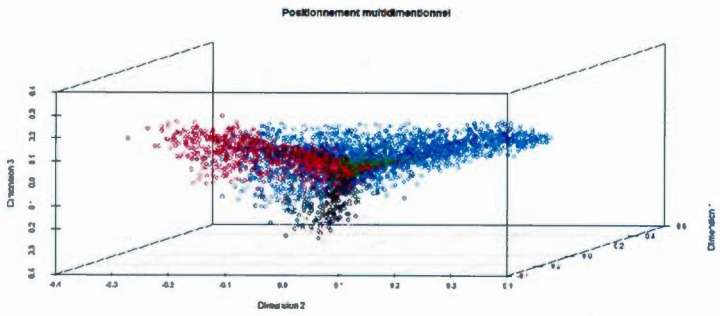


Figure 5.3: Graphique 3D de positionnement multidimensionnel

Pour vérifier l'association entre la variable y (présence ou non du gène RAS chez l'enfant) et chacune des premières composantes principales, nous avons d'abord regardé les valeurs-p issues du test de Pearson et corrigées avec Bonferroni. Ces dernières se situent entre 0.396 et 1.000 ce qui indique qu'aucun module n'est significatif (voir la figure (5.4)).

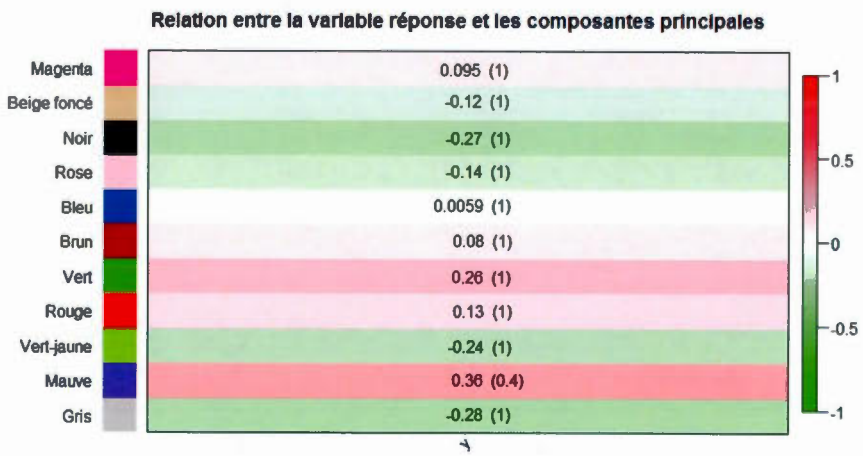


Figure 5.4: Relation entre la variable réponse et les composantes principales. Le premier chiffre de chaque ligne indique la corrélation de Pearson entre le module et la variable réponse et le deuxième (entre parenthèses) indique la valeur-p corrigée avec la méthode de Bonferroni.

Puisque la méthode basée sur la corrélation de Pearson corrigée avec la méthode de Bonferroni est conservatrice en présence d'un nombre élevé de modules et puisque le nombre de sujets est faible ($n = 34$), nous nous ne sommes pas limités à cette approche pour vérifier la significativité des groupes. En effet, nous avons appliqué quatre autres méthodes de sélection et nous avons vérifié leur chevauchement pour déterminer les groupes qui sont significatifs. Nous avons regardé le chevauchement de quatre méthodes plutôt d'une seule, car les tests sont peu puissants en raison du faible nombre d'individus (voir appendice C pour une simulation qui réflète le scénario de données réelles et qui montre l'impact du faible nombre de sujets sur la puissance). Les quatre méthodes que nous avons appliquées sont : une approche basée sur le taux de faux positifs, une recherche pas-à-pas progressive sur une régression logistique, une analyse discriminante progressive et une analyse discriminante pénalisée basée sur la fonction de Fisher.

Nous avons donc commencé par corriger les valeurs-p issues du test de Pearson pour les comparaisons multiples à l'aide de l'approche basée sur le taux de faux positifs (FDR). Cette approche est moins conservatrice que celle de Bonferroni, mais elle n'est pas très puissante dans ce cas-ci à cause du faible nombre de sujets présents dans l'étude. Nous n'avons donc gardé que les groupes de gènes avec les plus petites valeurs-q même si ces dernières ne sont pas inférieures à 0.05. Ainsi, nous avons retenu les groupes noir, vert, vert-jaune, mauve et gris (voir tableau (5.3)).

Nous avons ensuite fait une régression logistique entremêlée d'une recherche pas-à-pas progressive (« stepwise selection »). L'idée générale de cette approche est de sélectionner le meilleur modèle possible en ajoutant ou en éliminant un prédicteur à chaque étape du processus. Afin de déterminer la variable qui sera ajoutée ou éliminée, une statistique F associée à chacun des régresseurs est calculée. La variable ayant la plus grande valeur F et étant supérieure à un seuil prédéterminé est

Tableau 5.3: Valeurs-q obtenues grâce à l'approche FDR.

Module	valeur-q
Magenta	0.7160653
Beige foncé	0.7022144
Noir	0.3594011
Rose	0.7022144
Bleu	0.9735993
Brun	0.7160653
Vert	0.3594011
Rouge	0.7022144
Vert-jaune	0.3958405
Mauve	0.3594011
Gris	0.3594011

ajoutée au modèle tandis qu'une variable avec une statistique F non significative est éliminée. De plus, à chaque pas, le critère d'information d'Akaike (AIC) est calculé pour chacune des variables sélectionnées. Si l'AIC du modèle sans prédicteur est inférieur à l'AIC des modèles comprenant les variables sélectionnées, alors l'algorithme cesse et nous obtenons le meilleur modèle possible. Par exemple, dans le tableau (5.4), à la dernière étape du processus de recherche pas-à-pas, les trois régresseurs significatifs sont les modules gris, vert-jaune et mauve. De plus, le modèle sans prédicteur a un AIC égal à 41.452 et tous les autres modèles ont un AIC supérieur à cette valeur. Par conséquent, nous retenons le modèle composé des modules gris, vert-jaune et mauve. Nous avons également fait le test d'Hosmer-Lemeshow afin de tester l'adéquation du modèle composé de ces trois modules. Nous avons obtenu une statistique égale à 6.5837 avec valeur-p égale à 0.5821 ce qui signifie que le modèle s'ajuste bien aux données. Nous avons également vérifié le pouvoir prédictif du modèle réduit à trois modules en calculant l'aire sous la courbe ROC. Nous avons obtenu une aire égale à 0.8596 ce qui signifie que nous

retrouvons la bonne classe environ 9 fois sur 10.

Tableau 5.4: Résultats de la régression logistique entremêlée d'une recherche pas-à-pas progressive. Les colonnes 1 à 4 désignent le module, le degré de liberté, la déviance et l'AIC.

Module	Dl	Deviance	AIC
<aucun>		33.452	41.452
Gris	1	35.977	41.977
Vert-jaune	1	39.168	45.168
Mauve	1	41.254	47.254

La troisième approche que nous avons essayée pour trouver les modules significatifs est l'analyse discriminante ascendante. Cette approche ajoute une variable au modèle à chaque pas et elle calcule le critère de lambda de Wilks (Roever *et al.*, 2014). Elle retient le modèle qui minimise ce critère et dont la valeur-p associée à chacun des prédicteurs est significative. Nous pouvons voir dans le tableau (5.5) que les modules mauve, vert-jaune, gris et beige foncé ont été retenus, car ils ont tous une valeur-p inférieure à 0.05.

Tableau 5.5: Résultats de l'analyse discriminante progressive. Les colonnes 1 à 4 représentent respectivement le module, lambda de Wilks, la statistique F approximée pour le modèle sélectionné et la valeur-p de la statistique F.

Module	lambda de Wilks	Statistique F	Valeur-p
Mauve	0.870	4.794	0.036
Vert-jaune	0.741	5.414	0.010
Gris	0.664	5.067	0.006
Beige foncé	0.621	4.420	0.007

Nous avons terminé la sélection des groupes avec une analyse discriminante pénalisée basée sur la fonction de Fisher. Seuls les modules avec des coefficients $\hat{\beta}_k$ différents de 0 discriminent bien les individus avec ou sans mutation du gène RAS. Autrement dit, seuls les modules avec des coefficients différents de 0 sont sélectionnés. Ainsi, les modules noir, vert, vert-jaune, mauve et gris sont retenus par cette méthode (voir table (5.6)).

Tableau 5.6: Résultats de l'analyse discriminante pénalisée basée sur la fonction de Fisher.

Module	Coefficient
Magenta	-0.00
Beige foncé	0.00
Noir	0.34
Rose	0.00
Bleu	-0.00
Brun	-0.00
Vert	-0.33
Rouge	-0.00
Vert-jaune	0.20
Mauve	-0.78
Gris	0.37

Dans ces quatre situations, les modules jaune-vert, mauve et gris sont apparus significatifs. Étant donné que le module gris est composé de tous les gènes n'ayant pas de connectivité avec les autres gènes (Langfelder et Horvath, 2008b), nous n'avons gardé que les groupes jaune-vert et mauve comme modules significatifs. Pour chacun des gènes présents dans les modules significatifs, nous avons calculé leur importance (GS), leur connectivité standardisée (K) et leur fréquence de prédiction (f). Nous avons recueilli 62 gènes sur les 4384 avec $GS > 0.20$, $K > 0.20$

et $f > 750$. Ainsi, il y a 62 gènes à haute connectivité. Vous trouverez la liste de ces gènes dans l'appendice D. Plusieurs d'entre eux sont reliés au processus de la prolifération cellulaire, de la réplication de l'ADN ou de liaison des protéines pour n'en nommer que quelques-uns.

Afin de connaître les voies biologiques affectées par les 62 gènes trouvés, nous avons utilisé l'encyclopédie KEGG (Kyoto Encyclopedia of Genes and Genomes). Six voies biologiques ont été identifiées significatives à la suite de tests hypergéométriques (Zhang *et al.*, 2005).

Tableau 5.7: Voies biologiques significatives. Pour chacune des voies biologiques significatives, nous avons son numéro d'identification (ID), son nom, le nombre de gènes parmi les 62 trouvés qui font partie de cette voie, le nombre de gènes qui constituent la voie et la valeur-p obtenue à la suite d'un test hypergéométrique.

ID	Noms	Gènes trouvés	Total des gènes	Valeurs-p
04110	Cell cycle	9	128	9.646563e-14
04914	Progesterone-mediated oocyte maturation	4	87	2.241105e-06
04114	Oocyte meiosis	4	114	8.354937e-06
04115	p53 signaling pathway	3	69	2.696860e-05
03440	Homologous recombination	2	28	5.093630e-05
00670	One carbon pool by folate	1	18	1.024887e-03

Dans le tableau (5.7), une des voies biologiques a particulièrement attiré notre attention : la « p53 signaling pathway ». Cette dernière est une voie de transduction qui active plusieurs protéines, dont la p53, à la suite d'un stimulus tels un dommage à l'ADN ou l'activation d'oncogènes (Kyoto University Bioinformatics Center, 2016). Lors de la transduction du signal, plusieurs interactions de protéines

prennent place à l'intérieur de la cellule et une des réponses à ces interactions est l'activation de la transcription de la protéine p53. La protéine p53 a deux principaux rôles : l'arrêt du cycle cellulaire et l'apoptose. En effet, lorsque l'ADN subit un dommage, la p53 bloque l'évolution des cellules vers la division et active des gènes. Ces gènes réparent le matériel génétique et lorsque l'ADN est réparé, la cellule peut continuer son cheminement. Par contre, dans certains cas, les gènes sont incapables de réparer l'ADN. C'est pourquoi la p53 oblige la cellule à se « suicider » afin d'éviter la prolifération de cellules cancéreuses. C'est ce qu'on appelle la mort programmée ou l'apoptose (Fridman et Lowe, 2003). Malheureusement, il arrive que la protéine p53 soit mutée ou qu'elle soit tout simplement absente. Ceci entraîne une prolifération des cellules anormales potentiellement cancéreuses. Des mutations de la p53 ont été trouvées dans 50 % des cancers humains (Soussi, 2012).

Dans notre étude sur le jeu de données de la leucémie, trois gènes de la voie biologique « p53 signaling pathway » ont ressorti : GTSE1, CCNE2 et CDK1. Ces gènes sont tous reliés aux cellules lymphocytes B. De plus, le gène GTSE1 est déjà reconnu comme étant relié à la leucémie, plus spécifiquement à « poorly differentiated B cell origin leukemia » et à « acute lymphoblastic leukemia » (European Molecular Biology Laboratory, 2016). Ainsi, nos résultats vont dans le même sens que d'autres études faites auparavant.

CONCLUSION

Les données de grande dimension sont de plus en plus populaires et de plus en plus utilisées dans le monde de la recherche. En effet, les chercheurs recueillent des milliers de données pour mieux comprendre et pour mieux expliquer divers phénomènes complexes. Par exemple, en statistique génétique, les scientifiques collectent de l'information sur des milliers de gènes pour découvrir ceux qui sont impliqués dans un cancer. Le fait d'avoir des renseignements sur des milliers de gènes nous permet de mieux saisir les interactions entre les gènes et de mieux cibler les dysfonctionnements survenant dans un réseau de voies biologiques. Bien que le but de la recherche soit très noble, peu de sujets prennent part aux expériences pour diverses raisons. Nous nous retrouvons donc avec une matrice $\mathbf{X}^T\hat{\mathbf{W}}\mathbf{X}$ non inversible. Les méthodes classiques, comme la régression logistique et l'analyse discriminante linéaire, ne peuvent plus être appliquées directement sur ce type de données. Les méthodes de régularisation quant à elles peuvent être appliquées sur des données massives pour faire de la prédiction, mais elles ne permettent pas de tester l'association entre la variable réponse $\mathbf{y}$ et la matrice $\mathbf{X}$. Enfin, la méthode « Weighted Gene Co-expression Network Analysis » peut également être appliquée à des jeux de données de grande dimension, par contre elle ne permet ni de prédire ni de tester l'association entre une maladie et les gènes.

Dans ce mémoire, nous avons d'abord étudié des méthodes classiques, quelques méthodes de régularisation et la méthode WGCNA. Nous avons ensuite développé et proposé huit approches pouvant être appliquées à des données massives et pouvant tester l'association entre la variable réponse et les prédicteurs. Ces approches sont composées d'une méthode de régularisation ou de la méthode WGCNA et

d'une méthode classique. Nous avons découvert que nos approches constituées d'une méthode de régularisation ne sont pas optimales. En effet, pour effectuer ces approches, nous avons dû séparer nos données en groupes ce qui a réduit considérablement le nombre d'individus pour chacune des deux étapes (étape 1 : sélection des variables et étape 2 : test d'association). De plus, pour pouvoir effectuer les méthodes classiques à l'étape 2, il faut moins de gènes que d'individus. Ainsi, si les méthodes Elastic Net ou LASSO retenaient davantage de gènes à l'étape 1 que d'individus disponibles à l'étape 2, nous devions réduire la quantité de gènes en ne gardant que ceux avec les plus grands coefficients. Par conséquent, les puissances associées aux méthodes de régularisation sont plutôt faibles. La méthode qui ressort du lot est la WGCNA. Cette dernière, basée sur la co-expression de gènes, permet de trouver les modules significatifs avec de grandes puissances.

À la suite des tests d'association, nous avons comparé le pouvoir prédictif des méthodes de régularisation avec celui de l'approche WGCNA+EN. Le fait de calculer l'interconnectivité des gènes pour former les groupes et ensuite appliquer l'Elastic Net pour trouver les gènes à haute connectivité augmente grandement l'aire sous la courbe ROC. Les prédictions sont meilleures lorsque nous utilisons l'approche WGCNA+EN qu'une méthode de régularisation à elle seule.

Que ce soit pour l'association ou pour la prédiction, la méthode WGCNA est très efficace. Elle permet de réduire les informations sur les gènes sous forme de modules et de ne garder que les gènes avec une grande connectivité.

Dans le jeu de données réelles portant sur des enfants atteints de la leucémie, nous avons découvert, en utilisant l'approche WGCNA+EN, que 62 gènes sont reliés au gène RAS avec une connectivité élevée. Parmi ces 62 gènes, 23 d'entres eux font partie de voies biologiques significativement enrichies dont 3 prennent part à la voie biologique « p53 signaling pathway ». Cette dernière, présente dans plusieurs

types de cancer, semble avoir également un effet sur la leucémie.

Dans une prochaine recherche, il serait intéressant d'appliquer les approches proposées dans ce mémoire, en particulier celle composée de la WGCNA, dans des analyses GWAS (« genome-wide association scans ») afin de sélectionner des groupes de marqueurs génétiques qui agissent en groupes sur des maladies complexes et de construire ce qu'on appelle un « polygenic risk score » (PRS) pour bien prédire de telles maladies. Un PRS est défini comme une combinaison linéaire de plusieurs marqueurs, chacun pondéré par un poids selon sa significativité dans l'étude GWAS (Dudbridge, 2013).

APPENDICE A

TRANSCRIPTION ET TRADUCTION

Les gènes et les protéines sont étroitement liés dans les processus de transcription et de traduction. En effet, ce sont les gènes qui contiennent l'information nécessaire à la synthèse des protéines bien que ce ne sont pas eux qui les construisent directement. C'est l'acide ribonucléique (ARN) qui joue le rôle d'intermédiaire entre l'ADN et la synthèse des protéines. L'ADN et l'ARN sont tous les deux formés de bases azotées, par contre, l'ADN est constitué d'adénine, de guanine, de cytosine et de thymine tandis que l'ARN est composé d'adénine, de guanine, de cytosine et d'uracile. Le phénomène qui permet la création de l'ARN à partir de l'ADN est appelé transcription et celui qui permet la synthèse des polypeptides est appelé traduction.

A.1 Transcription

Pour que la transcription ait lieu, plusieurs protéines, appelées facteurs de transcription, doivent se lier au promoteur de l'ADN. Le promoteur est en fait une séquence d'ADN qui indique l'endroit où la synthèse de l'ARN commence. Lorsqu'il y a suffisamment de facteurs de transcription, une enzyme, appelée ARN polymérase II, se fixe au promoteur. Une fois cette dernière fixée, l'ADN se déroule

et ses deux brins s'éloignent un de l'autre au fur et à mesure que l'ARN polymérase II avance sur l'ADN. Ceci permet aux nucléotides de s'attacher au brin codant de l'ADN pour former l'ARN prémessager. Dans les cellules eucaryotes, l'ARN prémessager se détache de l'ADN avant même que l'ARN polymérase II ait terminé la transcription. Il y a donc des nucléotides qui viennent s'attacher aux deux extrémités de l'ARN prémessager pour ainsi former l'ARN messager (ARNm) pendant la période de maturation. Par la suite, l'ARN messager sort du noyau de la cellule et se déplace au cytoplasme pour entamer la synthèse d'un polypeptide (Campbell et Reece, 2007).

A.2 Traduction

La synthèse d'un polypeptide à partir de l'ARN messager est appelée la traduction. Cette dernière est composée de trois principales phases : l'initiation, l'élongation et la terminaison. Lors de l'étape de l'initiation, la petite sous-unité ribosomique s'attache à l'ARNm et elle avance jusqu'au codon AUG qui est le codon de départ de la traduction. Puis, un ARN de transfert (ARNt) portant de la méthionine à une extrémité et ayant comme anticodon d'initiation UAC s'attache au codon de départ. De manière générale, l'ARNt porte à une de ses extrémités un acide aminé tandis que l'autre est composée d'un anticodon. Par la suite, la grande sous-unité ribosomique se lie aux autres éléments du complexe d'initiation grâce à des protéines appelées facteurs d'initiation. Le ribosome ainsi formé est composé de trois sites (E, P et A). À cette étape, le site P est occupé par l'ARNt d'initiation et les deux autres sites sont libres. Lors de la phase de l'élongation, l'anticodon d'un ARNt s'attache au codon de l'ARNm au site A, puis à cet acide aminé s'attache l'acide aminé situé au site P. Ces acides aminés forment un polypeptide. L'ARNt du site P passe au site E et se détache du ribosome tandis que l'ARNt du site A

passe au site P pendant qu'un autre ARNt vient s'attacher au site A. Ce cycle se poursuit jusqu'à ce que le codon d'arrêt de l'ARNm atteint le site A. Ainsi, lors de la phase de terminaison, une molécule d'eau s'attache à la chaîne polypeptidique grâce à des facteurs de terminaison et le polypeptide se détache du ribosome. Enfin, le polypeptide se replie sur lui-même pour former une molécule tridimensionnelle et peut subir quelques modifications post-traductionnelles avant de devenir une protéine fonctionnelle (Campbell et Reece, 2007).

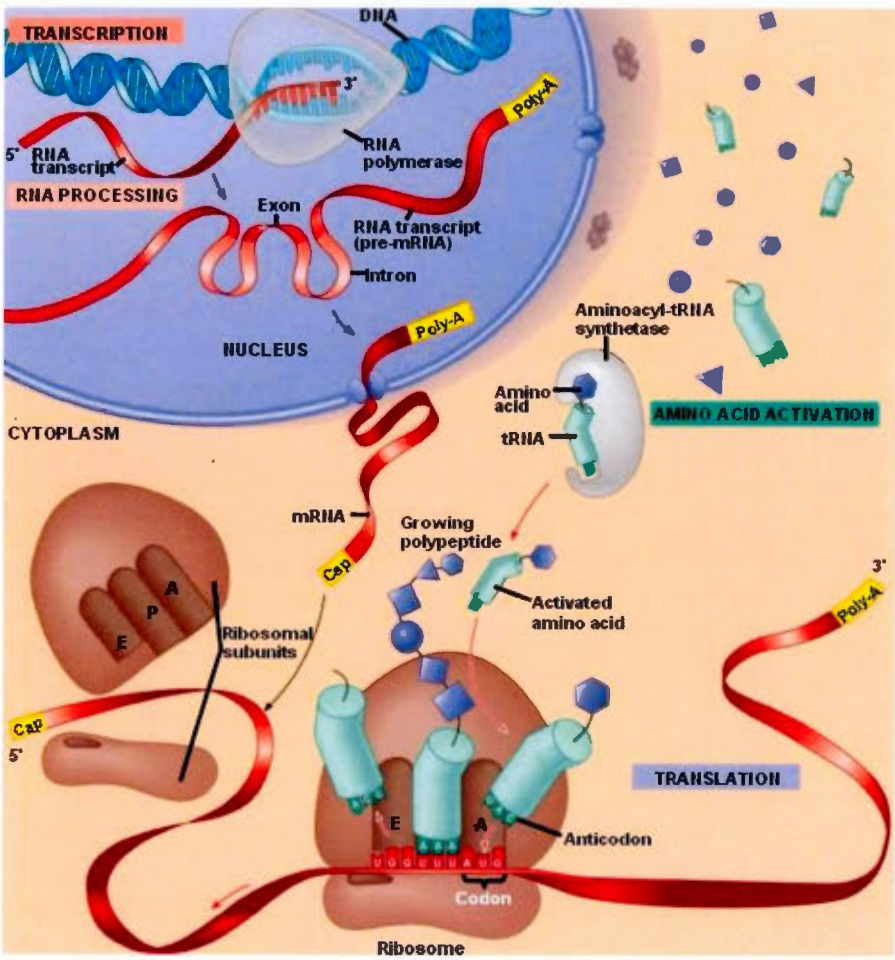


Figure A.1: Transcription et traduction. (Campbell et Reece, 2005)

APPENDICE B

ESTIMATEURS OBTENUS PAR LA RÉGRESSION RIDGE

B.1 Estimateurs

Pour estimer les paramètres d'une régression Ridge, il faut d'abord trouver la log-vraimsemblance avec la pénalité comme ceci :

$$l^{\lambda}(\boldsymbol{\beta}) = l(\boldsymbol{\beta}) - \lambda||\boldsymbol{\beta}||^2, \tag{B.1}$$

où

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{n} [y_i \log \pi_i + (1 - y_i) \log(1 - \pi_i)]$$

et

$$||\boldsymbol{\beta}||^2 = \sum_{i=1}^{p} \beta_j^2.$$

Le paramètre qui maximise l'équation (B.1) est noté $\hat{\boldsymbol{\beta}}^{\lambda}$ et il peut être obtenu à l'aide de l'algorithme de Newton-Raphson.

La première dérivée de $l^{\lambda}(\boldsymbol{\beta})$, notée $U^{\lambda}(\boldsymbol{\beta})$, est définie par

$$\begin{aligned} U^{\lambda}(\boldsymbol{\beta}) &= \mathbf{X}^T(\mathbf{y} - \boldsymbol{\pi}) - 2\lambda\boldsymbol{\beta} \\ &= U(\boldsymbol{\beta}) - 2\lambda\boldsymbol{\beta}, \end{aligned} \tag{B.2}$$

où $U(\boldsymbol{\beta})$ est la dérivée de la log-vraisemblance non restreinte, $\mathbf{y}$ est le vecteur contenant les y_i et $\boldsymbol{\pi}$ est le vecteur contenant les probabilités π_i.

L'opposé de la dérivée seconde, noté $\Omega^{\lambda}(\boldsymbol{\beta})$, est défini par

$$\begin{aligned}\Omega^{\lambda}(\boldsymbol{\beta}) &= \mathbf{X}^T\mathbf{W}\mathbf{X} + 2\lambda\mathbf{I} \\ &= \Omega(\boldsymbol{\beta}) + 2\lambda\mathbf{I}, \end{aligned} \tag{B.3}$$

où $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_p)$ est une matrice de dimension $n \times p$, $\mathbf{W}$ est une matrice diagonale de dimension $n \times n$ avec $w_{ii} = \pi_i(1 - \pi_i)$ et $\mathbf{I}$ est une matrice identité.

Nous pouvons maintenant utiliser la série de Taylor pour la première dérivée de la log-vraisemblance pénalisée autour du vrai paramètre $\boldsymbol{\beta_0}$ (Le Cessie et Van Houwelingen, 1992) :

$$U^{\lambda}(\hat{\boldsymbol{\beta}}^{\boldsymbol{\lambda}}) = U^{\lambda}(\boldsymbol{\beta}_0) - (\hat{\boldsymbol{\beta}}^{\lambda} - \boldsymbol{\beta}_0)\Omega^{\lambda}(\boldsymbol{\beta}_0) + O(||\hat{\boldsymbol{\beta}}^{\lambda} - \boldsymbol{\beta}_0||). \tag{B.4}$$

En posant $U^{\lambda}(\hat{\boldsymbol{\beta}}^{\boldsymbol{\lambda}}) = 0$, nous obtenons

$$\begin{aligned} 0 &\approx U^{\lambda}(\boldsymbol{\beta}_0) - (\hat{\boldsymbol{\beta}}^{\lambda} - \boldsymbol{\beta}_0)\Omega^{\lambda}(\boldsymbol{\beta}_0) \\ U^{\lambda}(\boldsymbol{\beta}_0) &\approx (\hat{\boldsymbol{\beta}}^{\lambda} - \boldsymbol{\beta}_0)\Omega^{\lambda}(\boldsymbol{\beta}_0) \\ \hat{\boldsymbol{\beta}}^{\lambda} &\approx \boldsymbol{\beta}_0 + \Omega^{\lambda}(\boldsymbol{\beta}_0)^{-1}U^{\lambda}(\boldsymbol{\beta}_0). \end{aligned}$$

Ainsi, en utilisant les équations (B.2) et (B.3), nous pouvons approximer $\hat{\boldsymbol{\beta}}^{\lambda}$ par :

$$\begin{aligned} \hat{\boldsymbol{\beta}}^{\lambda} &\approx \boldsymbol{\beta}_0 + (\Omega(\boldsymbol{\beta}_0) + 2\lambda\mathbf{I})^{-1}(U(\boldsymbol{\beta}_0) - 2\lambda\boldsymbol{\beta}_0) \\ &\approx (\Omega(\boldsymbol{\beta}_0) + 2\lambda\mathbf{I})^{-1}(\Omega(\boldsymbol{\beta}_0) + 2\lambda\mathbf{I})\boldsymbol{\beta}_0 + (\Omega(\boldsymbol{\beta}_0) + 2\lambda\mathbf{I})^{-1}(U(\boldsymbol{\beta}_0) - 2\lambda\boldsymbol{\beta}_0) \\ &\approx (\Omega(\boldsymbol{\beta}_0) + 2\lambda\mathbf{I})^{-1}(\Omega(\boldsymbol{\beta}_0)\boldsymbol{\beta}_0 + 2\lambda\boldsymbol{\beta}_0 + U(\boldsymbol{\beta}_0) - 2\lambda\boldsymbol{\beta}_0) \\ &\approx (\Omega(\boldsymbol{\beta}_0) + 2\lambda\mathbf{I})^{-1}(U(\boldsymbol{\beta}_0) + \Omega(\boldsymbol{\beta}_0)\boldsymbol{\beta}_0). \end{aligned} \tag{B.5}$$

L'estimateur de premier ordre pour le maximum de vraisemblance non restreint est

$$\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}_0 + \Omega^{-1}(\boldsymbol{\beta}_0)(U(\boldsymbol{\beta}_0),$$

que nous pouvons réécrire sous la forme

$$U(\beta_0) = \Omega(\beta_0)(\hat{\beta} - \beta_0). \tag{B.6}$$

En utilisant les équations (B.5) et (B.6), nous obtenons

$$\begin{aligned}\hat{\beta}^{\lambda} &= (\Omega(\beta_0) + 2\lambda\mathbf{I})^{-1}(\Omega(\beta_0)(\hat{\beta} - \beta_0) + \Omega(\beta_0)\beta_0) \\ &= (\Omega(\beta_0) + 2\lambda\mathbf{I})^{-1}\Omega(\beta_0)\hat{\beta}.\end{aligned}$$

B.2 Biais et variance

Puisque sous certaines conditions $\hat{\beta}$ est asymptotiquement sans biais de variance $\Omega(\beta_0)^{-1}$ (Le Cessie et Van Houwelingen, 1992), le biais asymptotique de l'estimateur Ridge est

$$\begin{aligned}\text{Biais}\left[\hat{\beta}^{\lambda}\right] &= \mathbb{E}\left[\hat{\beta}^{\lambda} - \beta_0\right] \\ &= \mathbb{E}\left[\{\Omega(\beta_0) + 2\lambda\mathbf{I}\}^{-1}\Omega(\beta_0)\hat{\beta} - \beta_0\right] \\ &= \{\Omega(\beta_0) + 2\lambda\mathbf{I}\}^{-1}\Omega(\beta_0)\beta_0 - \beta_0 \\ &= \{\Omega(\beta_0) + 2\lambda\mathbf{I}\}^{-1}\Omega(\beta_0)\beta_0 - \{\Omega(\beta_0) + 2\lambda\mathbf{I}\}^{-1}\{\Omega(\beta_0) + 2\lambda\mathbf{I}\}\beta_0 \\ &= \{\Omega(\beta_0) + 2\lambda\mathbf{I}\}^{-1}\left[\Omega(\beta_0)\beta_0 - \Omega(\beta_0)\beta_0 - 2\lambda\beta_0\right] \\ &= -2\lambda\{\Omega(\beta_0) + 2\lambda\mathbf{I}\}^{-1}\beta_0\end{aligned}$$

et la variance asymptotique de $\hat{\beta}^{\lambda}$ est

$$\begin{aligned}\text{Var}\left[\hat{\beta}^{\lambda}\right] &= \text{Var}\left[\{\Omega(\beta_0) + 2\lambda\mathbf{I}\}^{-1}\Omega(\beta_0)\hat{\beta}\right] \\ &= \left[\{\Omega(\beta_0) + 2\lambda\mathbf{I}\}^{-1}\Omega(\beta_0)\right]\text{Var}\left[\hat{\beta}\right]\left[\{\Omega(\beta_0) + 2\lambda\mathbf{I}\}^{-1}\Omega(\beta_0)\right]^T \\ &= \{\Omega(\beta_0) + 2\lambda\mathbf{I}\}^{-1}\Omega(\beta_0)\Omega(\beta_0)^{-1}\Omega(\beta_0)^T\left[\{\Omega(\beta_0) + 2\lambda\mathbf{I}\}^{-1}\right]^T \\ &= \{\Omega(\beta_0) + 2\lambda\mathbf{I}\}^{-1}\Omega(\beta_0)\{\Omega(\beta_0) + 2\lambda\mathbf{I}\}^{-1}.\end{aligned}$$

APPENDICE C

SIMULATION AVEC UN FAIBLE NOMBRE DE SUJETS

Nous avons simulé des données telles que présentées dans le tutoriel d'Horvath et de Langfelder. Par contre, nous avons posé $n = 34$ sujets et $p = 4384$ prédicteurs pour avoir les mêmes dimensions que dans notre jeu de données réelles. Nous avons ensuite calculé les puissances pour les 8 approches décrites dans le chapitre 4. Nous remarquons dans la table (C.1) que les puissances sont beaucoup plus petites que celles que nous avions obtenues avec $n = 150$ et $p = 1000$.

Tableau C.1: Puissance des 8 approches lorsqu'il y a 34 sujets et 4384 gènes.

Méthode	<0.1	<0.05	<0.025	<0.01	<0.005
WGCNA + Rlog	0.374	0.250	0.166	0.094	0.052
WGCNA + ADL	0.374	0.250	0.166	0.094	0.052
WGCNA + corrélation	0.360	0.242	0.174	0.094	0.066
K-moyennes + Rlog	0.281	0.171	0.098	0.047	0.030
K-moyennes + ADL	0.281	0.171	0.098	0.047	0.030
K-moyennes + corrélation	0.251	0.168	0.111	0.058	0.035
LASSO + Rlog	0.108	0.059	0.027	0.010	0.007
Elastic Net + Rlog	0.104	0.049	0.019	0.007	0.003

APPENDICE D

LISTE DES GÈNES À HAUTE CONNECTIVITÉ

La liste des gènes à connectivité élevée est présentée dans la table ci-dessous. Vous y trouverez non seulement son annotation « Ensembl », mais aussi le module auquel il appartient, son importance génétique (GS), sa connectivité standardisée (K) et sa fréquence de prédiction (f).

Tableau D.1: Liste des gènes à connectivité élevée

Module	Ensembl	GS	K	f
Jaune-vert	ENSG00000175175	-0.3646800	0.4449538	998
Jaune-vert	ENSG00000171992	-0.3397147	0.5044471	924
Jaune-vert	ENSG00000237636	-0.3162609	0.2791868	780
Jaune-vert	ENSG00000185737	-0.4608809	0.2065167	1000
Mauve	ENSG00000156970	0.3540073	0.8338320	773
Mauve	ENSG00000115163	0.2711292	0.6103249	755
Mauve	ENSG00000154839	0.3011605	0.6587690	773
Mauve	ENSG00000151725	0.2998627	0.4358405	773
Mauve	ENSG00000198826	0.2650693	0.4853515	755
Mauve	ENSG00000175567	-0.3976958	0.2079781	999
Mauve	ENSG00000122966	0.2417919	0.3164491	773

Module	Ensembl	GS	K	f
Mauve	ENSG00000119326	0.4507020	0.5622652	999
Mauve	ENSG00000143228	0.3711762	0.7733656	773
Mauve	ENSG00000267383	0.2664460	0.2327725	773
Mauve	ENSG00000267041	0.3311844	0.4730614	773
Mauve	ENSG00000129173	0.2705664	0.5779882	773
Mauve	ENSG00000176890	0.2559278	0.4953512	767
Mauve	ENSG00000162437	0.4100003	0.2396603	999
Mauve	ENSG00000075218	0.2926328	0.4979688	773
Mauve	ENSG00000118193	0.3539786	0.8693818	773
Mauve	ENSG00000153044	0.3164334	0.4950081	773
Mauve	ENSG00000104738	0.2587923	0.5168858	773
Mauve	ENSG00000154920	0.3314301	0.3255550	773
Mauve	ENSG00000138658	0.4068288	0.4364291	915
Mauve	ENSG00000093009	0.3210356	0.5754355	773
Mauve	ENSG00000144395	0.4234393	0.4746736	945
Mauve	ENSG00000072571	0.2870101	0.6978128	773
Mauve	ENSG00000121211	0.3228286	0.4362292	773
Mauve	ENSG00000123975	0.2446980	0.3160999	773
Mauve	ENSG00000184661	0.3330331	0.9057205	773
Mauve	ENSG00000100526	0.2971980	0.6237556	773
Mauve	ENSG00000183850	0.3246480	0.2959714	773
Mauve	ENSG00000175305	0.3541588	0.3653456	916
Mauve	ENSG00000136982	0.3156104	0.7746977	773
Mauve	ENSG00000177602	0.2910346	0.5492373	773
Mauve	ENSG00000138778	0.4741865	0.5843134	1000
Mauve	ENSG00000169679	0.3353536	0.5411078	773

Module	Ensembl	GS	K	f
Mauve	ENSG00000142731	0.3805819	0.6861894	773
Mauve	ENSG00000104147	0.3195799	0.6835086	773
Mauve	ENSG00000136492	0.4284996	0.8242763	928
Mauve	ENSG00000011426	0.3817368	0.7877223	773
Mauve	ENSG00000137804	0.2907766	0.8625247	771
Mauve	ENSG00000111247	0.4201574	0.8568484	927
Mauve	ENSG00000065328	0.2933592	0.7799694	771
Mauve	ENSG00000189057	0.2836941	0.6243901	773
Mauve	ENSG00000109805	0.3672326	0.8362003	773
Mauve	ENSG00000139618	0.3741474	0.5748266	773
Mauve	ENSG00000051341	0.2995503	0.6108476	773
Mauve	ENSG00000174371	0.3456158	0.8653935	773
Mauve	ENSG00000161888	0.2597933	0.5678048	773
Mauve	ENSG00000112742	0.4185939	0.8703328	913
Mauve	ENSG00000164109	0.3658348	0.7049722	773
Mauve	ENSG00000123485	0.3055176	0.8626594	773
Mauve	ENSG00000145386	0.3036067	0.9109576	767
Mauve	ENSG00000126787	0.3843969	0.9330087	773
Mauve	ENSG00000131747	0.3650638	1.0000000	773
Mauve	ENSG00000066279	0.3372248	0.9012754	773
Mauve	ENSG00000024526	0.3961359	0.6687241	915
Mauve	ENSG00000170312	0.3643564	0.8761297	773
Mauve	ENSG00000237649	0.2716337	0.4633074	773
Mauve	ENSG00000112984	0.3235227	0.5916386	773
Mauve	ENSG00000124942	-0.2998476	0.2499520	773

RÉFÉRENCES

Albert, R. (2005). Scale-free networks in cell biology. *Journal of cell science*, *118*(21), 4947–4957.

Albert, R. et Barabasi, A.-L. (2002). Statistical mechanics of complex networks. *Reviews of modern physics*, *74*, 47–97.

Barabási, A.-L. et Albert, R. (1999). Emergence of scaling in random networks. *science*, *286*(5439), 509–512.

Campbell, N. et Reece, J. (2005). *Biology*. Biology. Pearson, Benjamin Cummings.

Campbell, N. A. et Reece, J. B. (2007). *Biologie* (3e éd.). Édition du Renouveau Pédagogique Inc.

Chou, W.-C., Cheng, A.-L., Brotto, M. et Chuang, C.-Y. (2014). Visual gene-network analysis reveals the cancer gene co-expression in human endometrial cancer. *BMC genomics*, *15*(1), 1.

Dudbridge, F. (2013). Power and predictive accuracy of polygenic risk scores. *PLOS Genetics*, *9*(3), e1003348.

European Molecular Biology Laboratory (2015-2016). The european bioinformatics institute. Récupéré de `http://www.ebi.ac.uk/`

Fridman, J. S. et Lowe, S. W. (2003). Control of apoptosis by p53. *Oncogene*, *22*(56), 9030–9040.

Friedman, J., Hastie, T. et Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, *33*(1), 1–22. Récupéré de `http://www.jstatsoft.org/v33/i01/`

Hastie, T., Tibshirani, R. et Friedman, J. (2009). *The Elements of Statistical Learning : Data Mining, Inference, and Prediction* (2e éd.). Springer-Verlag New York.

Hoerl, A. E. et Kennard, R. W. (1970). Ridge regression : Biased estimation for nonorthogonal problems. *Technometrics*, *12*(1), 55–67.

Horvath, S. (2011). *Weighted Network Analysis* (1ère éd.). [Document électronique]. Springer-Verlag New York.

Hosmer, D. W. et Lemeshow, S. (2000). *Applied Logistic Regression* (2e éd.). John Wiley & Sons, Inc.

Kyoto University Bioinformatics Center (1995-2016). Kegg : Kyoto encyclopedia of genes and genomes. Récupéré de `http://www.genome.jp/`

Langfelder, P. et Horvath, S. (2008a). Tutorials for the wgcna package. Récupéré de `https://labs.genetics.ucla.edu/horvath/CoexpressionNetwork/Rpackages/WGCNA/Tutorials/`

Langfelder, P. et Horvath, S. (2008b). Wgcna : an r package for weighted correlation network analysis. *BMC bioinformatics*, *9*(1), 1.

Le Cessie, S. et Van Houwelingen, J. C. (1992). Ridge estimators in logistic regression. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, *41*(1), 191–201.

National Human Genome Research Institute. (2015). *Biological Pathways*. Récupéré de `https://www.genome.gov/27530687`

Rencher, A. C. (2002). *Methods of Multivariate Analysis*. John Wiley & Sons, Inc.

Roever, C., Raabe, N., Luebke, K., Ligges, U., Szepannek, G. et Zentgraf, M. (2014). Package klar.

Société canadienne du cancer (2016). Leucémie chez l'enfant. Récupéré de `http://www.cancer.ca/fr-ca/?region=pe`

Soussi, T. (1994-2012). The tp53 web site. Récupéré de `http://p53.free.fr/`

Tabachnick, B. et Fidell, L. (2007). *Using Multivariate Statistics* (5e éd.). Pearson Education.

Tamassia, R. (2013). *Handbook of graph drawing and visualization*. CRC press.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 267–288.

Tkačik, G. et Bialek, W. (2009). *Encyclopedia of Complexity and Systems Science*, chapitre Cell Biology : NetworksNetwork, RegulationRegulation and Pathways-Pathways, 719–741. Springer New York : New York, NY

Tseng, P. (2001). Convergence of a block coordinate descent method for nondif-

ferentiable minimization. *Journal of Optimization Theory and Applications*, *109*(3), 475–494.

Witten, D. M. et Tibshirani, R. (2011). Penalized classification using fisher's linear discriminant. *Journal of the Royal Statistical Society : Series B (Statistical Methodology*, *73*(5), 753–772.

Wu, J. (2012). *Advances in K-means clustering : a data mining thinking*. Springer Science & Business Media.

Zhang, B. et Horvath, S. (2005). A general framework for weighted gene co-expression network analysis. *Statistical applications in genetics and molecular biology*, *4*(1).

Zhang, B., Kirov, S. et Snoddy, J. (2005). Webgestalt : an integrated system for exploring gene sets in various biological contexts. *Nucleic Acids Research*, *33*, W741–W748.

Zou, H. et Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, *67*(2), 301–320.