Stabilité de la sélection de variables pour la régression et la classification de données corrélées en grande dimension

Emeline Perthame

ANNÉE 2015

THÈSE / UNIVERSITÉ DE RENNES 1
sous le sceau de l'Université Européenne de Bretagne

pour le grade de

DOCTEUR DE L'UNIVERSITÉ DE RENNES 1

Mention : Mathématiques et applications
Ecole doctorale Matisse

présentée par

Emeline Perthame

Préparée à l'IRMAR (UMR CNRS 6625)
Institut de Recherche Mathématique de Rennes
Laboratoire de Mathématiques Appliquées d'Agrocampus-Ouest

Stabilité de la sélection de variables pour la régression et la classification de données corrélées en grande dimension

Thèse soutenue à Agrocampus-Ouest le 16 octobre 2015

devant le jury composé de :

Stéphane ROBIN
DR, AgroParisTech/INRA / rapporteur
Korbinian STRIMMER
PR, Université de Leipzig / rapporteur
Michel DELECROIX
PR, UPMC / examinateur
Anne PHILIPPE
PR, Université de Nantes / examinatrice
Sylvain SARDY
PR, Université de Genève / examinateur
David CAUSEUR
PR, Agrocampus-Ouest / directeur de thèse

REMERCIEMENTS

I would like to sincerely thank Stéphane Robin and Korbinian Strimmer for reporting this thesis as well as Sylvain Sardy, Anne Philippe and Michel Delecroix for their kind participation as examiners in the Ph.D. defense.

I extend my thanks to Ching-Fan Sheu, for enabling this successful collaboration between the Department of statistics of Agrocampus-Ouest and the National Cheng Kung University at Taiwan, for his expertise and for providing the data which stimulated the statistical questions of this thesis.

Je remercie très sincèrement David Causeur, mon directeur de thèse, pour tout ce qu'il m'a appris d'un point de vue scientifique mais aussi pour ses précieuses qualités humaines. Je te remercie pour ta confiance, car tu m'as orientée tout en me laissant libre dans mes choix. Merci infiniment d'être toujours encourageant, rassurant et de t'être démené pendant mon stage de M2 pour trouver un financement de thèse.

Je voudrais adresser des remerciements chaleureux à Michel Delecroix pour avoir pris soin des nombreuses promotions de l'ISUP en tant que directeur et enseignant et plus personnellement, pour m'avoir mise en contact avec David lors de ma recherche de stage de M2.

J'adresse également des remerciements à l'ensemble des membres du département de mathématiques appliquées d'Agrocampus-Ouest : Karine et Elizabeth, sans qui le département ne tournerait pas aussi bien, François, Julie, Sébastien et Mathieu avec qui j'ai partagé mon quotidien : pauses cafés, pauses déjeuners, repas de Noël ou encore anniversaires. Mine de rien, je me suis attachée à vous tous. Je retiens de ces 3 années passées parmi vous une ambiance amicale à laquelle vous participez tous à votre manière. Enfin, j'adresse un petit mot particulier à Magalie dont j'ai partagé le bureau pendant ces trois ans. Je n'ai qu'une chose à dire : merci beaucoup, c'était vraiment (vraiment) super d'être ta co-bureau !

Je remercie ma collègue (et grande soeur de thèse) Chloé. Nous avons essentiellement travaillé à distance et j'ai apprécié nos coups de téléphone-points avancement d'article-papotages divers et variés. Tu m'as beaucoup appris et je te remercie pour ta patience, ta disponibilité et pour être toujours rassurante.

Je ne peux écrire ces remerciements sans évoquer Vincent et Tam, mes acolytes tout au long de ces 3 ans : on est arrivé presque en même temps à Rennes et on va en partir presque en même temps. J'ai passé avec vous de très bons moments au bureau, au restau, au cinéma, à l'accrobranche etc. J'attends avec impatience nos retrouvailles lors d'un voyage au Vietnam par exemple ! Je pense aussi aux quelques jeunes rencontrés à Rennes et avec qui j'ai tissé (tricoté ?) une belle amitié, Leslie et Guillaume et plus particulièrement Marie (que je remercie entre autres *pour ce magnifique gabarit*) et Margot, à qui je souhaite bon courage pour les années à venir ! J'adresse une pensée à Samuel et Cyril, les copains de l'ENSAI, à mon petit réseau parisien et grenoblois : Cécile, Eric, Xavier et Tim et à tous ceux que je n'ai pas cité dans ces remerciements et avec qui j'ai passé de bons moments de passage à Rennes, en conférence deci delà ou ailleurs.

Je n'oublie pas de remercier de tout cœur mes parents qui me soutiennent toujours sur tout et ma petite soeur (qui n'est plus petite depuis longtemps) et dont je suis très fière.

Et enfin pour finir, merci à Gaspar.

Stabilité de la sélection de variables pour la régression et la classification de données corrélées en grande dimension

Les données à haut-débit, par leur grande dimension et leur hétérogénéité, ont motivé le développement de méthodes statistiques pour la sélection de variables. En effet, le signal est souvent observé simultanément à plusieurs facteurs de confusion. Les approches de sélection habituelles, construites sous l'hypothèse d'indépendance des variables, sont alors remises en question car elles peuvent conduire à des décisions erronées.

L'objectif de cette thèse est de contribuer à l'amélioration des méthodes de sélection de variables pour la régression et la classification supervisée, par une meilleure prise en compte de la dépendance entre les statistiques de sélection. L'ensemble des méthodes proposées s'appuie sur la description de la dépendance entre covariables par un petit nombre de variables latentes. Ce modèle à facteurs suppose que les covariables sont indépendantes conditionnellement à un vecteur de facteurs latents.

Une partie de ce travail de thèse porte sur l'analyse de données de potentiels évoqués (ERP). Les ERP sont utilisés pour décrire par électro-encéphalographie l'évolution temporelle de l'activité cérébrale. Sur les courts intervalles de temps durant lesquels les variations d'ERPs peuvent être liées à des conditions expérimentales, le signal psychologique est faible, au regard de la forte variabilité inter-individuelle des courbes ERP. En effet, ces données sont caractérisées par une structure de dépendance temporelle forte et complexe. L'analyse statistique de ces données revient à tester pour chaque instant un lien entre l'activité cérébrale et des conditions expérimentales. Une méthode de décorrélation des statistiques de test est proposée, basée sur la modélisation jointe du signal et de la dépendance à partir d'une connaissance préalable d'instants où le signal est nul.

Ensuite, l'apport du modèle à facteurs dans le cadre général de l'Analyse Discriminante Linéaire est étudié. On démontre que la règle linéaire de classification optimale conditionnelle aux facteurs latents est plus performante que la règle non-conditionnelle. Un algorithme de type Expectation-Maximization pour l'estimation des paramètres du modèle est proposé. La méthode de décorrélation des données ainsi définie est compatible avec un objectif de prédiction.

Enfin, on aborde de manière plus formelle les problématiques de détection et d'identification de signal en situation de dépendance. On s'intéresse plus particulièrement au Higher Criticism (HC), défini sous l'hypothèse d'un signal rare de faible amplitude et sous l'indépendance. Il est montré dans la littérature que cette méthode atteint des bornes théoriques de détection. Les propriétés du HC en situation de dépendance sont étudiées et les bornes de détectabilité et d'estimabilité sont étendues à des situations arbitrairement complexes de dépendance. Dans le cadre de l'identification de signal, une adaptation de la méthode Higher Criticism Thresholding par décorrélation par les innovations est proposée.

Mots clés : grande dimension, dépendance, sélection de variables, modèle à facteurs latents, régression, classification supervisée, tests multiples

Stability of variable selection in regression and classification issues for correlated data in high dimension

The analysis of high throughput data has renewed the statistical methodology for feature selection. Such data are both characterized by their high dimension and their heterogeneity, as the true signal and several confusing factors are often observed at the same time. In such a framework, the usual statistical approaches are questioned and can lead to misleading decisions as they are initially designed under independence assumption among variables.

The goal of this thesis is to contribute to the improvement of variable selection methods in regression and supervised classification issues, by accounting for the dependence between selection statistics. All the methods proposed in this thesis are based on a factor model of covariates, which assumes that variables are conditionally independent given a vector of latent variables.

A part of this thesis focuses on the analysis of event-related potentials data (ERP). ERPs are now widely collected in psychological research to determine the time courses of mental events. In the significant analysis of the relationships between event-related potentials and experimental covariates, the psychological signal is often both rare, since it only occurs on short intervals and weak, regarding the huge between-subject variability of ERP curves. Indeed, this data is characterized by a temporal dependence pattern both strong and complex. Moreover, studying the effect of experimental condition on brain activity for each instant is a multiple testing issue. We propose to decorrelate the test statistics by a joint modeling of the signal and time-dependence among test statistics from a prior knowledge of time points during which the signal is null.

Second, an extension of decorrelation methods is proposed in order to handle a variable selection issue in the linear supervised classification models framework. The contribution of factor model assumption in the general framework of Linear Discriminant Analysis is studied. It is shown that the optimal linear classification rule conditionally to these factors is more efficient than the non-conditional rule. Next, an Expectation-Maximization algorithm for the estimation of the model parameters is proposed. This method of data decorrelation is compatible with a prediction purpose.

At last, the issues of detection and identification of a signal when features are dependent are addressed more analytically. We focus on the Higher Criticism (HC) procedure, defined under the assumptions of a sparse signal of low amplitude and independence among tests. It is shown in the literature that this method reaches theoretical bounds of detection. Properties of HC under dependence are studied and the bounds of detectability and estimability are extended to arbitrarily complex situations of dependence. Finally, in the context of signal identification, an extension of Higher Criticism Thresholding based on innovations is proposed.

Keywords : high dimension, dependence, variable selection, factor model, regression, supervised classification, multiple testing

TABLE DES MATIÈRES

CHAPITRE 1

INTRODUCTION

RÉSUMÉ : le recours de plus en plus fréquent à des technologies produisant des données à haut-débit - comme la spectroscopie proche infra-rouge, l'imagerie par résonance magnétique fonctionnelle ou l'électro-encéphalographie - a généré de nouvelles questions de recherche en statistique spécifiques de ces données dites de grande dimension, caractérisées par leur nombre de variables très supérieur à celui des individus. En particulier, un grand nombre de méthodes de prédiction, fondées sur des modèles de régression ou de classification supervisée, se sont développées en s'appuyant sur une hypothèse dite de parcimonie des modèles. En effet, ces méthodes supposent que peu de prédicteurs mesurés sont pertinents. Dès lors, une part importante de la problématique d'ajustement d'un modèle de prédiction en grande dimension repose sur une étape de sélection de ces variables. Un très grand nombre de méthodes de sélection ont ainsi été définies avec pour objectif essentiel de garantir une bonne performance de prédiction, le plus souvent sans se soucier de la pertinence des prédicteurs sélectionnés ou encore de la reproductibilité de la sélection. Cependant, la très haute résolution des données à haut-débit se traduit souvent par une grande dépendance entre les variables, dépendance affectant à la fois les performances de prédiction mais aussi la stabilité des méthodes de sélection de variables. L'objectif de cette introduction est de présenter différentes approches de prise en compte de la dépendance dans les procédures de sélection de variables, et ainsi de montrer qu'il est possible de tirer avantage de la corrélation pour améliorer l'estimation du support d'un signal.

Sommaire

1 Contexte

1.1 Sélection de variables

La problématique définissant le cadre général de cette thèse est la prise en compte de la dépendance dans les procédures de sélection de variables pour la prédiction en grande dimension, en régression et en classification supervisée. Dans la plupart des situations abordées ci-après, les données peuvent être décrites comme une série de n couples indépendants (X, Y) composés d'un profil de prédicteurs $X = (X_1, \ldots, X_m)$ de dimension $m \gg n$ et d'une variable réponse Y, soit quantitative soit catégorielle.

L'identification d'un sous-ensemble pertinent de prédicteurs est un des objectifs majeurs d'une analyse de régression ou de classification supervisée, et ce même en situation de "petite dimension" ($n \geq m$). Dans ce contexte plus traditionnel, on recense plusieurs méthodes de sélection consistant à comparer les modèles construits sur des sous-ensembles de prédicteurs selon un critère de qualité d'ajustement pénalisé par le nombre de prédicteurs. Ainsi, dans le contexte du modèle linéaire généralisé, la minimisation des critères AIC (introduit par Akaike (1973)) ou BIC (proposé par Schwarz (1978)), versions pénalisés de la déviance du modèle par la norme ℓ_0 du vecteur β des paramètres de régression, préfigurent les méthodes d'estimation par régularisation devenues si populaires pour les données de grande dimension, pour lesquelles la pénalisation est plus volontiers définie par les normes ℓ_1 (Tibshirani (1996)) ou ℓ_2 (Hoerl and Kennard (1970)) de β.

En effet, l'optimisation de critères pénalisés par :

$$||\beta||_0 \quad = \#\left\{j \in [1; m],\ \beta_j \neq 0\right\},$$

où $\#A$ désigne le cardinal d'un ensemble A, nécessite l'ajustement de tous les sous-modèles possibles (2^m modèles), ce qui pose des problèmes numériques, pour des valeurs mêmes modérées de m. Certes, la sélection pas à pas constitue une alternative raisonnable d'un point de vue calculatoire, mais le parcours par cet algorithme séquentiel d'une part très faible du graphe des sous-modèles, au mieux $m(m+1)/2$ sous-modèles, génère une instabilité de la procédure, d'autant plus grande que m est lui-même grand (voir Breiman (1996) et Fan and Li (2001)).

1.2 Grande dimension

Les progrès technologiques en terme de recueil et de stockage de données, notamment en biologie moléculaire pour l'étude du génome (voir par exemple Shalon et al. (1996) pour les puces à ADN et Baron et al. (2006) pour l'étude épigénétique de la méthylation de l'ADN), ou en neurosciences, pour l'analyse de l'activité cérébrale par électro-encéphalographie (Handy (2004)) ou imagerie par résonance magnétique (Poldrack et al. (2011)), ont conduit à des évolutions importantes de la méthodologie statistique pour l'adapter à des situations caractérisées par un nombre important de variables. Dans les cas abordés dans cette thèse, le nombre de variables est de l'ordre de plusieurs milliers. Les méthodes classiques, notamment celles dont l'objectif est l'identification de variables d'intérêt par sélection ou tests multiples, ont des propriétés analytiques éprouvées en situation asymptotique, lorsque le nombre n d'individus tends vers l'infini et que le nombre de variables est fixe. Cependant, ces méthodes se montrent peu performantes en grande dimension. Par exemple, la propriété de consistance d'estimation du support par le critère BIC (Shao (1997), Yang (2005)) se perd lorsque le nombre de variables n'est pas fixé (voir par exemple Broman and Speed (2002), Casella et al. (2009), Kim et al. (2012)). A l'instar de la problématique abordée plus haut pour évoquer la sélection pas à pas, un des problèmes est l'explosion combinatoire des associations possibles de variables sélectionnées, qui nécessite aussi le contrôle par des méthodes adaptées du nombre de sélections erronées, ou faux positifs. Une autre raison plus spécifique du paradigme $n \ll m$ est liée à l'instabilité voire l'impossibilité numérique de l'ajustement de modèles dont le nombre de paramètres dépasse celui des individus par des méthodes impliquant le plus souvent l'inversion de la matrice de variance-covariance des prédicteurs (par exemple la méthode des moindres carrés en régression). Ainsi, au-delà de la recherche de solutions statistiques performantes, un des défis de l'analyse de données de grande dimension est également la simplicité algorithmique des méthodes, garantissant la possibilité effective de leur mise en œuvre.

La sélection de prédicteurs pertinents s'apparente à la problématique souvent associée aux tests multiples, dont le but est une identification aussi complète que possible du support du signal, tout en contrôlant le nombre de prédicteurs sélectionnés par erreur. Les premières réflexions autour de ces questions de contrôle du taux d'erreur de type I pour un grand nombre de tests ont conduit à revoir l'objectif d'un contrôle de la probabilité d'un faux positif, le Family-Wise Error Rate (FWER), pour s'orienter vers un objectif moins conservateur de contrôle de la proportion de faux positifs dans l'ensemble sélectionné, le False Discovery Rate (FDR). La méthode de référence pour le contrôle du FDR, la procédure de Benjamini-Hochberg

(Benjamini and Hochberg (1995)), s'est ainsi imposée comme une méthode standard en analyse de données génomiques (voir par exemple l'ouvrage de van der Laan and Dudoit (2007)), préférée à la méthode de Bonferroni (Bonferroni (1936)) plus traditionnellement utilisée lorsque le nombre de tests est plus modérés. Benjamini and Hochberg (1995) démontre que leur méthode de détermination du seuil de sélection sur les statistiques de tests contrôle effectivement le FDR sous une hypothèse d'indépendance ou de faible dépendance. La dépendance étant dès lors perçue comme un obstacle potentiel au contrôle du FDR, de nombreux auteurs se sont attachés à étendre la méthode de Benjamini-Hochberg de telle sorte qu'elle garantisse le contrôle du FDR sous certaines hypothèses de dépendance (voir par exemple Benjamini and Yekutieli (2001)). En pratique, ces approches de protection contre les effets de la dépendance sur le contrôle du FDR ont le plus souvent conduit à des méthodes très conservatives. Plus récemment, s'appuyant sur la démonstration que le classement des statistiques de test en situation de dépendance n'est pas consistant, au sens statistique où il n'est pas conforme à l'amplitude du signal testé, quelques auteurs ont privilégié une autre approche, ne visant pas à une modification de la procédure de Benjamini-Hochberg, mais à une décorrélation des statistiques de test (voir Zuber and Strimmer (2009) et Hall and Jin (2010) pour une procédure de tests ajustés sur la corrélation, Kustra et al. (2006), Leek and Storey (2007), Carvalho et al. (2008), Friguet et al. (2009), Sun et al. (2012) et plus récemment Allen et al. (2014) et Houseman et al. (2015) pour une modélisation par des facteurs latents de la dépendance). Les différentes méthodes se différencient essentiellement par le modèle de variance utilisé et surtout par la technique d'estimation jointe du signal et de la variance.

A l'instar des procédures de tests multiples, Donoho and Jin (2004) définissent une procédure de sélection de variables pour la détection d'un signal, le Higher Criticism Thresholding (HCT). Les auteurs s'appuient sur l'idée proposée par Tukey (1976) que la détection statistique d'un signal, c'est à dire le test global de son existence, peut reposer sur le vecteur des statistiques de test des composantes individuelles de ce signal. En situation d'indépendance entre les statistiques de sélection et dans le cadre général d'un signal à la fois rare et faible (paradigme "Rare-and-Weak"), Donoho and Jin (2008) démontrent l'optimalité de cette procédure de sélection, au sens où elle atteint les bornes optimales de détection de Ingster (1997). Hall and Jin (2008) et Hall and Jin (2010) montrent que ces résultats théoriques sont fortement affectés par une dépendance entre les statistiques de sélection et proposent une extension à des cas particuliers de dépendance, dont la structure auto-régressive d'ordre 1. Par ailleurs, Ahdesmäki and Strimmer (2010) et Klaus and Strimmer (2013) étudient les propriétés de la méthode HCT pour l'identification du signal, à savoir l'estimation de son support et démontrent son équivalence avec une procédure de tests multiples contrôlant le False Non-Discovery Rate (FNDR).

La diversité des approches de prise en compte de la dépendance traduit de profondes divergences dans la communauté statistique, partagée entre une démarche naïve consistant à ignorer la corrélation et un point de vue opposé justifiant une modélisation jointe de la variance et de l'espérance pour améliorer l'identification du signal. Ainsi, en particulier dans un contexte d'analyse discriminante linéaire, les

tenants d'une approche dite *naive Bayes* montrent la supériorité de ce point de vue en terme de performance de classification (voir notamment Tibshirani et al. (2003), Bickel and Levina (2004), Efron (2008)). Ces méthodes reposent sur une hypothèse erronée d'indépendance entre les variables dont une alternative consiste à estimer la matrice de covariance par des méthodes de *shrinkage*, sans hypothèse particulière de structure de la dépendance. Le principe de ces méthodes est de s'affranchir de la propriété de non biais de l'estimateur empirique pour diminuer la variance d'estimation. Ainsi, l'estimateur *ridge* (Hoerl and Kennard (1970)) du vecteur β des coefficients de régression linéaire, qui résulte de la minimisation de la déviance du modèle pénalisée par $||\beta||_2$, prend une forme similaire à celle de l'estimateur des moindres carrés, où la matrice de covariance empirique S est remplacée par l'expression suivante :

$$\hat{\Sigma}_\gamma = S + \gamma \mathbb{I}_m,$$

où $\mathbb{I}_m$ désigne la matrice identité d'ordre m et $\gamma \geq 0$. Le paramètre γ de régularisation introduit ci-dessus permet bien un compromis entre deux points de vue extrêmes de la dépendance, à savoir l'indépendance pour de grandes valeurs de γ et la structure de covariance la plus complexe estimée par S pour $\gamma = 0$. On retrouve cette idée dans de nombreuses méthodes de régression ou de classification supervisée, dont dans les *Correlation Adjusted T-scores* (CAT-scores, Zuber and Strimmer (2009)) utilisés dans l'étape de sélection de variables de la méthode *Shrinkage Discriminant Analysis* (SDA, Ahdesmäki and Strimmer (2010)). Ici, les auteurs proposent une expression analytique pour un estimateur du paramètre de *shrinkage* γ. Cette idée se retrouve aussi dans la méthode *shrunken centroids regularized discriminant analysis* (SCRDA, Guo et al. (2007)), dans laquelle la matrice de covariance empirique est remplacée par

$$\hat{\Sigma}_\alpha = \alpha S + (1 - \alpha)\mathbb{I}_m,$$

où $0 \leq \alpha \leq 1$.

L'estimation par *shrinkage*, du type de la méthode *ridge*, apporte une solution essentiellement numérique à la problématique de la grande dimension, qui se montre souvent performante en terme de précision de la règle de décision qui s'en déduit. L'estimation *ridge* de modèles de régression ou de classification supervisée s'impose notamment comme la référence pour les questions relatives à la sélection génomique, dont l'objectif est l'estimation de la valeur génétique d'un animal ou d'une plante à partir de données de génotypage à l'échelle de son génome. Toutefois, la recherche de zones d'intérêt du génome appelés Quantitative Trait Loci (QTL) ou de manière équivalente la recherche de la signature moléculaire associée à un stress d'intérêt de l'organisme nécessite des approches plus exigeantes dont l'objectif est certes de garantir une bonne prédiction mais aussi d'identifier les leviers de cette prédiction, en d'autres termes les prédicteurs pertinents. La prise en compte simultanée de ces deux objectifs par la méthode dite LASSO, pour *Least Absolute Shrinkage and Selection Operator* (Tibshirani (1996)), qui consiste à minimiser un critère d'ajustement, la déviance par exemple, pénalisé par $||\beta||_1$ explique sa grande popularité dans de nombreux domaines associés à des technologies à haut débit. Dans le contexte de

la classification supervisée, des méthodes d'analyse linéaire discriminante pénalisée ont été développées (voir Tibshirani et al. (2002) pour une version *ridge* des plus proches voisins ou Witten and Tibshirani (2011) et Clemmensen et al. (2011) pour une approche lasso).

Pourtant, une dépendance forte entre prédicteurs affecte notoirement les propriétés de la méthode LASSO, notamment dans sa capacité à déterminer le support d'un signal (Van de Geer (2010), Fan and Lv (2010)). La pénalisation par combinaison convexe de $||\beta||_1$ et $||\beta||_2$ dans la méthode *elastic net* (Zou and Hastie (2005)) vise justement à apporter plus de stabilité à la méthode. D'autres extensions plus récentes, basées sur du ré-échantillonnage, ont directement visé à améliorer la reproductibilité de la sélection par LASSO, en cherchant à réduire le sous-ensemble des variables sélectionnées à celles les plus souvent retenues (voir notamment Bach (2008) pour la méthode bolasso et Meinshausen and Bühlmann (2010) pour la méthode *stability selection*).

1.3 Illustrations de situations de dépendance

On présente ici quelques situations dans lesquelles on cherche à identifier un signal biologique. Ce signal est assimilable à un lien entre une variable réponse quantitative ou catégorielle et des variables explicatives nombreuses, mesurées par une technologie à haut débit, et présentant une structure de dépendance forte. Dans un premier temps, on s'intéresse à des données mesurées par électroencéphalographie (EEG) de l'activité du cerveau en psychologie expérimentale. Ces données de potentiels évoqués, ou encore ERP (Event-Related Potentials, Handy (2004)), décrivent avec une très forte résolution, jusqu'à une mesure toute les demi-millisecondes (Groppe et al. (2011a) et Groppe et al. (2011b)), l'activité cérébrale en des électrodes localisées très précisément sur le crâne, pour un nombre limité de sujets, entre 10 et 20 en général. L'analyse de ces données vise généralement à identifier les intervalles de temps pour lesquels l'association avec une réponse expérimentale, par exemple un score évaluant un comportement ou l'appartenance à une catégorie particulière de population, est significative. Une expérience ayant fait l'objet d'une collaboration avec National Cheng-Kung University, Tainan (Taiwan) et le problème cognitif associé sont détaillés dans le Chapitre 2.

La Figure 1.1 révèle une structure de dépendance temporelle forte entre les mesures de l'activité cérébrale : l'histogramme montre qu'une grande proportion des corrélations entre les ERPs mesurés sur l'électrode CZ (milieu de la région centrale de la tête) sont élevées et positives. En gris, la distribution des corrélations d'une matrice de même dimension sous l'hypothèse d'indépendance est tracée pour comparaison. On remarque une forte asymétrie à droite de la distribution. D'après l'image de la matrice des corrélations, il semble que l'auto-corrélation génère un grand nombre de corrélations proches de 1 sur les bandes proches de la diagonale. On remarque aussi des blocs de corrélations élevées et positives, correspondant à une synchronisation de l'activité cérébrale sur des intervalles de temps, et une auto-corrélation croissante au cours du temps. La structure est donc plus complexe que celle produite par un processus auto-régressif d'ordre 1 souvent utilisé pour

modéliser de telles données (voir Yeung et al. (2004), Guthrie and Buchwald (1991) et Bugli and Lambert (2006)). Lors de l'expérience, l'activité cérébrale est mesurée sur plusieurs électrodes placées sur le crâne du sujet. Sur les autres électrodes, on remarque des structures similaires.

La prise en compte de la dépendance dans les études d'association ou de sélection suscite également de nombreux développements en matière d'analyse de données génomiques, en particulier pour l'analyse du transcriptome à partir de *microarrays* (Shalon et al. (1996)). On se réfère notamment à Lee and Batzoglou (2003) et Teschendorff et al. (2011) pour une application de l'analyse en composantes indépendantes, Schäfer and Strimmer (2005), Opgen-Rhein and Strimmer (2007), Zuber and Strimmer (2009) et Ahdesmäki and Strimmer (2010) pour une définition de statistiques de tests décorrélées (CAT-scores) par un estimateur de type James-Stein de la matrice de covariance et Kustra et al. (2006), Leek and Storey (2007),Carvalho et al. (2008), Friguet et al. (2009), Sun et al. (2012) et plus récemment Allen et al. (2014) pour une modélisation par des facteurs latents de la dépendance. Cette même approche est aussi utilisée très récemment pour l'analyse de données de méthylation de l'ADN par Houseman et al. (2015). On propose d'explorer dans ce paragraphe la distribution des corrélations d'une série de données publiques, utilisées à des fins d'illustration dans des packages `R` ou `Matlab`, pour des méthodes de classification supervisée. Ces données ont été choisies pour la variété des situations qu'elles représentent, notamment par leurs dimensions résumées dans la Table 1.1, ainsi que le nombre de classes de la variable réponse. Ces données sont associées à des problématiques d'études du cancer du colon (Alon et al. (1999)), du sein (West et al. (2001)), de la leucémie (Golub et al. (1999)), du lymphome (Chung and Keles (2010)), du cancer de la prostate (Singh et al. (2002)) et de cancers chez l'enfant auquel on se réfère dans la suite par SRBCT (Khan et al. (2001)).

La Figure 1.2 présente en noir les histogrammes des corrélations entre variables (gènes) pour chacune des situations introduites ci-dessus et en bleu la distribution des corrélations pour des données de mêmes dimensions sous l'hypothèse d'indépendance. On remarque une diversité de profils de dépendance, certaines s'éloignant de manière remarquable de l'indépendance, comme pour le cancer du colon, Figure 1.2(a), le cancer du sein, Figure 1.2(b) et dans une moindre mesure sur la leucémie, Figure 1.2(c) pour lesquelles la distribution des corrélations révèle une sur-représentation de corrélations fortes et positives. La distribution des corrélations pour les données de lymphome, Figure 1.2(d) et SRBCT, Figure 1.2(e) semble en revanche symétrique avec une proportion notable de corrélations modérées. Enfin, il est intéressant de remarquer que les corrélations entre gènes pour le cancer de la prostate, Figure 1.2(f), s'ajustent bien à la distribution des corrélations sous l'indépendance. Ces exemples illustrent que, pour une même technologie et des problématiques similaires, des profils de dépendance très différents peuvent être observés, qu'il convient de prendre en compte lors de l'analyse statistique. De nombreux auteurs ont récemment émis l'hypothèse que ces dépendances résultent d'effets latents de processus biologiques non maitrisés par les dispositifs expérimentaux, susceptibles de masquer partiellement le signal biologique d'intérêt. Ces propos sont

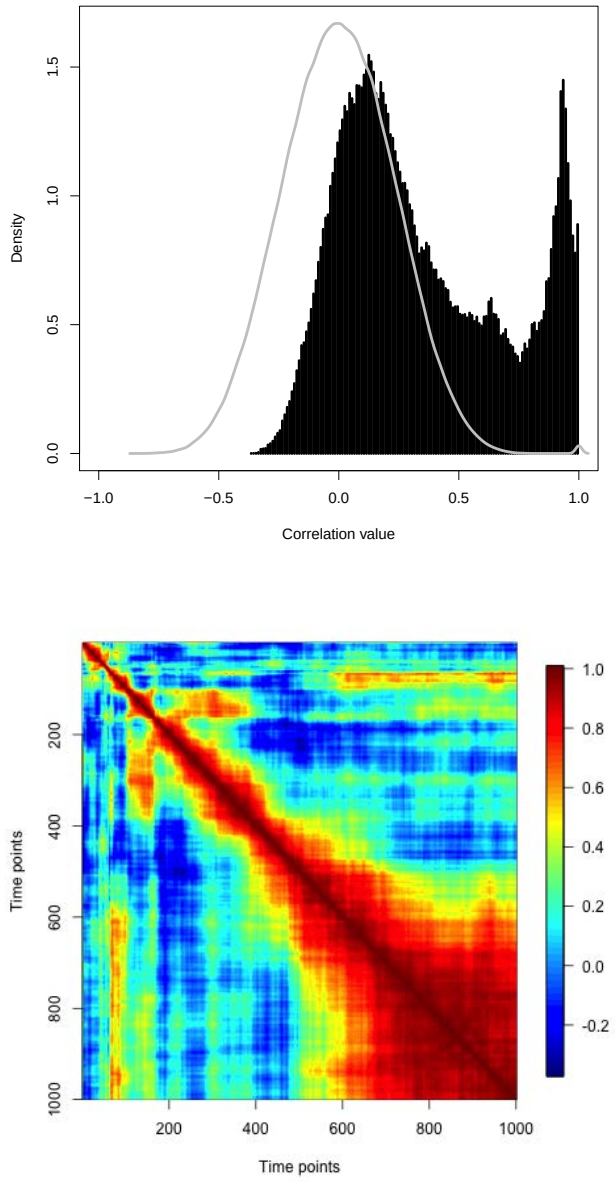


FIGURE 1.1 – Haut : histogramme des corrélations résiduelles entre les mesures de l'activité cérébrale pour les données d'ERPs mesurées à l'électrode `CZ` (en noir) comparé à la distribution des corrélations d'une matrice de mêmes dimensions sous hypothèse d'indépendance. Bas : image de la matrice des corrélations

TABLE 1.1 – Dimensions de données publiques associées à des études sur le cancer.

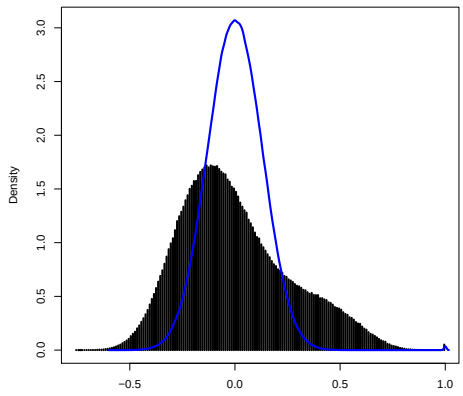
	Colon	Sein	Leucémie	Lymphome	Prostate	SRBCT
Nbre de var.	2000	7129	7129	4026	6033	2308
Nbre d'obs.	62	44	38	62	102	63
Nbre de classes	2	2	2	3	2	4

en particulier tenus par Kustra et al. (2006), Leek and Storey (2007), Pournara and Wernisch (2007), Carvalho et al. (2008), Friguet et al. (2009) et plus récemment par Sun et al. (2012) et Houseman et al. (2015).

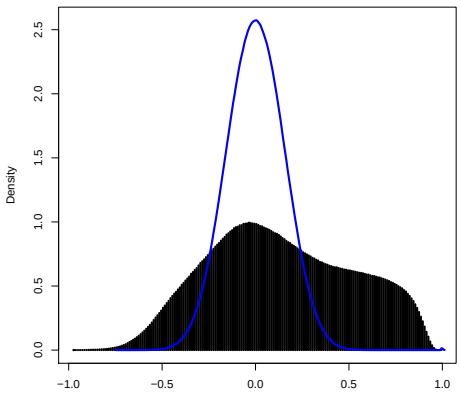
Dans la suite du manuscrit, les exemples ci-dessus sont utilisés à des fins d'illustration de l'impact de la dépendance sur les procédures de sélection de variables. En particulier, le Chapitre 2 introduit la notion d'instabilité des procédures de tests multiples, principale conséquence d'une forte dépendance entre les statistiques de sélection. Le contrôle du taux de faux positifs n'étant en revanche pas remis en cause par les formes de dépendance temporelle étudiées dans ce Chapitre, la probabilité pour qu'une procédure de type Benjamini-Hochberg détecte un signal peut être très faible en situation de dépendance et, conditionnellement à la détection d'un signal, son identification, à savoir l'estimation de son support, est moins précise qu'en situation d'indépendance. De même, dans un problème de sélection de variables en classification supervisée abordé dans le Chapitre 3, on illustrera que la dépendance affecte à la fois le nombre de variables sélectionnées et le rang des variables sélectionnées par des méthodes d'estimation régularisée notamment la méthode Lasso (Tibshirani (1996)). Aussi, on observera sur des données réelles que l'ensemble des variables sélectionnées par ces méthodes n'est pas reproductible. Enfin, le Chapitre 4 est dédié à l'étude de la méthode HCT pour l'identification d'un signal, dans le paradigme "Rare-and-Weak" proposé par Donoho and Jin (2004). On montre que la méthode HCT est très conservative lorsque les variables sont très corrélées. Finalement, dans les problèmes de tests multiples comme dans ceux de sélection de variables, l'impact de la dépendance se traduit par une non-consistance du classement des variables par leur pouvoir prédictif ou discriminant.

2 Prise en compte de la dépendance

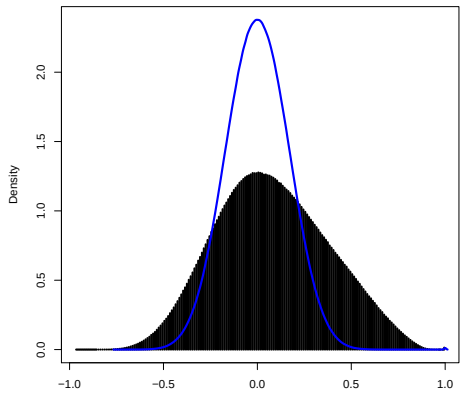
L'impact négatif de la dépendance sur la précision des procédures de tests multiples, en particulier due à l'instabilité du rang des statistiques de sélection, a suscité le développement de nombreuses approches innovantes. La dépendance entre les statistiques de tests ou de sélection étant directement héritée de celle entre les variables explicatives, ces approches visent essentiellement à estimer la structure de dépendance entre les variables pour construire des stratégies de décorrélation. Cette thèse vise à une contribution à l'optimisation de ces méthodes de décorrélation, dans le but de rétablir les propriétés théoriques et pratiques des méthodes élaborées sous l'hypothèse d'indépendance. On peut distinguer deux types d'approches pour la décorrélation. Le premier se rapproche des méthodes d'analyse de séries chronologiques, au sens où l'on cherche à construire la transformation linéaire des données



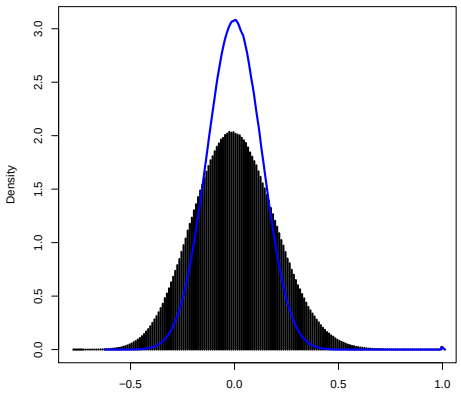
(a) Cancer du colon



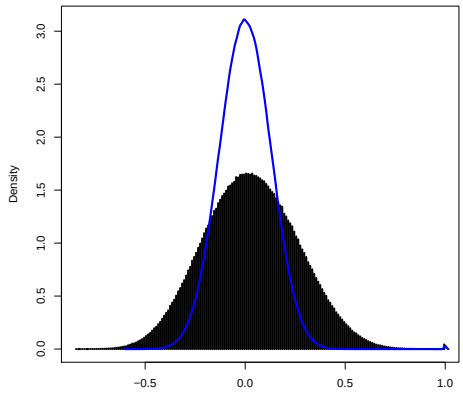
(b) Cancer du sein



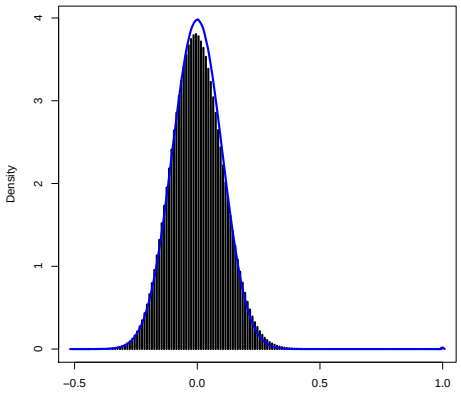
(c) Leucémie



(d) Lymphome



(e) SRBCT



(f) Cancer de la prostate

FIGURE 1.2 – Histogramme des corrélations entre variables (en noir) comparée à la distribution sous indépendance (en bleu) pour des données d'expression de gènes.

conduisant à des innovations indépendantes (Ahdesmäki and Strimmer (2010), Zuber and Strimmer (2009), Hall and Jin (2010)). Le second type s'appuie sur l'hypothèse d'effets latents affectant de manière linéaire la dépendance entre les statistiques de sélection (Friguet et al. (2009), Leek and Storey (2007), Leek and Storey (2008), Sun et al. (2012)).

2.1 Décorrélation par les innovations

Si l'on considère un vecteur de covariables X de matrice de covariance Σ, le principe général des méthodes de décorrélation par les innovations s'appuie sur l'existence d'une matrice L $(m \times m)$ telle que :

$$\Sigma^{-1} \quad = \quad LL',$$

où L' désigne la transposée de L. Cette matrice L permet de définir des covariables décorrélées X^*, appelées innovations, telles que :

$$X^* \quad = \quad L'X.$$

Ainsi, $\mathbb{V}(X^*) = L'\Sigma L = L'(L')^{-1}L^{-1}L = \mathbb{I}_m$.

Correlation-adjusted t-scores (CAT scores) On s'attarde dans ce paragraphe sur la présentation des CAT-scores, initialement proposés par Zuber and Strimmer (2009) dans un contexte de comparaisons multiples puis repris dans la méthode *Shrinkage Discriminant Analysis* (SDA, voir Ahdesmäki and Strimmer (2010)) en analyse linéaire discriminante.

On considère ici une variable réponse Y, catégorielle à K groupes et un $m-$profil x de variables explicatives distribué, conditionnellement au groupe $Y = y$, selon une loi normale d'espérance :

$$\mathbb{E}(x \mid Y = y) \quad = \quad \mu_y$$

et de même variance :

$$\mathbb{V}(x \mid Y = y) \quad = \quad \Sigma$$

pour tout $y \in \{1, \ldots, K\}$. On définit, pour chaque groupe $y \in \{1, \ldots, K\}$, des t-scores τ_y correspondant au m-vecteur des statistiques de test associées à la comparaison de la moyenne globale $\mu = \mathbb{E}(x)$ à μ_y :

$$\tau_y \quad = \quad \left\{ \left(\frac{1}{n_y} - \frac{1}{n} \right) D \right\}^{-1/2} (\mu_y - \mu),$$

où $D = \text{diag}(\Sigma)$, n est le nombre total d'observations indépendantes de (x, Y) et n_y est le nombre d'observations dans le groupe y. Zuber and Strimmer (2009) définissent les *correlation-adjusted t-scores* ou CAT-scores par :

$$\tau_y^{adj} \quad = \quad R^{-1/2}\tau_y,$$

où R désigne la matrice de corrélation de x telle que :

$$\Sigma = D^{1/2} R D^{1/2}.$$

Pour chaque variable explicative, une statistique du test d'association avec Y est calculée à partir d'une estimation par *shrinkage* des CAT-scores. Il est intéressant de noter qu'après décorrélation du profil des variables explicatives, la sélection de variables pour la classification supervisée s'apparente à un problème de tests multiples. Dans le cas présent, pour chaque variable, Ahdesmäki and Strimmer (2010) recommande d'estimer le taux de faux positifs (FDR) local et de sélectionner les variables explicatives par seuillage de ce FDR local, à un niveau suffisamment élevé pour garantir la sélection d'une part importante du support du signal.

Pour permettre le calcul des CAT-scores en grande dimension, Ahdesmäki and Strimmer (2010) proposent une méthode d'estimation par *shrinkage*, de type James-Stein. Ainsi, l'estimateur de la matrice de correlation prend la forme suivante :

$$\hat{R}_\gamma = \gamma \mathbb{I}_m + (1-\gamma)\hat{R},$$

où $\hat{R}$ est la matrice des corrélations empiriques et $0 \leq \gamma \leq 1$ est le paramètre de régularisation. Schäfer and Strimmer (2005) propose une expression analytique de la valeur optimale de γ, au sens de la performance de la méthode de classification supervisée. Cette approche par *shrinkage* permet le calcul explicite de $\hat{R}^\alpha_\gamma$, pour toute puissance α. Ainsi, Zuber and Strimmer (2009) définit la matrice Z suivante :

$$\begin{aligned} Z &= \frac{1}{\gamma}\hat{R}_\gamma \\ &= \mathbb{I}_m + UMU', \end{aligned}$$

où M est une matrice symétrique définie positive et U est une base orthonormale. La puissance α de la matrice Z peut-être calculée par la formule suivante :

$$Z^\alpha = \mathbb{I}_m - U(\mathbb{I}_r - (\mathbb{I}_r + M)^\alpha)U'$$

où r est le rang de M. Cette formule ne fait appel qu'au calcul de puissance de la matrice $\mathbb{I}_r + M$. Après sélection des variables explicatives selon la procédure décrite ci-dessus, l'étape de classification s'appuie sur une approche d'analyse discriminante linéaire pour laquelle tous les paramètres sont estimés par *shrinkage* (Hausseur and Strimmer (2009)). Cette méthode est implémentée dans le package `R sda` (Ahdesmäki et al. (2014)).

innovated HCT Comme mentionné ci-dessus, la décorrélation du profil des variables explicatives permet de considérer la question de la sélection de variables comme un problème de tests multiples, pour lequel la sélection d'une variable repose sur un seuillage des statistiques de tests ou des p-values associées. Alors que Ahdesmäki and Strimmer (2010) propose de définir le seuil en s'appuyant sur l'estimation des risques d'erreur de sélection, la méthode Higher Criticism Thresholding (HCT, voir Donoho and Jin (2015) pour une revue récente) repose sur la maximisation d'une fonction objectif. Si $(p_1, \ldots, p_m)$ désigne le vecteur des probabilités

critiques issues de tests d'hypothèses et $(p_{(1)}, \ldots, p_{(m)})$ leur statistique d'ordre, HCT définit le seuil optimal comme l'indice des p-values pour lequel l'écart entre la fonction de répartition empirique et celle de la loi uniforme, en d'autres termes la loi d'une p-value sous l'hypothèse nulle, est maximal :

$$j^* = \underset{j:1/m \leq p_{(j)} \leq 1/2}{\operatorname{argmax}} \left\{ \sqrt{m} \frac{j/m - p_{(j)}}{\sqrt{p_{(j)}(1-p_{(j)})}} \right\}.$$

L'ensemble des variables sélectionnées est $\{j, p_j \leq p_{(j*)}\}$. Les variantes et les propriétés de HCT sont présentées en détails dans le Chapitre 4.

En situation de dépendance, Hall and Jin (2010) s'appuie sur la factorisation de Cholesky de la matrice de covariance pour définir la méthode *innovated HCT* (iHCT), une variante de HCT. Hall and Jin (2010) suppose que le vecteur T des statistiques de test est gaussien tel que :

$$T = \mu + Z \text{ où } Z \sim \mathcal{N}(0, \Sigma) \text{ et } \Sigma \neq \mathbb{I}_m, \tag{1.1}$$

où le signal μ est un vecteur parcimonieux possédant k coordonnées non nulles parmi m, de même amplitude A_m. Les statistiques de tests sont décorrélées par la factorisation de Cholesky inverse de Σ telle que $U_m \Sigma U_m' = \mathbb{I}_m$ et Hall and Jin (2010) considère le problème suivant, équivalent à (1.1) :

$$U_m T = U_m \mu + U_m Z \text{ où } U_m Z \sim \mathcal{N}(0, \mathbb{I}_m). \tag{1.2}$$

Les résultats établis par Hall and Jin (2010) reposent sur une hypothèse sur la matrice de covariance particulièrement adaptée à des structures de dépendance temporelle de type auto-régressif et garantissant que le signal transformé $U_m \mu$ a un support similaire à celui de μ. Afin de satisfaire cette hypothèse, les auteurs proposent un lissage de la matrice de décorrélation U_m et définissent une matrice $\tilde{U}_m$ dont le terme général $\tilde{u}_{k,j}$ est défini par :

$$\tilde{u}_{k,j} = \begin{cases} u_{kj} & \text{si } k - b_m + 1 \leq j \leq k \\ 0 & \text{sinon,} \end{cases}$$

où u_{kj} est le terme général (k, j) de la matrice U_m et b_m est une fenêtre recommandée entre 1 et $\log(m)$. Enfin, les statistiques de tests T sont décorrélées par la matrice V_m définie par :

$$V_m(b_m) = \overline{U}_m' U_m,$$

où $\overline{U}_m$ est la matrice $\tilde{U}_m$ dont les colonnes ont été normalisées à 1. Si $b_m = 1$, *innovated HCT* revient à appliquer HCT aux statistiques de test décorrélées :

$$V_m X = V_m \mu + V_m Z,$$

dont de nouvelles probabilités critiques $(p_1, \ldots, p_m)$ sont déduites. Hall and Jin (2010) recommande $b_m = \log(m)$ et dans ce cas, le seuil optimal $iHC_m^*(b_m)$ de *innovated HCT* est défini comme suit :

$$iHC_m^*(b_m) = \frac{1}{\sqrt{2b_m - 1}} \max_{j:1/m \leq p_{(j)} \leq 1/2} \left\{ \sqrt{m} \frac{j/m - p_{(j)}}{\sqrt{p_{(j)}(1-p_{(j)})}} \right\}.$$

Hall and Jin (2010) démontre des propriétés d'optimalité de *innovated HCT* en terme de détection et d'estimation du signal sous certaines conditions sur Σ.

Enfin, on verra dans le Chapitre 4 que le modèle à facteurs présenté dans la section suivante permet également une expression analytique de la racine carrée de la matrice de covariance, afin de développer une procédure similaire au *innovated HCT* (Hall and Jin (2010)) ou aux CAT-scores (Zuber and Strimmer (2009)).

2.2 Modélisation de la structure de dépendance

Dans certains domaines d'application, en particulier celui de l'analyse de données génomiques, il est pertinent de considérer que la dépendance est structurée par l'effet de variables latentes. Le modèle correspondant à ce type de point de vue sur la dépendance, appelé modèle à facteurs, conduit à la réduction du rang de la matrice de variance-covariance entre les variables explicatives. Les composantes de la dépendance, si possible en nombre modéré, sont assimilables à autant de sources de variabilité générant une hétérogénéité de la distribution jointe des variables explicatives. En effet, l'effet de ces facteurs peut se confondre avec le vrai signal et expliquer la non-consistance du classement des p-values par des méthodes de tests qui ignorent la dépendance. Par exemple, en analyse de données génomiques, l'expression des gènes peut être associée à une condition expérimentale mais aussi activée par l'activité biologique de l'individu (la bibliographie est riche sur le sujet, voir Kustra et al. (2006), Leek and Storey (2008), Carvalho et al. (2008), Friguet et al. (2009), Teschendorff et al. (2011), Sun and Cai (2009), Pournara and Wernisch (2007), Blum et al. (2010)). Enfin, des articles récents illustrent l'apport du modèle à facteurs en épigénétique, sur des données de méthylation de l'ADN (Houseman et al. (2015)).

2.2.1 Modèle à facteurs

Le modèle à facteurs suppose l'existence de variables latentes, non observées, qui peuvent avoir un effet linéaire sur la variable réponse. Ce modèle suppose que la dépendance peut être décrite dans un espace linéaire de dimension modérée. Ainsi, si on considère que le profil X des variables explicatives suit une loi normale telle que :

$$X \sim \mathcal{N}_m(0, \Sigma), \tag{1.3}$$

le modèle à facteurs suppose l'existence d'un vecteur de $q \ll m$ variables latentes $Z = (Z_1, Z_2, \ldots, Z_q)$, que l'on suppose distribué selon une loi normale d'espérance nulle et de variance $\mathbb{I}_q$, décrivant la dépendance entre les variables explicatives. Conditionnellement aux facteurs latents Z, les variables explicatives sont en effet indépendantes et suivent une loi normale telle que :

$$X|Z = z \sim \mathcal{N}_m(Bz, \Psi), \tag{1.4}$$

où B est une matrice $(m \times q)$ de *loadings* représentant la dépendance commune aux m variables. En effet,

$$Cov(X, Z) = B.$$

Ψ est une matrice diagonale représentant la variance spécifique aux variables explicatives.

De manière équivalent, le modèle (1.4) conduit à la décomposition suivante de la matrice de variance-covariance Σ :

$$\Sigma \quad = \quad \Psi + BB'.$$

D'un point de vue numérique, cette décomposition est intéressante car elle permet notamment l'inversion de Σ en ne faisant appel qu'à l'inversion de matrices diagonales ou de petite dimension $(q \times q)$ par la formule de Woodbury (Press et al. (2007)) :

$$\Sigma^{-1} \quad = \quad \Psi^{-1} - \Psi^{-1}B(\mathbb{I}_q + B'\Psi^{-1}B)^{-1}B'\Psi^{-1}.$$

Plusieurs méthodes de tests ou de sélection de variables s'appuient sur l'identification du noyau de dépendance BZ pour décorréler les variables explicatives : on dénote entre autres SVA pour *surrogate variable analysis* (Leek and Storey (2007)), FAMT pour *factor analysis for multiple testing* (Friguet et al. (2009)), LEAPP pour *latent effect adjustment after primary projection* (Sun et al. (2012)). La principale différence entre ces méthodes réside sur la technique de séparation du signal et du bruit, structuré par l'effet des facteurs latents sur la variable réponse dans la procédure d'estimation des paramètres du modèle (1.4). Toutes les méthodes supposent que le signal est parcimonieux et l'estimation de la structure de dépendance repose sur l'identification des variables hors du support du signal, assimilables à du bruit pur.

Plusieurs techniques existent pour identifier ces variables, à partir desquelles on peut identifier la structure de dépendance du bruit. Dans la procédure FAMT (*factor analysis for multiple testing*), proposée par Friguet et al. (2009), les variables explicatives sans effet sur la variable réponse sont identifiées par seuillage sur les statistiques de test. Causeur et al. (2012) propose une adaptation de FAMT pour l'analyse de potentiels évoqués cognitifs. La méthode FAMT est implémentée dans le package `R FAMT` (Causeur et al. (2011)). La méthode SVA (*surrogate variable analysis* , Leek and Storey (2007), Leek and Storey (2008)) estime les coefficients associés à la covariable sans ajustement sur la dépendance puis isole itérativement les facteurs latents en pondérant par un poids faible les variables pour lesquelles l'effet est non nul. Cette méthode est implémentée dans le package `R sva` (Leek et al. (2014)). Enfin, dans la procédure d'estimation LEAPP (*latent effect adjustment after primary projection*), Sun et al. (2012) introduisent une matrice de rotation transformant les données de telle manière à concentrer l'ensemble du signal sur une seule variable. Les facteurs latents sont alors estimés à partir des autres variables transformées par un modèle de régression à effets mixtes. Enfin, la structure en facteurs est intégrée dans l'estimation de l'effet de la variable transformée concentrant le signal. Cette méthode est implémentée dans le package `R leapp` (Sun et al. (2014)).

2.2.2 Estimation des paramètres

La littérature sur l'estimation des paramètres du modèle à facteurs, en particulier pour ses applications traditionnelles en psychologie, est vaste (Mardia et al. (1979)). L'estimation par maximum de vraisemblance introduite par Jöreskog (1967) est adaptée à l'hypothèse de normalité des variables explicatives introduite plus haut. Cependant, la maximisation directe de la vraisemblance n'est pas possible en grande dimension. Friguet et al. (2009) proposent un algorithme EM (Rubin and Thayer (1982)) s'appuyant sur un parallèle entre facteurs latents et données manquantes. Lorsque B et Ψ sont connus, les facteurs latents sont estimés par les scores de Thomson (Thomson (1951)) et par la formule d'inversion de Woodbury :

$$\begin{aligned} \mathbb{E}(Z|X=x) &= B'\Sigma^{-1}x \\ &= (\mathbb{I}_q + B'\Psi^{-1}B)^{-1}B'\Psi^{-1}x. \end{aligned}$$

2.2.3 Nombre de facteurs

Le choix du nombre de facteurs q à retenir dans le modèle à facteurs est un point crucial de l'analyse. Extraire un trop grand nombre de facteurs peut rendre l'estimation des variances résiduelles spécifiques Ψ artificiellement faibles et mener à des décisions erronées lors de l'étape de sélection de variables.

Par analogie avec l'analyse en composantes principales, les méthodes les plus classiques pour estimer le nombre de facteurs reposent sur l'examen de la séquence ordonnée dans l'ordre décroissant des valeurs propres de la matrice de variance-covariance. Certaines procédures simples comme la méthode de Kaiser, retiennent le nombre de valeurs propres supérieures à 1 ou à la moyenne des valeurs propres (Kaiser (1960)). Pour l'essentiel toutefois, les méthodes de détermination du nombre de facteurs cherchent à identifier une rupture dans la décroissance de la séquence des valeurs propres, aussi appelée coude. Certains auteurs ont proposé des procédures de détermination automatique de ce coude. Zoski and Jurs (1993) proposent ainsi une méthode basée sur des régressions sur des séquences emboîtées de valeurs propres successives, le nombre de facteurs étant alors déterminé par le nombre de valeurs propres dans la séquence pour laquelle la différence de pente entre deux régressions successives est maximale. De nombreuses autres méthodes privilégient une approche par tests de comparaison de modèles emboîtés (voir entre autres Anderson (1963), Bartlett (1950), Bartlett (1951), Lawley (1956)). Ces méthodes sont toutefois peu utilisées en grande dimension, principalement car elles surestiment le nombre de facteurs. Dans une revue détaillée de nombreuses méthodes, Ford et al. (1986) suggèrent que l'analyse parallèle (Buja and Eyuboglu (1992)) décrite à présent est la méthode la plus performante. Buja and Eyuboglu (1992) proposent de retenir le nombre de valeurs propres de la matrice de covariance empirique supérieures à la moyenne (ou à un quantile) des valeurs propres de matrices de mêmes dimensions obtenues par ré-échantillonnage sous l'hypothèse d'absence de structure en facteurs. Il s'agit de la méthode utilisée dans SVA par Leek and Storey (2007) et dans LEAPP par Sun et al. (2012).

Enfin, dans le contexte des tests multiples en grande dimension, Friguet et al. (2009) proposent une méthode pour estimer le nombre de facteurs en grande dimension, par minimisation du critère d'inflation de la variance du nombre de faux positifs. En situation d'indépendance, ce nombre de faux positifs suit une loi Binomiale dont on peut donner explicitement l'expression de la variance. Friguet et al. (2009) donnent une expression analytique ν_k pour l'inflation de la variance du nombre de faux positifs, lorsque la covariance entre les statistiques de tests est décrit par un modèle à k facteurs. Ils montrent ainsi que cette inflation de variance est une fonction croissante du niveau de dépendance entre les statistiques de tests. Les auteurs déterminent alors le nombre de facteurs par minimisation de la variance du nombre de faux positifs, estimée quand les tests sont calculés à partir des résidus du modèle à k facteurs : $\hat{\varepsilon} - \hat{B}_k\hat{Z}_k$ pour chaque modèle à k facteurs (B_k, Ψ_k). Dans la suite, on utilisera cette méthode pour déterminer le nombre de facteurs.

2.2.4 Décorrélation par ajustement des effets des facteurs latents

Après estimation des paramètres du modèle à facteurs et des variables latentes, l'idée principale des méthodes de sélection est de travailler sur des données décorrélées. On définit les données ajustées sur l'effet de ces facteurs latents par :

$$\tilde{X} \quad = \quad X - BZ.$$

En pratique, on approche les facteurs latents Z par $\tilde{Z} = \mathbb{E}(Z|X)$. Pour faire le lien avec les méthodes de décorrélation basées sur les innovations, on peut remarquer que :

$$X - B\tilde{Z} \quad = \quad \Psi\Sigma^{-1}X.$$

Ainsi, les méthodes statistiques initialement développées sous l'hypothèse d'indépendance ou de faible dépendance peuvent être appliquées aux données obtenues après ajustement de l'effet des facteurs latents. Friguet et al. (2009) montrent les bonnes propriétés de la méthode de Benjamini-Hochberg appliquée aux statistiques de tests calculées sur les données décorrélées. De la même manière, dans une optique de classification supervisée, Leek et al. (2014) proposent de combiner à l'étape de décorrélation une analyse diagonale discriminante dans laquelle les paramètres sont estimées par *shrinkage* (Tibshirani et al. (2002)).

3 Synthèse : fardeau ou aubaine ?

Il est intéressant de remarquer que peu d'articles proposent des résultats théoriques sur les performances de procédures de sélection de variables dans un cadre général de dépendance : la plupart des résultats supposent en effet l'indépendance entre variables explicatives (le bien connu $X_1, \ldots, X_n$ *i.i.d.*), une faible dépendance ou une corrélation très structurée, comme par exemple une structure auto-régressive (Hall and Jin (2008), Hall and Jin (2010)). Certaines méthodes ignorant la corrélation entre statistique de tests ou de sélection s'avèrent d'ailleurs robustes à la dépendance (Ahdesmäki and Strimmer (2010), Efron (2008),

Hand (2006), Efron (1975)), ce qui peut laisser penser que cette dépendance peut être négligée en pratique.

Pourtant, sur la base à la fois d'un argumentaire théorique et d'observations pratiques, il semble dangereux de généraliser à des situations arbitrairement complexes de dépendance les propriétés des procédures établies sous l'hypothèse d'indépendance. Dans les contextes de données d'ERP ou génomiques évoquées dans ce manuscrit, l'impact de la dépendance sur les propriétés des procédures standards est en effet clairement négatif : perte de stabilité des procédures de tests ou de sélection, diminution des niveaux de performance. En d'autres termes, si l'on s'intéresse à la perturbation engendrée par la dépendance sur les propriétés de méthodes construites idéalement sous l'hypothèse d'indépendance, il peut être tentant de conclure que la dépendance est un fardeau.

En revanche, l'intégration d'un modèle de la dépendance dans la construction des méthodes permet d'atteindre non seulement l'objectif de restaurer les propriétés de stabilité sous l'hypothèse d'indépendance mais aussi de réduire le bruit, de mieux identifier le signal et donc d'augmenter la puissance des procédures de tests et de sélection. C'est aussi le constat fait par Hall and Jin (2010) dans un paragraphe intitulé *correlation, curse or blessing ?* où les auteurs démontrent selon deux raisonnements que face à un problème de tests multiples, la situation est plus favorable lorsque les statistiques de test sont corrélées. S'inspirant du paradigme du signal *Rare and Weak* de Donoho and Jin (2008), Hall and Jin (2010) considèrent qu'un vecteur de statistiques de test est un vecteur gaussien que l'on peut écrire comme le modèle suivant :

$$X = \mu + Z$$

où $Z \sim \mathcal{N}_m(0, \Sigma)$. Le signal est un vecteur μ parcimonieux dont k coordonnées sont non nulles parmi m et de même amplitude A_m, les indices des coordonnées non nulles sont notés $\ell_1, \ldots, \ell_K \in \{1, \ldots, m\}$, l'ensemble de ces indices formant le support du signal. Σ est une matrice de corrélation et l'on cherche à comparer la situation où $\Sigma = \mathbb{I}_m$ à celle où $\Sigma \neq \mathbb{I}_m$, du point de vue de l'estimation du support du signal. Comme les moyennes μ sont les mêmes dans les deux problèmes, une manière de les comparer est de mesurer la quantité d'incertitude du bruit Z par l'entropie différentielle. Pour une loi de probabilité à densité ϕ, l'entropie différentielle s'écrit :

$$h(Z) \quad = \quad -\int_{\mathbb{R}^m} \phi(z) \log(\phi(z)) dz.$$

Dans le cas d'une loi normale multivariée, $h(Z)$ est proportionnelle au déterminant de la matrice de corrélation Σ (Cover and Thomas (2006)).

$$h(Z) \quad \propto \quad |\Sigma|.$$

Enfin, le déterminant d'une matrice de corrélation est maximum lorsqu'elle est égale à l'identité. En effet, si $\lambda_1, \ldots, \lambda_m$ sont les valeurs propres de Σ et $|\Sigma|$ désigne le

déterminant de Σ, par convexité de la fonction $x \mapsto \log(x)$ on a :

$$\begin{aligned} \log|\Sigma| &= \sum_{k=1}^{m} \log(\lambda_k) \\ \log|\Sigma| &\leq m \log(\frac{1}{m}\sum_{k=1}^{m} \lambda_k) = 0, \end{aligned}$$

donc

$$|\Sigma| \leq 1 = |\mathbb{I}_m|.$$

Ainsi, le cas de statistiques de test indépendantes contient plus d'incertitude que le cas corrélé et peut donc être considéré, selon ce point de vue, comme plus difficile.

Une autre façon de montrer que le cas où $\Sigma \neq \mathbb{I}_m$ est une situation favorable est de reprendre l'approche suggérée par Hall and Jin (2010) de décorrélation par la factorisation de Cholesky inverse (on rappelle que $U_m \Sigma U_m' = \mathbb{I}_m$) :

$$U_m X = U_m \mu + U_m Z$$

avec $U_m Z \sim \mathcal{N}(0, \mathbb{I}_m)$. La structure de covariance étant la même, on peut comparer l'amplitude du signal sur son support $(\ell_1, \ldots, \ell_K)$. Hall and Jin (2010) montrent que le signal dans le cas décorrélé est plus fort. En effet, si l'on s'intéresse aux coordonnées $(\ell_1, \ldots, \ell_K)$ du vecteur $(U_m \mu)$, alors :

$$\begin{aligned} (U_m\mu)_{l_k} &= A_m \sum_{j=1}^{K} U_m(l_k, l_j) \\ &= A_m U_m(l_k, l_k) + A_m \sum_{j=1}^{K-1} U_m(l_j, l_k) \end{aligned}$$

car la matrice U_m issue de la factorisation de Cholesky inverse est triangulaire inférieure. On peut maintenant remarquer que les termes diagonaux de U_m sont supérieurs à 1 car Σ est une matrice de corrélation. Ainsi,

$$U_m(l_k, l_k) \geq 1.$$

Enfin, les auteurs introduisent des hypothèses de décroissance "rapide" des termes extra-diagonaux des matrices Σ et U_m. Sous cette condition, et comme le signal est rare, les termes $U_m(l_j, l_k)$ sont proches de 0 :

$$U_m(l_j, l_k) \approx 0.$$

Finalement, les auteurs en concluent que :

$$(U_m\mu)_{l_k} \gtrsim A_m.$$

Autrement dit, le cas corrélé est une situation plus favorable, ce qui apporte une autre justification aux approches de décorrélation.

On retrouve des conclusions similaires dans Verzelen (2012), où l'auteur étudie les limites de détectabilité et d'estimabilité d'un signal dans un problème général de régression, en fonction de sa parcimonie et de son amplitude. Il établit des bornes minimax du risque et en déduit une caractérisation des situations de grande dimension dans lesquelles le signal n'est ni détectable, ni estimable. Ainsi, la "ultra haute dimension" est définie de manière formelle en fonction d'un paramètre de parcimonie k (le nombre de coordonnées non nulles du vecteur des coefficients de régression) et du nombre d'individus n :

$$k \log(\frac{m}{k}) \gg n.$$

Verzelen (2012) établit aussi des bornes minimax en situation de dépendance et en conclut que la dépendance favorise la détectabilité du signal. En effet, il démontre que l'amplitude minimale du signal nécessaire pour séparer l'hypothèse alternative de l'hypothèse nulle est supérieure lorsque les variables sont indépendantes.

Ce manuscrit apporte de nouvelles illustrations des résultats précédents. Ainsi, on donne dans le Chapitre 3 l'expression des taux d'erreurs minimaux de classification par la règle de Bayes lorsque le profil des variables explicatives présente une dépendance structurée par l'effet linéaire de facteurs latents. Dans ces mêmes conditions, on donne également dans le Chapitre 4 les bornes de détectabilité d'un signal, tel que défini par Donoho and Jin (2004) et on montre que le plus petit signal détectable est plus faible en situation de forte dépendance.

4 Organisation de la thèse

Le Chapitre 2 présente une procédure d'identification du support d'un signal pour des données de potentiels évoqués cognitifs (ERPs) issues d'une expérience visant à détecter les différences d'activité cérébrales entre différentes conditions expérimentale. Les données d'ERP sont caractérisées par une structure en blocs d'intervalles de temps de leur dépendance, combinée avec une composante autorégressive d'ordre 1. La méthode proposée exploite cette structure de covariance remarquable pour estimer conjointement le signal et les corrélations résiduelles du modèle, en prenant pour hypothèses une connaissance a priori d'intervalles de temps pour lesquels le signal est nul et l'existence d'une structure à facteurs pour la matrice de covariance résiduelle. Les résultats de ce chapitre ont fait l'objet d'un article en seconde révision (Sheu et al. (2015)).

Le Chapitre 3 est dédié à l'étude de l'impact de la dépendance sur la stabilité de la sélection de variable pour la classification supervisée. La méthode proposée s'appuie sur l'introduction de facteurs latents de dépendance dans un modèle linéaire généralisé. La méthode d'ajustement alterne une étape d'estimation de la structure de covariance et des moyennes par groupe avec une étape d'estimation des probabilités individuelles d'appartenance aux groupes, intervenant dans l'ajustement des données par les effets latents. Les résultats de ce chapitre ont fait l'objet d'un

article (Perthame et al. (2015)) et d'un package `R` intitulé `FADA` disponible sur le `CRAN` (Perthame et al. (2014)).

Le Chapitre 4 présente une adaptation de la méthode *Higher Criticism Thresholding* à une situation de dépendance entre statistique de tests. Donoho and Jin (2004) montrent que cette méthode, développée initialement pour la détection d'un signal puis pour l'estimation du support du signal dans un contexte général de comparaisons multiples, est optimale sous une hypothèse d'indépendance. Cependant, certains auteurs suggèrent que ces propriétés sont considérablement remises en cause lorsque la dépendance entre les tests est importante. Dans ce chapitre, on étudiera l'apport d'une approche HCT basée sur une décorrélation préalable pour la détection d'un signal cognitif dans le cadre de mesures d'ERPs. La modélisation par un modèle à facteurs de la dépendance offre en effet un cadre théorique favorable, dans lequel le calcul des innovations est rendu possible par le calcul explicite d'une racine de l'inverse de la matrice de covariance. Dans ce contexte, on donne aussi l'expression de nouvelles bornes de détectabilité et d'estimabilité du support du signal prenant en compte la structure de dépendance. Les résultats de ce chapitre font l'objet de la rédaction d'un article.

Enfin, le Chapitre 5 rappelle les principaux points évoqués et conclut ce manuscrit par des perspectives.

Ce travail de recherche a été diffusé sous forme d'articles et de communications orales lors de conférences et de séminaires, dont une liste est présentée dans le Chapitre 6.

CHAPITRE 2

TESTS MULTIPLES POUR DES DONNÉES DE POTENTIELS ÉVOQUÉS COGNITIFS

RÉSUMÉ : les potentiels évoqués sont des mesures temporelles de l'activité électrique cérébrale et permettent des associations entre des occurrences moteurs ou cognitives et des mécanismes cérébraux. Le signal sous-jacent aux ERPs est souvent rare et faible au regard de la grande variabilité inter-individuelle. Ainsi, l'identification d'associations significatives entre les mesures d'ERP et des variables comportementales (ou expérimentales) d'intérêt représente un véritable problème statistique. Une approche de prise en compte de la dépendance par un modèle à facteurs est justifiée par la structure de dépendance observée sur les données ERP, caractérisée par des blocs de corrélations et de fortes auto-corrélations. Une procédure adaptative d'ajustement sur l'effet de facteurs latents est proposée. Cette procédure est fondée sur une estimation jointe du signal et du bruit sous-jacent aux données, à partir d'une connaissance préalable d'intervalles de temps durant lesquels du bruit seul est observé. Une étude par simulations est réalisée à partir d'un signal connu imposé intégré à la structure de dépendance extraite des données réelles. La procédure proposée atteint de bonnes performances relativement aux autres méthodes comparées et apparait plus puissante pour détecter des intervalles de temps pertinents.

Sommaire

Le contenu de ce chapitre correspond à un article intitulé "Accounting for the dependence in large-scale multiple testing of event-related potential data" actuellement en seconde révision dans la revue *Annals of Applied Statistics* (Sheu et al. (2015)).

1 Introduction

Les études cliniques et la recherche médicale font de plus en plus appel aux données à haut-débit telles que les potentiels évoqués-cognitifs (ERPs, Handy (2004)) et l'imagerie par résonance magnétique fonctionnelle (fMRI, Poldrack et al. (2011)). En effet, les ERPs permettent d'étudier très précisément l'évolution temporelle d'un processus cérébral et l'imagerie par fMRI permet de localiser des zones du cerveau activées par des conditions expérimentales. Ces données étant recueillies massivement, les chercheurs font face à des problèmes de correction des tests multiples : la recherche d'effets significatifs parmi des milliers de comparaisons demande un équilibre entre le contrôle d'un faible taux de faux positifs tout en maintenant une puissance de détection suffisante. Le but de ce chapitre est d'atteindre cet objectif sur des données d'ERPs présentant une forte et complexe structure de dépendance temporelle.

Un état de l'art sur l'analyse univariée de données ERPs est réalisé dans Groppe et al. (2011a) et Groppe et al. (2011b). Les auteurs se concentrent sur une comparaison entre des méthodes contrôlant le taux de faux positifs (FDR) (Benjamini and Hochberg (1995)) et des méthodes fondées sur du ré-échantillonnage (Blair and Karniski (1993)). Cependant, ces articles ne font pas mention de la dépendance entre les tests, héritée de la forte dépendance temporelle des données ERPs. En effet, une forte corrélation entre les variables affecte la précision de l'estimation du FDR et la stabilité des résultats (c'est-à-dire, la variance de la proportion de découvertes) (voir Efron (2007)). Ainsi, négliger la dépendance entre tests réduit la capacité de détection des procédures (Leek and Storey (2008)).

La remarquable structure de dépendance temporelle des ERPs entraine une forte régularité temporelle des statistiques de test. On observe alors des intervalles de temps, en dehors du support du signal, pour lesquels les probabilités critiques sont très faibles. Différentes approches spécifiques aux ERPs existent pour résoudre ce problème de dépendance entre statistiques de test. Le test de Guthrie-Buchwald (Guthrie and Buchwald (1991)), proposé avant l'essor des méthodes contrôlant le FDR, considère qu'un intervalle de temps est significatif lorsque les probabilités critiques sont inférieures à un certain seuil (0.05 par exemple) et s'il est suffisamment long. La durée de l'intervalle de temps est alors comparée à la distribution des durées de processus auto-régressifs simulés sous l'hypothèse nulle. Cependant, cette procédure n'est pas construite pour contrôler le taux de faux positifs. Dans le même contexte, Sun and Cai (2009) propose une méthode utilisant un modèle à chaine de Markov cachée pour prendre en compte la corrélation. Ce modèle suppose l'existence de classes latentes, structurant les données. Enfin, les méthodes fondées sur un modèle à facteurs latents (SVA Leek and Storey (2007), LEAPP Sun et al. (2012), FAMT Friguet et al. (2009)) présentées dans le Chapitre 1, Section 2.2 permettent aussi de prendre en compte cette dépendance multivariée.

En particulier, une adaptation de la procédure FAMT (Friguet et al. (2009), Causeur et al. (2011)) au contexte des ERPs est proposée par Causeur et al. (2012). Sur des simulations d'ERPs, cette méthode donne de meilleurs résultats de détection du signal que des procédures standards telles que la correction de Benjamini-Hochberg (Benjamini and Hochberg (1995)). Excepté le test de Guthrie-Buchwald, toutes ces méthodes ne proposent pas de tirer profit de la composante temporelle de la dépendance pour séparer le signal du bruit. Le but de ce chapitre est d'illustrer comment la dépendance induit une régularité dans l'estimation du signal, qui mène à une mauvaise identification du support du signal, entrainant ainsi une confusion entre l'effet des covariables et les facteurs latents. Ensuite, un algorithme alternant estimation des paramètres du modèle à facteurs et effet des covariables sachant la structure de covariance et une connaissance a priori d'intervalles de temps sous l'hypothèse nulle est proposé.

La Section 2 de ce chapitre décrit l'expérience d'oubli direct mise en place par les psychologues puis le modèle utilisé pour l'analyse de données ERPs est présenté. Durant cette expérience, les ERPs correspondent à l'activité cérébrale mesurée au cours du temps. La variable comportementale d'intérêt est une mesure de la mémoire de reconnaissance. On s'intéresse ici à la structure de dépendance des statistiques de

tests qui permettent d'étudier l'association entre les ERPs et ce score de reconnaissance. La Section 3 illustre l'impact de la dépendance sur des procédures classiques de comparaisons multiples. La procédure adaptative d'ajustement sur les facteurs latents proposée est ensuite détaillée dans la Section 4. Cette procédure capture de manière itérative la corrélation résiduelle des données par un modèle à facteurs et corrige l'estimation du signal de la régularité induite par la corrélation de l'activité cérébrale. La Section 5 présente les résultats de simulations visant à comparer la méthode proposée à des procédures standards de correction des probabilités critiques et à d'autres procédures reposant sur un modèle à facteurs. Enfin, la Section 6 présente les résultats de l'analyse des données réelles d'ERPs recueillies lors de l'expérience d'oubli direct. On remarquera que les procédures standards contrôlant le FDR ne détectent aucun intervalle de temps durant lequel l'activité cérébrale est significativement liée au score de performance. Inversement, la méthode proposée déclare significative une vague de corrélations autour de 400 ms, ce qui peut être expliqué par la composante FN400, déjà mentionnée dans la littérature sur la mémoire de reconnaissance (voir Rugg and Curran (2007)).

2 Modèle pour l'analyse de données ERPs

2.1 Expérience d'oubli direct

Le développement de la méthode proposée a été motivé par une étude de données ERPs dont le but est d'explorer les processus électro-physiologiques associés au phénomène d'oubli direct. L'étude fait appel à un paradigme expérimental similaire à celui décrit par Lee et al. (2013) pour étudier la capacité d'oubli ou de mémorisation intentionnels d'objets présentés à des sujets. L'expérience est composée de deux phases. Dans la phase d'apprentissage, on demande à 20 participants de mémoriser (signalé par un "+") ou d'oublier (signalé par un "X") un mot brièvement projeté sur un écran. Les ERPs sont mesurés pendant 1000 millisecondes (ms) pour chacun des 90 mots, 45 mots à oublier (TBF, *to-be-forgotten*) et 45 à retenir (TBR, *to-be-remembered*). Ensuite, on teste la capacité des sujets à reconnaître si un mot leur a déjà été présenté (ancien ou nouveau mot). Durant la phase de test, 90 nouveaux mots sont ajoutés aux 90 anciens. Pour mesurer la capacité de reconnaissance, on calcule la proportion de bonnes réponses (le sujet reconnaît qu'un ancien mot lui a déjà été présenté) moins la proportion de fausses alertes (le sujet croit à tort qu'un nouveau lui a déjà été présenté). Les données ERPs mesurées une fois par milli-seconde sur neuf électrodes (3 sur la région frontale, 3 sur la région centrale et 3 sur la région postérieure) sont pré-traitées. Pour chaque participant et chaque condition, les signaux sont moyennés sur l'ensemble des mots. La Figure 2.1 présente les vingt courbes (une par sujet) de la condition TBR de l'électrode CZ (milieu de la région centrale).

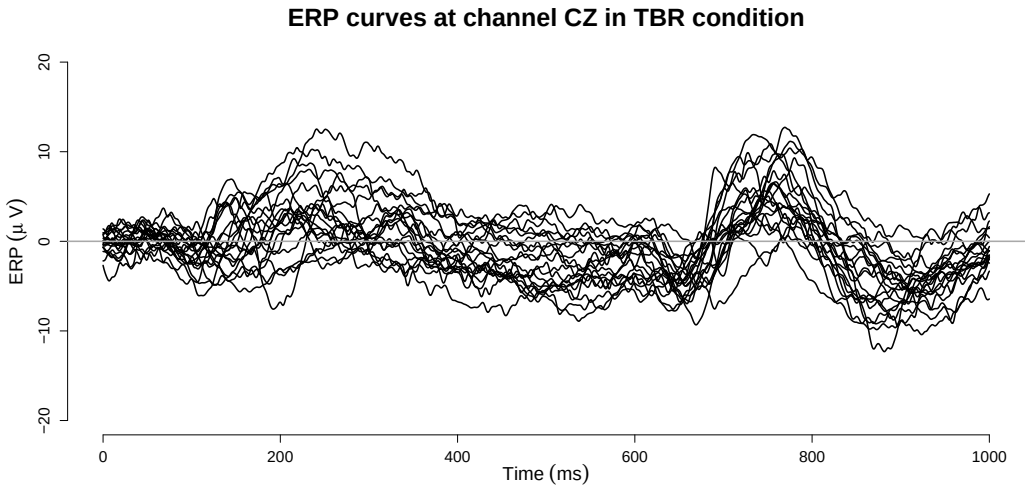


FIGURE 2.1 – Courbes d'ERPs de la condition TBR pour l'électrode CZ (milieu de la région centrale) lors de l'expérience d'oubli direct. Un point correspond à 1 ms.

2.2 Modèle et statistique de tests

2.2.1 Modèle générale

On suppose que l'activité cérébrale diffère si on a demandé au sujet d'oublier ou de mémoriser les mots et que cette différence se traduit par des intervalles de temps différents durant lesquels l'activité cérébrale est significativement corrélée au score de performance. Pour chaque électrode placée sur la tête du sujet, cette recherche d'intervalles peut se traduire par un problème de comparaisons multiples analysant la relation entre le score de reconnaissance x et les ERPs sous les deux instructions. Ainsi, pour la mesure Y_{jkt} de l'individu j, j variant de 1 à n, à l'instant t pour la condition k ($k = 1$ pour TBR et $k = 2$ pour TBF), on a le modèle :

$$Y_{jkt} \quad = \quad \alpha_{0,t} + \beta_{jt} + \gamma_{kt} + \alpha_{kt} x_{jk} + \varepsilon_{jkt}, \tag{2.1}$$

où l'effet individuel d'un sujet est $\beta_t = (\beta_{1t}, \ldots, \beta_{nt})$ avec la contrainte d'unicité $\sum_j \beta_{jt} = 0$, l'effet de la condition expérimentale est $\gamma_t = (\gamma_{1t}, \gamma_{2t})$ avec la contrainte d'unicité $\gamma_{1t} + \gamma_{2t} = 0$ et α_{kt} est l'effet du score de reconnaissance au temps t pour la condition k avec la contrainte $\alpha_{1t} + \alpha_{2t} = 0$ pour tout t variant de 1 à T. Le bruit ε_{jkt} est supposé gaussien et l'on suppose dans un premier temps que $\mathrm{Cov}(\varepsilon_{jkt}, \varepsilon_{jkt'}) = 0$. On verra par la suite qu'on peut relâcher cette hypothèse.

2.2.2 Allure de la statistique de test

A chaque instant t pour la condition k, le problème revient à tester l'hypothèse nulle H_0^{kt}, sous laquelle le score de reconnaissance n'a pas de lien avec la mesure ERPs à l'instant t, contre l'alternative H_1^{kt} :

$$\begin{aligned} H_0^{kt} \quad &: \quad \alpha_{kt} = 0 \\ H_1^{kt} \quad &: \quad \alpha_{kt} \neq 0, \end{aligned}$$

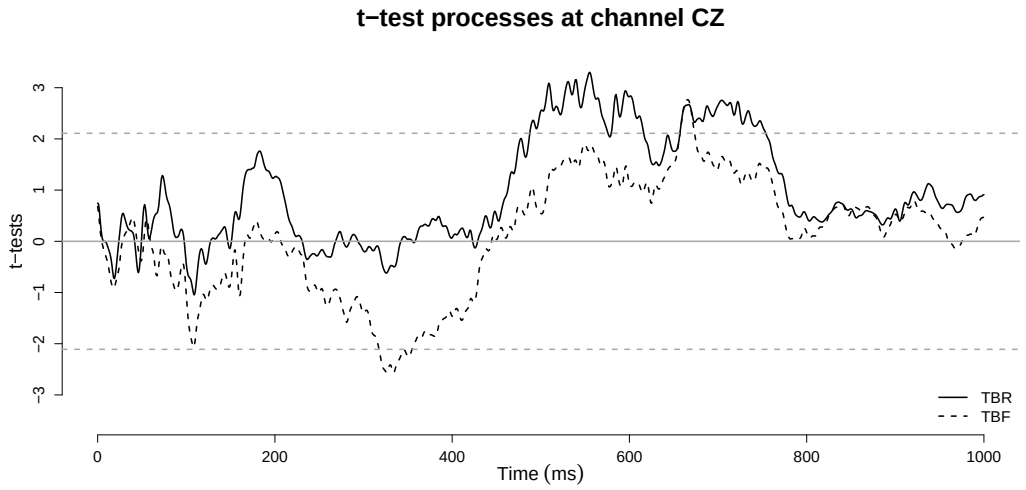


FIGURE 2.2 – Statistique de tests de significativité de α_t pour l'électrode CZ pour les conditions TBR (trait plein) et TBF (pointillés). Les lignes horizontales correspondent aux quantiles à 2.5 et 97.5 % de la distribution sous l'hypothèse nulle.

pour $k = 1, 2$ et $t = 1, \ldots, T$. La t-statistique associée à ce test pour l'électrode CZ est présentée sur la Figure 2.2. On remarque peu d'intervalles de temps significatifs, surtout pour la condition TBF. De plus, la régularité des courbes est incompatible avec celle d'une collection de statistiques de tests indépendantes suivant une loi de Student. Or, il est connu qu'une forte dépendance entre les tests modifie la distribution des statistiques de tests sous l'hypothèse nulle (Friguet and Causeur (2011)).

Cette forte dépendance temporelle est aussi confirmée par la Figure 1.1 interprétée dans le Chapitre 1 page 18, présentant l'histogramme des corrélations résiduelles ainsi qu'une image de la matrice de corrélation pour l'électrode CZ.

2.2.3 Modèle pour l'analyse de l'expérience d'oubli direct

Pour analyser les données ERPs, il est nécessaire de prendre en compte la dépendance dans le modèle (2.1). Si Y_{it} désigne la mesure de l'activité cérébrale pour le sujet $i = 1, \ldots, n$ à l'instant $t = 1, \ldots, T$ où T est le nombre de mesures. Une expérience de 1 000 ms avec une mesure d'ERP toutes les 10 ms mène par exemple à $T = 100$ mesures. On pose un modèle linéaire multivarié pour expliquer le lien entre les mesures d'ERPs et les covariables $x_i = (x_{i1}, \ldots, x_{ip})$ ajusté éventuellement sur l'effet de covariables d'ajustement $u_i = (u_{i1}, \ldots, u_{ir})$:

$$Y_{it} \quad = \quad \alpha_{0,t} + \alpha_t' x_i + a_t' u_i + \varepsilon_{it}, \tag{2.2}$$

où $\alpha_{0,t}$ est la constante du modèle, pour chaque instant t, α_t et a_t, des vecteurs de tailles p et r, sont les coefficients de régression associés à l'effet sur les mesures ERPs des covariables x et u respectivement et ε_{it} est un bruit gaussien centré d'écart-type σ_t. L'indépendance temporelle entre les termes d'erreur ε_{it} est généralement

supposée. Pour chaque sujet i, le vecteur $\varepsilon_i = (\varepsilon_{i1}, \ldots, \varepsilon_{iT})$ suit une loi normale multivariée centrée de variance :

$$D_\sigma = \text{diag}(\sigma_1^2, \ldots, \sigma_T^2),$$

où diag(.) est l'opérateur transformant un vecteur en une matrice dont les termes diagonaux sont les termes du vecteur. Afin de prendre en compte la dépendance des données d'ERPs, on relâche l'hypothèse d'indépendance des résidus dans le modèle (2.2) et on suppose que :

$$\begin{aligned} \text{Var}(\varepsilon) &= \Sigma \\ &= D_\sigma^{1/2} R D_\sigma^{1/2}, \end{aligned}$$

où R est la matrice $T \times T$ des corrélations résiduelles.

Deux types de covariables sont introduites dans le modèle (2.2) : x, dont les effets sur les mesures ERPs nous intéressent, et u, des covariables auxiliaires. Dans l'expérience d'oubli direct, u contient l'effet du sujet et l'effet principal de la condition (TBR et TBF). La seule variable d'intérêt x est le score de reconnaissance. Le cas $p = 1$ recouvre un grand nombre de situations dans lesquelles x peut être numérique (comme le score de reconnaissance) ou une variable catégorielle à deux modalités. Ces conditions expérimentales sont les plus courantes dans les expériences utilisant les ERPs (Handy (2004)) (on peut avoir $p > 1$ lorsque la covariable d'intérêt est une variable de groupes avec un nombre de groupes $K > 2$). Dans la suite, le terme *signal* désigne la matrice $\alpha(T \times p)$ dont les lignes sont les p-vecteurs α_t.

Sur les données ERPs, le signal est souvent à la fois rare et faible. Un signal rare signifie que pour la plupart des instants t, l'hypothèse nulle $H_{0t} : \alpha_t = 0$ est vraie (c'est-à-dire qu'il n'y a pas de signal sur la plupart de la durée de l'expérience). Un signal faible signifie qu'il est difficile de détecter les instants pour lesquels H_{0t} est fausse, étant donné le faible nombre d'observations et la grande variabilité résiduelle. Dans le cadre du modèle linéaire, la sélection d'instants significatifs est fondée sur une série de tests d'hypothèse sur les termes de la matrice $\alpha(T \times p)$, dont la t-ème ligne notée α_t représente les effets du vecteur de covariables x sur l'activité cérébrale au temps t. On note $\hat{\alpha}(T \times p)$ la matrice du signal observé, dont les lignes $\hat{\alpha}_t$ sont calculées par la méthode des moindres carrés appliquée au modèle (2.2) :

$$\hat{\alpha}_t = (x' P_u x)^{-1} x' P_u Y_t,$$

où $P_u = \mathbb{I}_n - U(U'U)^{-1}U'$ avec U est la matrice $n \times (r+1)$ dont la ligne i est $(1, u_i)$, $Y_t = (Y_{1t}, \ldots, Y_{nt})$ et x est la matrice $n \times p$ dont la ligne i est x_i. Le vecteur $\mathcal{T} = (\mathcal{T}_t)_{t=1,\ldots,T}$ des statistiques de test associées aux hypothèses de test H_{0t} est donné par la statistique de Fisher suivante :

$$\mathcal{T}_t = \frac{1}{p} \frac{\hat{\alpha}_t' x' P_u x \hat{\alpha}_t}{\hat{\sigma}_t^2}, \tag{2.3}$$

où $\hat{\sigma}_t^2$ est l'estimateur standard de la variance résiduelle du modèle (2.2).

Sous l'hypothèse nulle, chaque composante $\mathcal{T}_t$ de $\mathcal{T}$ suit une loi de Fisher à p et $d = n - p - r - 1$ degrés de liberté. Ici, $p = 1$, ce qui explique l'utilisation des statistiques de Student : celles-ci sont en fait la racine carrée signée des statistiques de test $\mathcal{T}_t$. Dans la suite, on désigne par p_t la probabilité critique associée à la statistique $\mathcal{T}_t$.

Sous les hypothèses du modèle linéaire, on peut confirmer que les statistiques de test héritent directement de la structure de dépendance des données par la matrice de corrélation R : sous l'hypothèse nulle globale $H_0 = \bigcap_t H_{0t}$, les composantes de $\mathcal{T}_t$ suivent une loi de Fisher dont la structure de corrélation est :

$$Cor(\mathcal{T}_t, \mathcal{T}_{t'}) \quad = \quad r_{tt'}^2 \left(\frac{1}{p} + \frac{1}{d}\right) \frac{p(d-4)}{p+d-2} \underset{d \to +\infty}{\sim} r_{tt'}^2$$

où $r_{tt'}$ est le terme (t, t') de la matrice R. La section suivante illustre l'impact de la dépendance sur des procédures classiques de tests multiples puis la méthode permettant de prendre en compte la dépendance au sein du test des coefficients du modèle (2.2) est présentée.

3 Dépendance temporelle entre statistiques de test

Ce paragraphe a pour but de rappeler quelques outils de la théorie des comparaisons multiples nécessaires dans la suite de ce chapitre.

3.1 Procédures standards de correction des probabilités critiques

La plupart des méthodes de correction des probabilités critiques consistent à rejeter l'hypothèse nulle $H_{0,t}$ si $p_t \leq p^*$ où p^* est un seuil choisi de telle façon que le nombre V_s de rejets à tort est contrôlé. Les méthodes les plus connues sont construites pour un nombre modéré de tests, comme pour la comparaison post-hoc en analyse de la variance. Leur but est de contrôler le FWER (family wise error rate) défini comme la probabilité d'obtenir au moins un faux positif :

$$\text{FWER}_s \quad = \quad \mathbb{P}(V_s \geq 1).$$

Les méthodes contrôlant le FWER garantissent que $\text{FWER}_s \leq \alpha$ où α est fixé à l'avance. Cependant les méthodes contrôlant le FWER telles que la correction de Bonferroni (voir Bonferroni (1936)) sont souvent trop conservatives quand le nombre de tests est grand. Depuis les 20 dernières années, les questions soulevées par les comparaisons multiples ont généré le développement d'un grand nombre de procédures de tests et de méthodes de seuillage pour les données en grande dimension (voir van der Laan and Dudoit (2007), Efron (2007) pour une revue des méthodes les plus utilisées, Groppe et al. (2011a), Groppe et al. (2011b), Lage-Castellanos et al. (2010) pour les données d'ERPs en psychologie. Parmi elles, on compte une nouvelle famille de méthodes visant à contrôler, non plus le FWER

mais le FDR (False Discovery Rate), défini comme l'espérance de la proportion de rejets à tort de l'hypothèse nulle parmi les rejets :

$$\mathrm{FDR}_s \quad = \quad \mathbb{E}(\mathrm{FDP}_s),$$

où la proportion de rejets à tort FDP (False Discovery Proportion) vaut :

$$\mathrm{FDP}_s \quad = \quad \begin{cases} 0, & \text{si } R_s = 0 \\ \frac{V_s}{R_s}, & \text{si } R_s > 0, \end{cases}$$

où R_s est le nombre de rejets. Une de ces méthodes est la correction très connue de Benjamini-Hochberg (BH) (voirBenjamini and Hochberg (1995)), pour laquelle les probabilités critiques corrigées s'expriment :

$$p_t^{BH} \quad = \quad p_t \frac{T}{\mathrm{rang}(p_t)},$$

où $\mathrm{rang}(p_t)$ est l'indice associé au rang de p_t. On s'intéresse ici aux méthodes contrôlant le FDR par la méthode de BH pour des tests corrélés. La méthode la plus connue est la méthode de Benjamini-Yekutieli (BY) (voir Benjamini and Yekutieli (2001)), qui modifie la procédure de BH afin de contrôler le FDR sous des hypothèses spécifiques de dépendance positive entre les tests. Les probabilités critiques corrigées sont définies dans ce cas par :

$$p_t^{BY} \quad = \quad p_t \frac{T}{\mathrm{rang}(p_t)} \sum_{t=1}^{T} \frac{1}{t}.$$

Dans le contexte des ERPs, la méthode de Guthrie-Buchwald (GB) (voir Guthrie and Buchwald (1991)) est la première à prendre en compte la dépendance temporelle en supposant que les statistiques de test suivent un processus auto-régressif d'ordre 1. Le test de Guthrie-Buchwald, proposé avant l'essor des méthodes contrôlant le FDR, considère qu'un intervalle de temps est significatif si les probabilités critiques sont inférieures à un certain seuil (0.05 par exemple) et s'il est suffisamment long. La durée d'un intervalle est comparée à la distribution des durées de processus auto-régressifs simulés sous l'hypothèse nulle. Cependant, cette procédure n'est pas construite pour contrôler le taux de faux positifs. Le but de la méthode est d'empêcher la détection d'intervalles de temps trop courts plutôt que de contrôler l'erreur de type I.

Dans la section suivante, on étudie les propriétés des corrections de Benjamini-Hochberg, Benjamini-Yekutieli et de la méthode de Guthrie Buchwald en situation de forte dépendance sur des simulations.

3.2 Impact de la dépendance sur les procédures de tests multiples

Afin d'illustrer l'impact de la dépendance temporelle sur la capacité des procédures à identifier un signal connu, on réalise une étude par simulations. On

génère $n \times T$ données d'ERPs selon le modèle (2.2), avec $n = 20$ sujets et une durée d'expérience de $T = 1\ 000$ ms, comme dans l'expérience d'oubli direct présentée dans la Section 2.1. Pour chaque jeu de données simulé, la seule covariable x est le score de reconnaissance centré, comme celui observé dans la condition TBR de l'expérience. On pose $\alpha_{0,t} = 0$, $\alpha_t = 0$ pour tout $t = 1, \ldots, T$ et les écart-types résiduels $\sigma_t, t = 1, \ldots, T$ correspondent aux estimations calculées sur les courbes d'ERPs observées à l'électrode CZ.

On considère deux matrices de corrélations résiduelles : une structure d'indépendance où la matrice de corrélation est l'identité $\mathbb{I}_T$ et la structure de dépendance observée des données ERPs présentée en Figure 1.1. Le vrai signal $t \mapsto \alpha_t$ correspond à un pic de support [450 ms; 550 ms] et dont l'amplitude varie en commençant par 0 puis de 1.5 à 12.5 avec un pas de 0.1. La Figure 2.3 montre la puissance associée aux tests de Student $H_{0t} : \alpha_t = 0$ pour un risque fixé à 5%. On simule 1 000 jeux de données pour chaque combinaison de force de signal et de structure de corrélation. Dans ces simulations, on compare les procédures de GB avec un seuil graphique de 0.05 et deux procédures contrôlant le FDR : BH et BY avec un contrôle du FDR à 5%.

A chaque étape de la simulation, les performances des procédures sont mesurées par le FDR, la PPV (Positive Predictive Value ou précision, qui est la proportion de rejets de l'hypothèse nulle corrects parmi les rejets) et la probabilité de non rejet (PNR, définie comme la proportion de jeux de données pour lesquels il n'y a aucun rejet à tort de l'hypothèse nulle). Les Figures 2.4, 2.5 et 2.6 résument les résultats pour le FDR, la PPV et la PNR.

L'application de ce traitement du signal étant biomédicale, on peut interpréter les procédures de tests multiples en terme de sensibilité et résolution, qui permettent de comparer la précision d'appareils de mesure. La sensibilité est définie ici comme la plus petite amplitude de signal pour laquelle la méthode commence à détecter de faibles impulsions (fausses ou justes). Dans la situation présente de comparaisons multiples, on propose de définir la sensibilité d'une procédure comme le pic minimum pour lequel $1 - \text{PNR}$ dépasse 0.1 :

$$\text{Sensibilité} \quad = \quad \min\{\max_t \alpha_t, 1 - \text{PNR} \geq 0.10\}. \tag{2.4}$$

De même, la résolution d'une procédure de tests multiples est l'amplitude minimale pour laquelle le support du signal est détecté. On propose de définir la résolution d'une méthode comme l'amplitude minimum pour laquelle la précision dépasse 0.9, c'est-à-dire que 90% du support est découvert :

$$\text{Résolution} \quad = \quad \min\{\max_t \alpha_t, \text{PPV} \geq 0.90\}. \tag{2.5}$$

La sensibilité et la résolution des trois procédures comparées dans cette étude par simulations sont rapportées dans la Table 2.1.

On remarque sans surprise que les procédures de BH et BY contrôlent le FDR. En effet, les deux méthodes contrôlent le FDR à un niveau inférieur de celui fixé (fixé

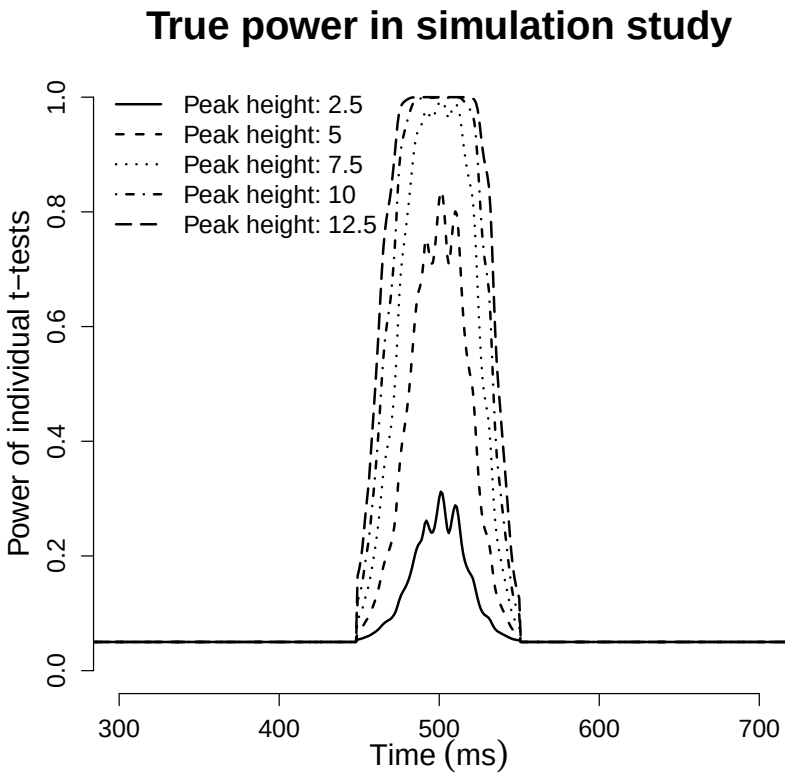


FIGURE 2.3 – Puissances théoriques associées à chaque test de l'hypothèse $H_{0,t}$: $\alpha_t = 0$ dans l'étude par simulations. La forme du signal est la même sur l'intervalle [450 ms, 550 ms] mais l'amplitude du pic varie de 0 à 12.5 : seules les courbes de puissance associées aux amplitudes 2.5, 5.0, 7.5, 10.0 et 12.5 sont représentées ici.

TABLE 2.1 – Sensibilité et résolution (rappel des définitions : (2.4) et (2.5)) pour 3 procédures de tests multiples (BH, BY and GB) calculées à partir de simulations de données d'ERP sous l'indépendance et sous dépendance. Pour les méthodes BH et BY, le FDR est contrôlé au niveau 5%. Le seuil graphique de la méthode de GB est fixé à 0.05.

	Indépendance		Dépendance	
Méthode	Sensibilité	Résolution	Sensibilité	Résolution
BH	2.6	6.3	3.6	9.1
BY	4.3	7.5	4.9	10.4
GB	3.7	5.9	0.0	7.9

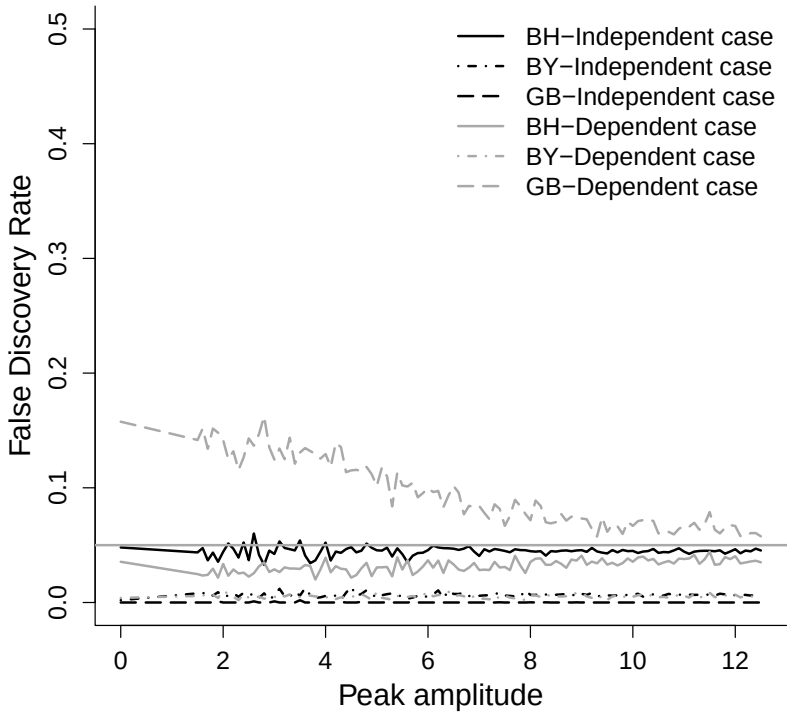


FIGURE 2.4 – Comparaison du taux de faux positifs (FDR) pour 3 procédures de tests multiples (trait plein pour BH, pointillés courts pour BY et pointillés longs GB) à partir de simulations d'ERPs sous l'indépendance (noir) et en situation de dépendance (gris). 1 000 jeux de données sont simulés pour chaque amplitude de pic $\max_t \alpha_t$.

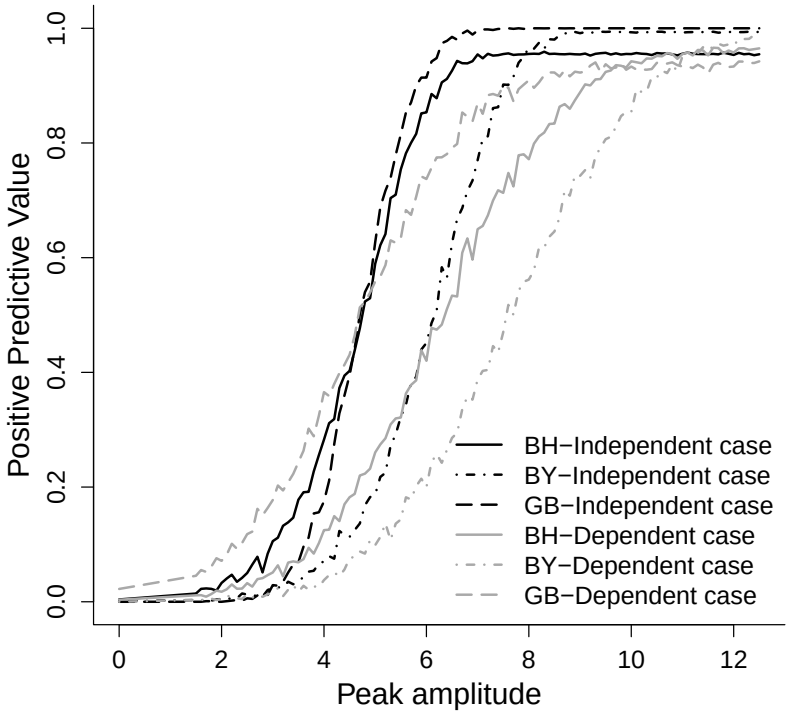


FIGURE 2.5 – Comparaison de la précision (PPV) pour 3 procédures de tests multiples (trait plein pour BH, pointillés courts pour BY et pointillés longs GB) à partir de simulations d'ERPs sous l'indépendance (noir) et en situation de dépendance (gris). 1 000 jeux de données sont simulés pour chaque amplitude de pic $\max_t \alpha_t$.

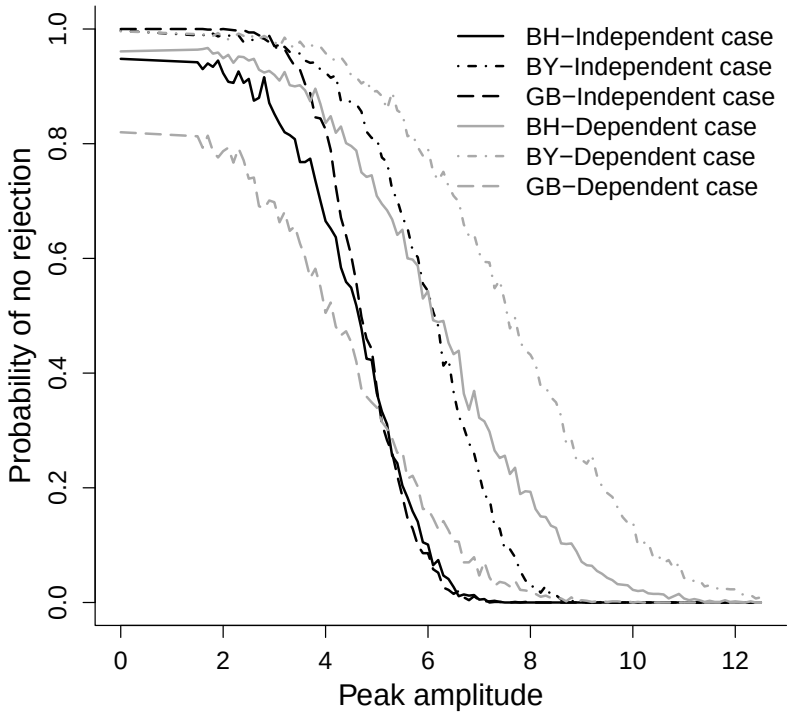


FIGURE 2.6 – Comparaison de la probabilité de non rejet (PNR) pour 3 procédures de tests multiples (trait plein pour BH, pointillés courts pour BY et pointillés longs GB) à partir de simulations d'ERPs sous l'indépendance (noir) et en situation de dépendance (gris). 1 000 jeux de données sont simulés pour chaque amplitude de pic $\max_t \alpha_t$.

à 0.05) pour le cas de dépendance, surtout BY. Ceci induit une remarquable perte de précision par rapport à BH. Ces résultats sont cohérents avec ceux rapportés par des études précédentes sur les méthodes de comparaisons multiples (voir Friguet et al. (2009)). La Table 2.1 confirme que la dépendance a tendance à réduire la sensibilité des procédures de BH et BY.

En revanche, la méthode de GB ne permet pas d'assurer le niveau de FDR fixé en cas de dépendance, surtout lorsque le signal est faible ou modéré, alors que la méthode semble sur-contrôler le FDR en cas d'indépendance. On peut aussi déduire de la Table 2.1 que la sensibilité de la méthode de GB augmente en présence de dépendance. En effet, en situation de dépendance, même quand le signal est nul durant toute l'expérience, la probabilité que GB détecte des instants significatifs dépasse 0.10. Ces observations sont aussi cohérentes avec celles rapportées par Causeur et al. (2012) pour lesquelles les données sont générées selon un modèle pour les ERPs (voir Yeung et al. (2004)) mêlé à un processus auto-régressif d'ordre 1.

Finalement, les trois procédures sont moins précises en présence de dépendance qu'en situation d'indépendance. La Table 2.1 nous permet de dire que la dépendance affecte la résolution des procédures de BH, BY et GB. Cependant, la méthode de GB présente une meilleure capacité à détecter le pic, en terme de précision, en cas de dépendance, et, si l'amplitude du signal n'est pas trop faible, aussi en cas d'indépendance.

De plus, la dépendance semble affecter la stabilité de la détection du signal des méthodes de BH et BY. En effet, la probabilité de non rejet diminue beaucoup plus lentement avec l'amplitude du signal lorsque les tests sont fortement corrélés. On rappelle que le FDR et le PNR sont liés par la formule

$$\mathrm{FDR} \quad = \quad \mathrm{pFDR}(1 - \mathrm{PNR})$$

où

$$\mathrm{pFDR} \quad = \quad \mathbb{E}(\mathrm{FDR}|R > 0)$$

est appelé *positive FDR* (voir Storey and Tibshirani (2003)). Bien que la dépendance ne semble pas affecter le contrôle global du niveau du FDR, elle a tendance à augmenter la PNR et le pFDR. En d'autres termes, lorsque les tests sont fortement dépendants, les chances de déclarer significatif au moins un instant est plus faible que dans le cas d'indépendance. De plus, quand certains instants sont déclarés significatifs, la proportion de faux positifs est plus grande.

3.3 Modèle à facteurs

On propose un modèle à facteurs latents pour les résidus du modèle (2.2) afin de prendre en compte la structure de dépendance temporelle complexe des données ERPs. Ensuite, on propose d'estimer le signal et la dépendance pour obtenir des statistiques de tests plus précises et affranchies le plus possible de la dépendance.

On suppose donc l'existence de q facteurs latents $f = (f_1, \ldots, f_q)$ distribués selon une loi normale centrée de variance $\mathbb{I}_q$ tels que, conditionnellement à z_i, x_i et f_i, la mesure d'ERPs Y_{it} pour le sujet i au temps t s'écrit :

$$Y_{it} = \alpha_{0,t} + \alpha_t' x_i + a_t' u_i + b_t' z_i + e_{it}, \tag{2.6}$$

où b_t est le q-vecteur des coefficients des facteurs pour Y_t et e_{it} est un terme d'erreur spécifique, distribué selon une loi normale centrée et de variance ψ_t^2. De plus, on suppose que les termes d'erreur sont indépendants deux à deux, ce qui permet d'écrire la décomposition suivante pour la matrice de covariance résiduelle Σ :

$$\Sigma = \Psi + BB',$$

où B est une matrice $T \times q$ dont les lignes sont les b_t et Ψ est une matrice diagonale dont les éléments diagonaux sont les ψ_t^2.

Afin d'illustrer la capacité du modèle (2.6) à capter la structure de dépendance complexe observée sur la Figure 1.1, on estime la matrice de ses corrélations résiduelles par un modèle à 1, 5 et 10 facteurs pour l'électrode CZ avec l'algorithme EM proposé par Friguet et al. (2009). Les résultats sont comparés sur la Figure 2.7. On constate que la structure de dépendance est globalement bien approchée avec un faible nombre de facteurs par rapport à la dimension de la matrice.

4 Estimation jointe du signal et de la structure de dépendance

4.1 Algorithme

La méthode proposée emploie un procédé itératif pour actualiser l'une après l'autre l'estimation du signal et des paramètres du modèle à facteurs. A chaque étape, sachant la connaissance d'intervalles de temps $\mathcal{T}_0$ où le signal est nul, l'estimation du bruit en dehors de $\mathcal{T}_0$ est mise à jour en utilisant sa corrélation avec l'intervalle $\mathcal{T}_0$.

On pose

$$\Delta = \hat{\alpha} - \alpha,$$

la matrice $T \times p$ d'erreurs d'estimation dont la t-ème ligne est

$$\delta_t = (x' P_z x)^{-1} x' P_z \varepsilon_t,$$

où $\varepsilon_t = (\varepsilon_{1t}, \ldots, \varepsilon_{nt})$ est le vecteur des résidus du modèle (2.2). En notant Vec(.) l'opérateur transformant une matrice en un vecteur en concaténant ses lignes, le pT-vecteur Vec(Δ) est distribué selon une loi normale centrée et de covariance $V_\delta = \Sigma \otimes (x' P_z x)^{-1}$ où $\otimes$ est le produit de Kronecker.

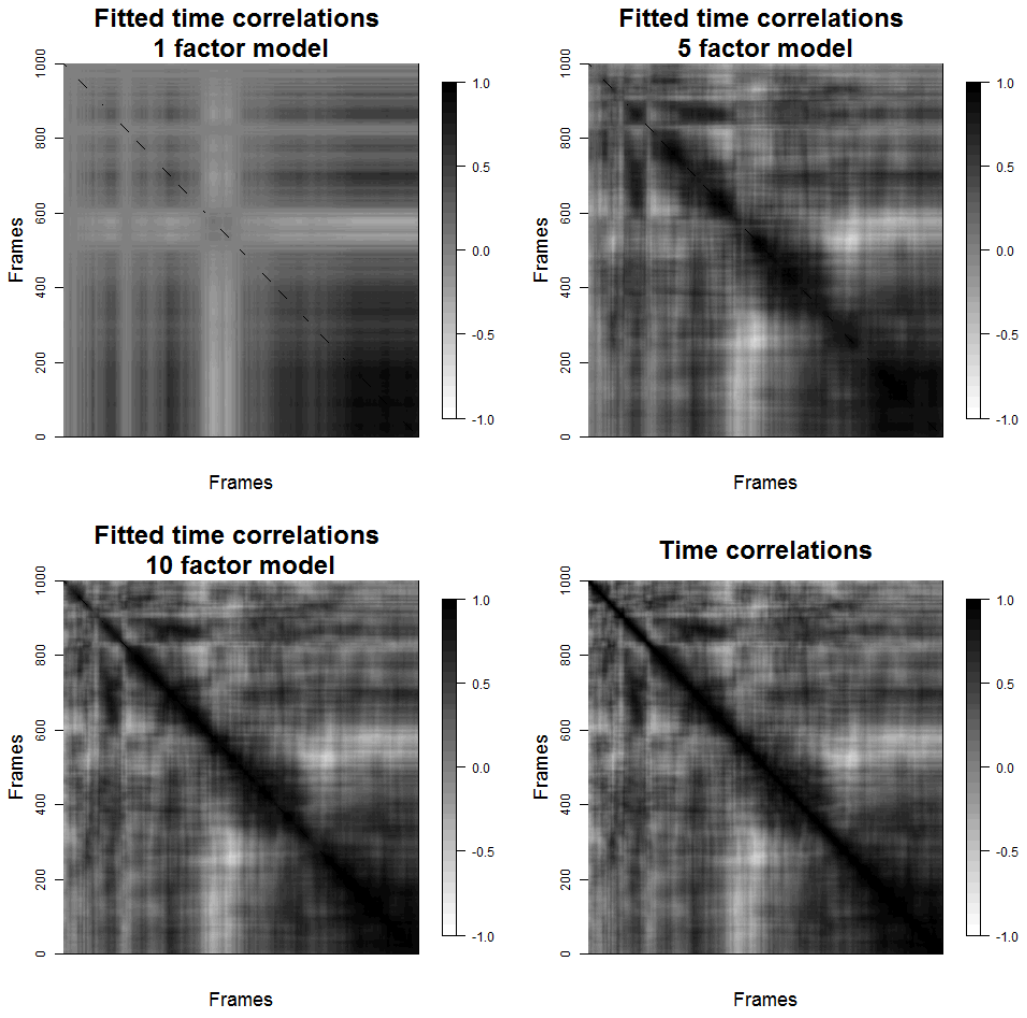


FIGURE 2.7 – Image plots of the fitted correlation matrix of the residuals of model (2.6) at channel CZ using factor models with 1, 5 and 10 factors : top and bottom-left panels, respectively. The bottom-right panel reproduces the right panel of Figure 1.1.

4.1.1 Correction de l'estimation du signal

Les précédentes expériences sur l'étude des ERPs permettent aux psychologues d'acquérir des notions sur l'instant auquel le signal doit commencer et combien de temps il devrait durer pour une condition expérimentale donnée. Lorsque cette connaissance a priori n'est pas acquise, on peut utiliser les résultats d'une étape de tests multiples préliminaire pour déterminer des instants où le signal a peu de chances d'être présent, c'est-à-dire des instants pour lesquels $\alpha_t = 0$ pour t appartenant à un ensemble $\mathcal{T}_0$. Ainsi, l'erreur d'estimation δ_t pour $t \in \mathcal{T}_0$ n'est pas mêlée au vrai signal α_t et donc $\Delta_0 = \hat{\alpha}_0$ où Δ_0 (resp. $\hat{\alpha}_0$) est la sous-matrice de Δ (resp. $\hat{\alpha}$) restreinte aux $t \in \mathcal{T}_0$. On note $\widetilde{\Delta}$ la matrice définie par :

$$\widetilde{\Delta} \;=\; \begin{pmatrix} \Delta_0 \\ \Delta_{-0} \end{pmatrix} \tag{2.7}$$

où Δ_{-0} est la sous-matrice de Δ dont les lignes sont δ_t, $t \notin \mathcal{T}_0$ et on note $\widetilde{V_\delta}$ un réarrangement des lignes de V_δ telle que

$$\widetilde{V_\delta} \;=\; \begin{pmatrix} \Sigma_{0,0} & \Sigma'_{-0,0} \\ \Sigma_{-0,0} & \Sigma_{-0,-0} \end{pmatrix} \otimes (x' P_z x)^{-1},$$

où $\Sigma_{0,0}$ (resp. $\Sigma_{-0,-0}$) est une sous-matrice de Σ restreinte aux lignes et colonnes correspondant aux temps t appartenant à $\mathcal{T}_0$ (resp. $t \notin \mathcal{T}_0$) et $\Sigma_{-0,0}$ est une sous-matrice de Σ restreinte aux lignes telles que $t \notin \mathcal{T}_0$ et aux colonnes telles que $t \in \mathcal{T}_0$.

Pour chaque $t \notin \mathcal{T}_0$, on prédit δ_t par Δ_0 par la prédiction linéaire :

$$\begin{aligned} \mathrm{vec}(\hat{\Delta}_{-0}) &= \big[\hat{\Sigma}_{-0,0} \otimes (x' P_z x)^{-1}\big]\big[\hat{\Sigma}_{0,0} \otimes (x' P_z x)^{-1}\big]^{-1} \mathrm{vec}(\Delta_0), \\ &= \big[\hat{\Sigma}_{-0,0} \otimes (x' P_z x)^{-1}\big]\big[\hat{\Sigma}_{0,0}^{-1} \otimes (x' P_z x)\big] \mathrm{vec}(\Delta_0), \\ &= \big[\hat{\Sigma}_{-0,0} \hat{\Sigma}_{0,0}^{-1}\big] \otimes \mathbb{I}_p \mathrm{vec}(\Delta_0), \end{aligned} \tag{2.8}$$

où $\mathbb{I}_p$ est la matrice identité $p \times p$ et $\hat{\Sigma}_{-0,0}$ et $\hat{\Sigma}_{0,0}$ sont des estimateurs de $\Sigma_{-0,0}$ et $\Sigma_{0,0}$ respectivement. La forme matricielle de l'Expression (2.8) est :

$$\hat{\Delta}_{-0} \;=\; \hat{\Sigma}_{-0,0} \hat{\Sigma}_{0,0}^{-1} \Delta_0. \tag{2.9}$$

L'Expression (2.9) fait appel à l'inversion de la matrice $\hat{\Sigma}_{0,0}$ ce qui peut poser des problèmes numériques du fait de la dimension du problème. On peut utiliser ici le modèle à facteurs (2.6) pour estimer la matrice complète Σ par $\hat{\Sigma} = \hat{\Psi} + \hat{B}\hat{B}'$ où $\hat{\Psi}$ est la matrice diagonale des variances spécifiques estimées $\hat{\psi}_t^2$ et $\hat{B}$ est la matrice $T \times q$ des *loadings* B estimés ($q \ll T$). La partition de Δ donne les partitions de Ψ et B suivantes :

$$\tilde{B} = \begin{pmatrix} B_0 \\ B_{-0,} \end{pmatrix}, \qquad \tilde{\Psi} = \begin{pmatrix} \Psi_0 & 0 \\ 0 & \Psi_{-0} \end{pmatrix}.$$

On en déduit des estimateurs de $\Sigma_{0,0}$ et $\Sigma_{-0,0}$:

$$\hat{\Sigma}_{-0,0} = \hat{B}_{-0}\hat{B}'_0, \qquad \hat{\Sigma}_{0,0} = \hat{\Psi}_0 + \hat{B}_0\hat{B}'_0.$$

Il est intéressant de noter que calculer $\hat{\Sigma}_{0,0}^{-1}$ dans l'Expression (2.9) ne fait appel qu'à l'inversion d'une matrice $q \times q$ selon la formule de Woodbury (Press et al. (2007)) :

$$\hat{\Sigma}_{0,0}^{-1} = \hat{\Psi}_0^{-1} - \hat{\Psi}_0^{-1}\hat{B}_0(I_q + \hat{B}_0'\hat{\Psi}_0^{-1}\hat{B}_0)^{-1}\hat{B}_0'\hat{\Psi}_0^{-1}.$$

Un estimateur $\hat{\Delta}^{(1)}$ de Δ est obtenu en remplaçant Δ_0 dans l'Expression (2.7) par $\hat{\Delta}_0$ obtenu par l'Expression (2.9). Un nouvel estimateur de α est obtenu en corrigeant l'estimateur précédent $\hat{\alpha}$ de l'erreur d'estimation prédite :

$$\hat{\alpha}^{(1)} \quad = \quad \hat{\alpha} - \hat{\Delta}^{(1)}.$$

Ce nouvel estimateur est utilisé pour actualiser le calcul des résidus $\hat{\varepsilon}$:

$$\hat{\varepsilon}^{(1)} = P_z(Y - x\hat{\alpha}^{(1)'}).$$

Un nouvelle décomposition en facteurs de la matrice de covariance des résidus est calculée, menant à une nouvelle estimation de Δ puis une nouvelle estimation $\hat{\alpha}^{(k)}$ du signal, où l'exposant k indique l'indice d'itération de l'algorithme. Ces étapes sont alternées jusqu'à ce que le critère de convergence de l'estimation de α soit atteint.

4.1.2 Décorrélation de la statistique de test par ajustement sur les facteurs

L'algorithme EM proposé par Rubin and Thayer (1982) et détaillé dans Friguet et al. (2009) est adapté au présent contexte. Une fois qu'un modèle à facteurs est estimé sur des covariables dépendantes et destinées à une procédure de tests multiples, les nouvelles statistique de test $\widetilde{\mathcal{T}}$ (supposées indépendantes) associées aux tests d'hypothèse $H_{0t}, t = 1, \ldots, T$ sont appelées *statistiques de tests ajustées sur les facteurs*.

Finalement, la procédure d'ajustement adaptatif sur les facteurs proposée alterne l'étape d'estimation du signal corrigé de l'erreur de prédiction (par modèle à facteurs de la structure de dépendance) et de calcul des statistiques de test ajustées sur les facteurs, utilisées pour actualiser l'ensemble $\mathcal{T}_0$. A l'étape k de l'algorithme, on a $\mathcal{T}_0^k$ et les estimateurs $(\hat{\Psi}_k, \hat{B}_k, \hat{F}_k)$ des paramètres du modèle à facteurs. L'étape $(k+1)$ se décompose en deux parties :

- Estimation de l'erreur d'estimation $\hat{\Delta}^{(k+1)}$ et actualisation de l'estimation du signal par $\hat{\alpha}^{(k+1)} = \hat{\alpha} - \hat{\Delta}_{(k+1)}$. Les résidus sont ensuite actualisés : $\hat{\varepsilon}^{(k+1)} = P_z(Y - x\hat{\alpha}^{(k+1)'})$;
- Calcul des paramètres du modèle à facteurs $(\hat{\Psi}_{k+1}, \hat{B}_{k+1}, \hat{F}_{k+1})$ à partir des nouveaux résidus $\hat{\varepsilon}^{(k+1)}$. Les statistiques de tests facteurs-ajustées sont calculées selon l'Expression (2.3) puis $\mathcal{T}_0$ est actualisé. La mise à jour de $\mathcal{T}_0^k$ est faite pour favoriser la sélection d'instants pour lesquels le signal a de grandes chances d'être nul. On préfère retenir un grand nombre de faux positifs plutôt qu'avoir une règle de sélection plus précise, qui recouvrerait mieux le vrai ensemble $\mathcal{T}_0$ mais qui augmenterait le risque de retenir le support du signal. On suggère la règle suivante : $\mathcal{T}_0^{k+1} = \left\{ t = 1, \ldots, T, \widetilde{p}_t^{(k+1)} \geq 0.2 \right\}$ où $p_t^{(k+1)}$ est la probabilité critique associée à la statistique de test à l'étape k pour le temps t.

Les itérations cessent à l'étape k si $\mathcal{T}_0^{k+1} = \mathcal{T}_0^k$.

En pratique, l'estimation finale de $\mathcal{T}_0$ est peu sensible au choix du seuil sur les probabilités critiques (0.2 proposé ici) sous réserve que le choix du seuil ne soit pas extrême : une trop petite valeur risque d'effacer le signal et de sur-contrôler le FDR alors qu'une valeur proche de 1 aura tendance à produire les mêmes résultats que la méthode choisie pour initialiser l'algorithme, à savoir la méthode des moindres carrés ici.

4.1.3 Illustration sur un exemple

Afin d'illustrer comment la régularité de la courbe du signal estimé induite par la forte dépendance dans la matrice V_δ peut générer une confusion entre le vrai signal α et l'erreur d'estimation Δ, on reprend un jeu de données simulé selon le plan de simulations décrit en Section 3.2 avec une amplitude de signal $\max_t \alpha_t = 5$ dont le support est l'intervalle [450 ms ;550 ms]. Le trait plein du graphique supérieur de la Figure 2.8 représente les valeurs de la statistique de test calculée à partir de l'estimation par la méthode des moindres carrés du signal. Les probabilités critiques brutes sont corrigées par la méthode de BH. On constate que la procédure ne détecte pas le support du signal et déclare significatifs des intervalles de temps en dehors du support de α. Cet exemple montre que le rang des probabilités critiques est affecté lorsque les procédures de comparaisons multiples sont appliquées sans prise en compte de la dépendance.

La partie inférieure de la Figure 2.8 montre que ni SVA (voir Leek and Storey (2007)) ni LEAPP (voir Sun et al. (2012)) ne parviennent à séparer le signal du bruit, ce qui entraine la déclaration à tort d'instants de temps significatifs.

Afin d'illustrer l'impact de la connaissance a priori de $\mathcal{T}_0$ sur la procédure d'estimation proposée dans ce chapitre, on suppose d'abord que le signal est nul sur l'ensemble $\mathcal{T}_0 = [1, 100] \cup [901, 1000]$. La courbe en pointillés courts sur la partie supérieure de la Figure 2.8 indique les valeurs de la statistique de test obtenues par la méthode AFA à partir d'un a priori correct. L'intervalle déclaré significatif après correction de BH des probabilités critiques associées à la statistique de test ajustée sur les facteurs avec un contrôle du FDR fixé à 0.05 indique que le support du signal est correctement détecté. La courbe en pointillés longs représente les valeurs de la statistique de test obtenues par la méthode AFA à partir d'un a priori faux $\mathcal{T}_0 = [450, 550]$, c'est-à-dire, exactement le support du signal. Initialisé avec un tel a priori, la méthode échoue en déclarant significatifs des intervalles où le vrai signal est absent.

5 Etude par simulations

5.1 Méthodes

L'estimation jointe du signal et de la matrice de covariance résiduelle afin de décorréler les statistiques de test peut être combinée à n'importe quelle méthode

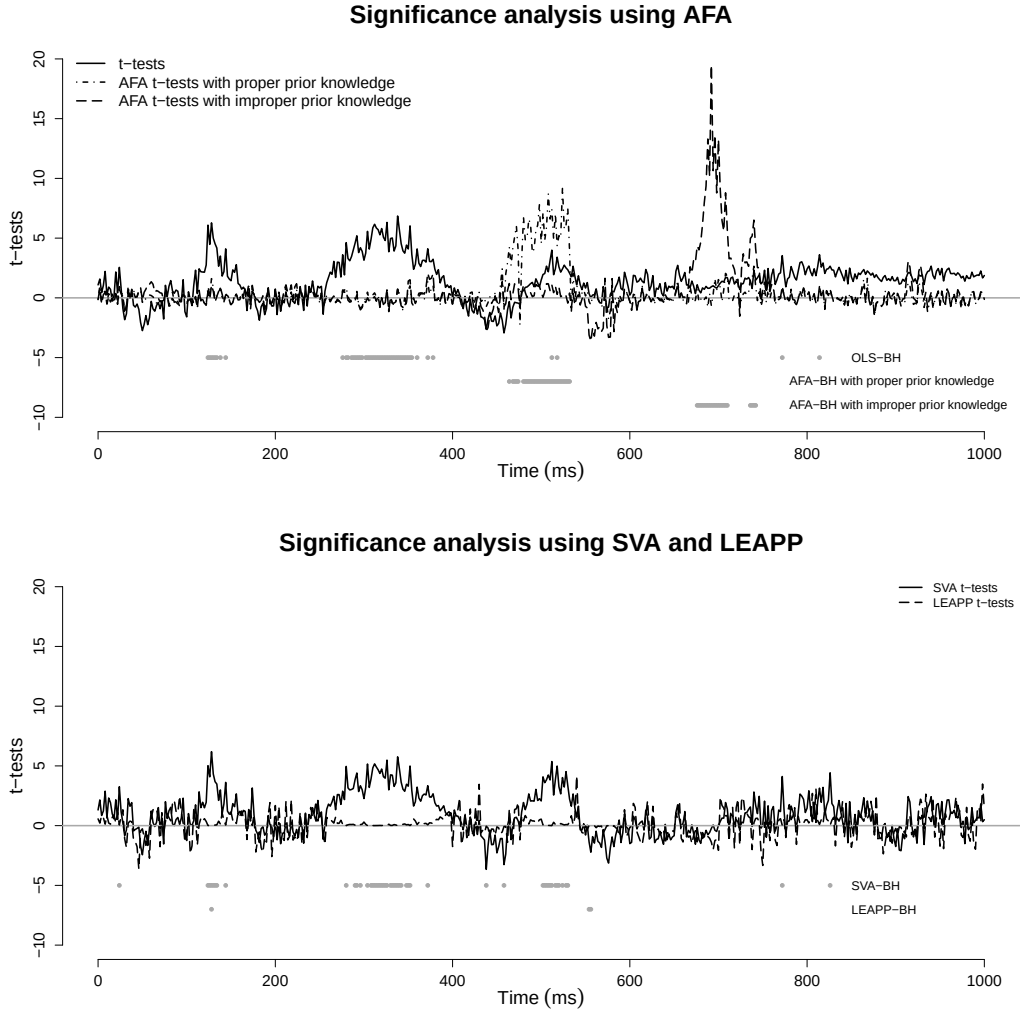


FIGURE 2.8 – t-statistics for a single simulation run. Top panel : the OLS estimation of the signal (solid curve) and t-statistics after Adaptive Factor Adjustment, based on $\mathcal{T}_0 = [1, 100] \cup [901, 1000]$ (dotted) and $\mathcal{T}_0 = [450, 550]$ (dashed). Bottom panel : t-statistics after SVA (solid) and LEAPP (dashed) adjustment. The gray dots above the x-axis indicate significant time points identified by the BH method controlling the FDR at 0.05 level.

de seuillage associée au contrôle de l'erreur de type I, du FDR ou du FWER. La procédure de BH est une méthode standard de correction des probabilités critiques en tests multiples lorsque les tests sont indépendants. Ainsi, la méthode de BH est la procédure choisie pour corriger les probabilités critiques produites par la méthode AFA. Cette combinaison de l'estimation adaptative des modèle à facteurs et de la procédure de BH est appelée dans la suite *procédure de tests multiples AFA*.

Les performances de la procédure AFA sont comparées à celles de méthodes existantes en tests multiples, soit parce qu'elles sont beaucoup utilisées, soit parce qu'elles sont construites pour prendre en compte la dépendance. La procédure de BY garantit le contrôle du FDR sous des conditions spécifiques sur la dépendance bien qu'elle ne corrige pas l'impact de la corrélation en ajustant les probabilités brutes. Les méthodes SVA et LEAPP représentent les approches développées récemment basées sur un modèle de régression à facteur similaire à (2.6) pour décorréler les statistiques de test. La méthode de Causeur et al. (2012) n'est pas inclue dans les comparaisons car elle est dépassée par AFA.

Pour résumer, on compare les performances des six méthodes de tests multiples pour les données ERPs suivantes (le niveau de contrôle du FDR fixé à 0.05 pour toutes les méthodes) :

1. BH : procédure de Benjamini-Hochberg (Benjamini and Hochberg (1995)) appliquée aux probabilités critiques brutes
2. BY : procédure de Benjamini-Yekutieli (Benjamini and Yekutieli (2001)) appliquée aux probabilités critiques brutes
3. GB : procédure de Guthrie-Buchwald (Guthrie and Buchwald (1991)) appliquée aux probabilités critiques brutes avec un seuil graphique fixé à 0.05
4. LEAPP : procédure d'ajustement sur des facteurs latents (*latent effect adjustment after primary projection*, Sun et al. (2014)) couplée à un contrôle du FDR par la méthode de BH
5. SVA : procédure d'ajustement sur des facteurs latents (*surrogate variable analysis*, Leek and Storey (2008)) couplée à un contrôle du FDR par la méthode de BH. On utilise les options par défaut du package `R sva` (Leek et al. (2014)) pour estimer le modèle et le nombre de facteurs.
6. AFA : méthode proposée d'ajustement adaptatif sur les facteurs avec un contrôle du FDR. L'a priori $\mathcal{T}_0$ est fixé à l'ensemble des temps pour lesquels les probabilités critiques du test de Student usuel sont supérieures ou égales à 0.2. Le nombre de facteurs est estimé pour chaque jeu de données simulé en minimisant le critère d'inflation de la variance (Friguet et al. (2009)) comme implémenté dans le package `R ERP` (Causeur and Sheu (2014)).

Le plan de simulations est le même que dans la Section 3.2. Les Figures 2.9, 2.10 et 2.11 présentent les résultats en terme de FDR, précision (espérance de la proportion de rejets corrects parmi le nombre de rejets) et PNR (probabilité de n'avoir aucun rejet à tort). La Table 2.2 présente la sensibilité et la résolution des 6 méthodes. La procédure AFA dépasse les autres méthodes selon les critères mesurés dans cette comparaison. La Figure 2.9 montre que la méthode AFA hérite

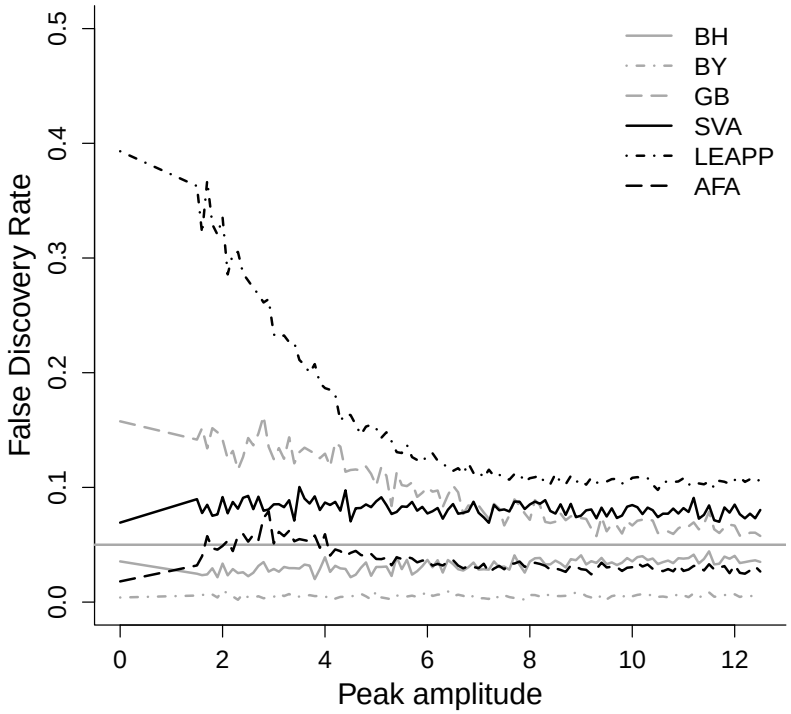


FIGURE 2.9 – Comparaison du taux de faux positifs (FDR) pour 6 procédures de tests multiples, calculé à partir de 1000 jeux de données ERPs simulés pour chaque amplitude de pic.

des bonnes propriétés de la procédure de BH en terme de contrôle du FDR, ce qui n'est pas le cas pour les méthodes de GB, LEAPP ou SVA, surtout dans les cas où le signal est faible ou modéré. De plus, ce contrôle du FDR n'est pas affecté par l'instabilité causée par la dépendance comme expliqué dans la Section 3.2. La Figure 2.11 montre que la probabilité de non rejet de AFA est parmi les plus faibles lorsque le signal est élevé ou modéré, ce qui assure que le *positive FDR* (l'espérance de la proportion de faux positifs sachant qu'on observe au moins un rejet) est proche du FDR. Sur la Figure 2.10, la courbe de précision confirme que AFA est performant pour détecter des signaux d'amplitude modérée à élevée.

5.2 RÉSULTATS

La Table 2.2 montre que les trois méthodes fondées sur de la décorrélation, SVA, LEAPP et AFA, sont assez sensibles, surtout LEAPP. Ceci est cohérent avec la Figure 2.9 où on remarque que LEAPP ne contrôle pas le FDR pour un signal faible ou modéré. En termes de résolution, AFA a le niveau le plus faible, suivi par la méthode de GB. Ceci confirme que la modélisation de la dépendance temporelle

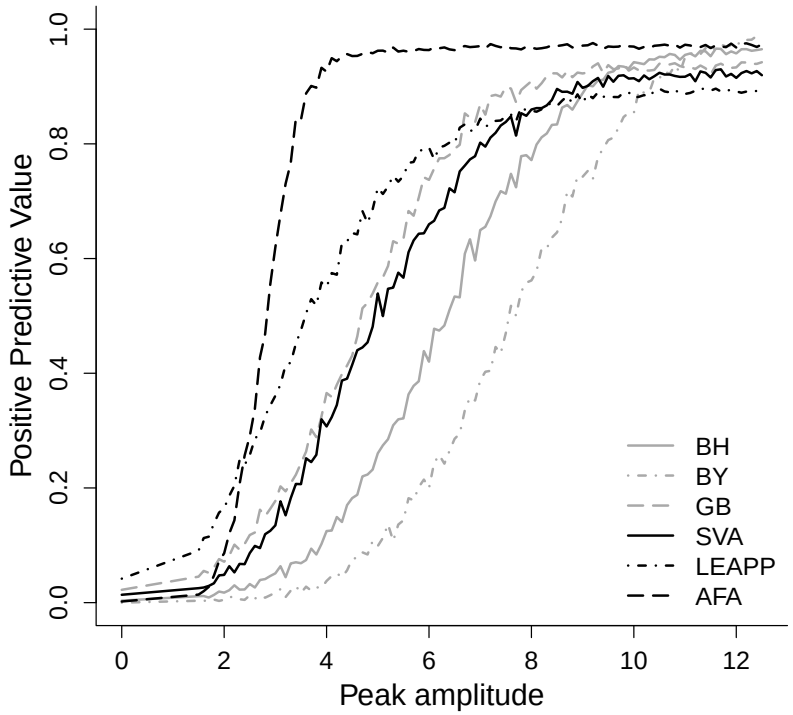


FIGURE 2.10 – Comparaison de la précision (PPV) pour 6 procédures de tests multiples, calculé à partir de 1000 jeux de données ERPs simulés pour chaque amplitude de pic.

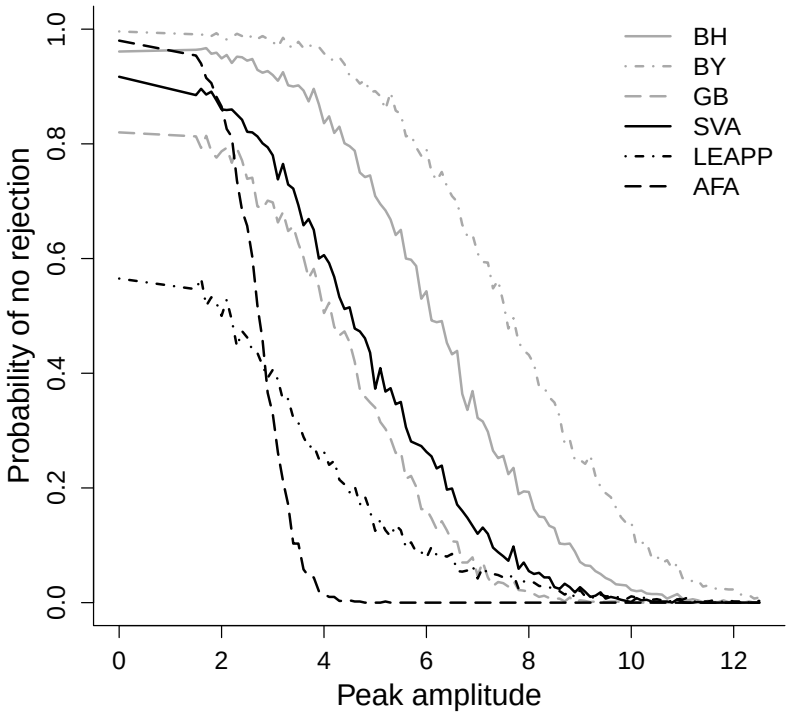


FIGURE 2.11 – Comparaison de la probabilité de non rejet (PNR) pour 6 procédures de tests multiples, calculé à partir de 1000 jeux de données ERPs simulés pour chaque amplitude de pic.

TABLE 2.2 – Sensibilité et résolution de 6 procédures de tests multiples calculées sur des simulations

	Méthodes					
	BH	BY	GB	SVA	LEAPP	AFA
Sensibilité	3.6	4.9	0.0	1.5	0.0	1.9
Résolution	9.1	10.4	7.9	8.9	>12.5	3.7

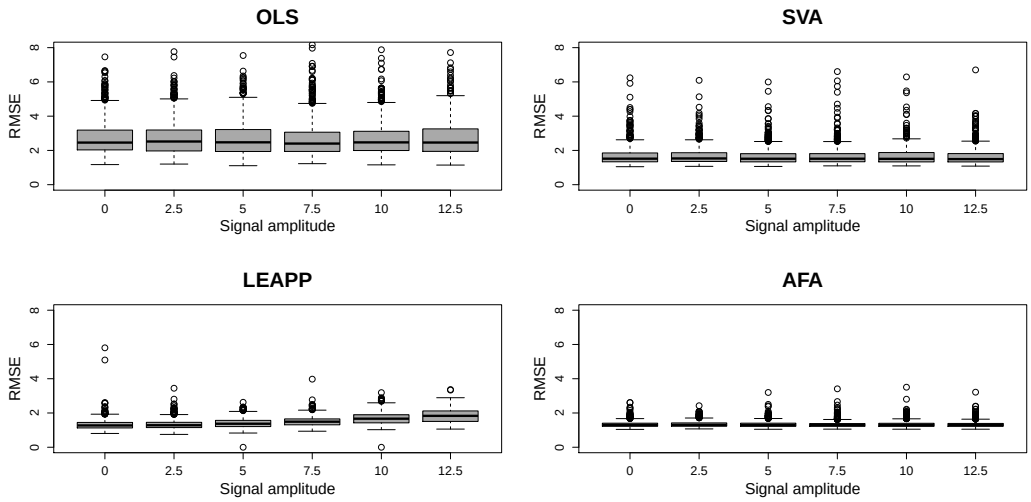


FIGURE 2.12 – Racine de l'erreur quadratique moyenne (RMSE) de l'estimation de α, calculée à partir de 1,000 jeux de données pour chaque amplitude de signal. La RMSE est comparée pour 4 méthodes : moindres carrés, SVA, LEAPP et AFA.

résiduelle est utile pour séparer le signal du bruit. On déduit de la Figure 2.12 la même conclusion. Les boîtes à moustaches de cette figure présentent la racine de l'erreur quadratique moyenne (RMSE) de l'estimation de α pour la méthode des moindres carrés, SVA, LEAPP et AFA. La méthode des moindres carrés, qui ignore la dépendance, est celle qui présente les moins bonnes performances d'estimation en moyenne. Les trois méthodes fondées sur la décorrélation possèdent un RMSE plus faible que les moindres carrés et AFA est celle dont le RMSE est le plus faible, quelle que soit l'amplitude du signal. On constate de nouveau que la méthode LEAPP extrait plus difficilement les signaux d'amplitude élevée car les boîtes à moustaches affichent des performances de moins en moins bonnes lorsque le signal augmente.

6 RÉSULTATS SUR LES ERPS

Les psychologues associés au projet proposent une interprétation des résultats de la méthode AFA appliquée aux données d'ERPs issues de l'expérience d'oubli direct. D'après la littérature, il semble que la capacité à reconnaître des mots qu'on nous a demandé d'oublier repose plus sur le sentiment de familiarité que la capacité

à reconnaître des mots qu'on nous a demandé de mémoriser (Gardiner et al. (1994)). Des études empiriques sur la mémoire de reconnaissance faisant appel aux ERPs ont indiqué que la phase précoce de reconnaissance faisant appel à la familiarité est associée à des modulations de la composante ERP FN400, une positivité augmentée pour les mots anciens par rapport aux nouveaux est observée approximativement 400 à 600 ms après le début du stimulus. Le fait que la composante FN400 augmente progressivement avec la reconnaissance suggère aussi que cette composante est un indice de familiarité (voir Rugg and Curran (2007)). Bien que l'expérience d'oubli direct décrite en Section 2.1 soit exploratoire, des études précédentes indiquent qu'on peut s'attendre à des intervalles de temps significatifs entre 400 et 600 ms. Qualitativement, on peut s'attendre à des intervalles de temps significatifs en retard pour la condition TBF pour les électrodes postérieures. La confirmation de ces attentes est importante pour la compréhension du mécanisme neurophysiologique concernant le contrôle intentionnel de la mémorisation ou de l'oubli.

Une application directe de la méthode de BH aux données ERPs issues de l'expérience d'oubli direct n'identifie aucun intervalle de temps significatif sur aucune des 9 électrodes. Afin d'appliquer la méthode AFA, la connaissance à priori $\mathcal{T}_0$ pour les électrodes frontales et centrales est fixée à $[1, 200] \cup [901, 1000$ ms et à $[1, 200]$ ms pour les électrodes postérieures. Le critère d'inflation de la variance suggère 2 facteurs pour toutes les électrodes. Les graphiques de la Figure 2.13 présentent les courbes de corrélation de l'électrode CZ pour les deux instructions obtenues par la méthode des moindres carrés et par la méthode AFA. La méthode AFA détecte, dans les deux conditions, l'intervalle de temps [400,700] ms significatif avec des corrélations positives et un large pic, précédé par un pic négatif pour la condition TBF.

La Figure 2.14 donne une représentation spatiale des électrodes et les courbes de corrélation associées obtenues par la méthode AFA. Des pics de corrélation positive significatifs apparaissent principalement sur l'intervalle [400,700] ms pour les deux conditions pour les neuf électrodes. De plus, la méthode AFA confirme que des intervalles de corrélations négatives apparaissent sur la plupart des électrodes pour la condition TBF uniquement. Cette inflexion des courbes de corrélation autour de 400 ms, plus visible dans la condition TBF, implique une relation entre l'instruction et la modulation de la composante FN400.

7 Conclusion

L'analyse à grande échelle des tests univariés des données de potentiels évoqués cognitifs cérébraux (Groppe et al. (2011a)) est reconnue comme un problème de recherche difficile car le signal sous-jacent aux ERPs est souvent rare, intervenant seulement durant de brefs instants durant l'expérience, et faible par rapport à la variabilité inter-individuelle (voir Donoho and Jin (2008) et Jin (2009) pour la terminologie sur le modèle *rare and weak features*). Lorsque l'on teste simultanément plusieurs hypothèses sur un grand nombre de mesures au cours du temps (Woolrich et al. (2009)), la nécessité de contrôler le taux de faux positifs est confrontée à celle

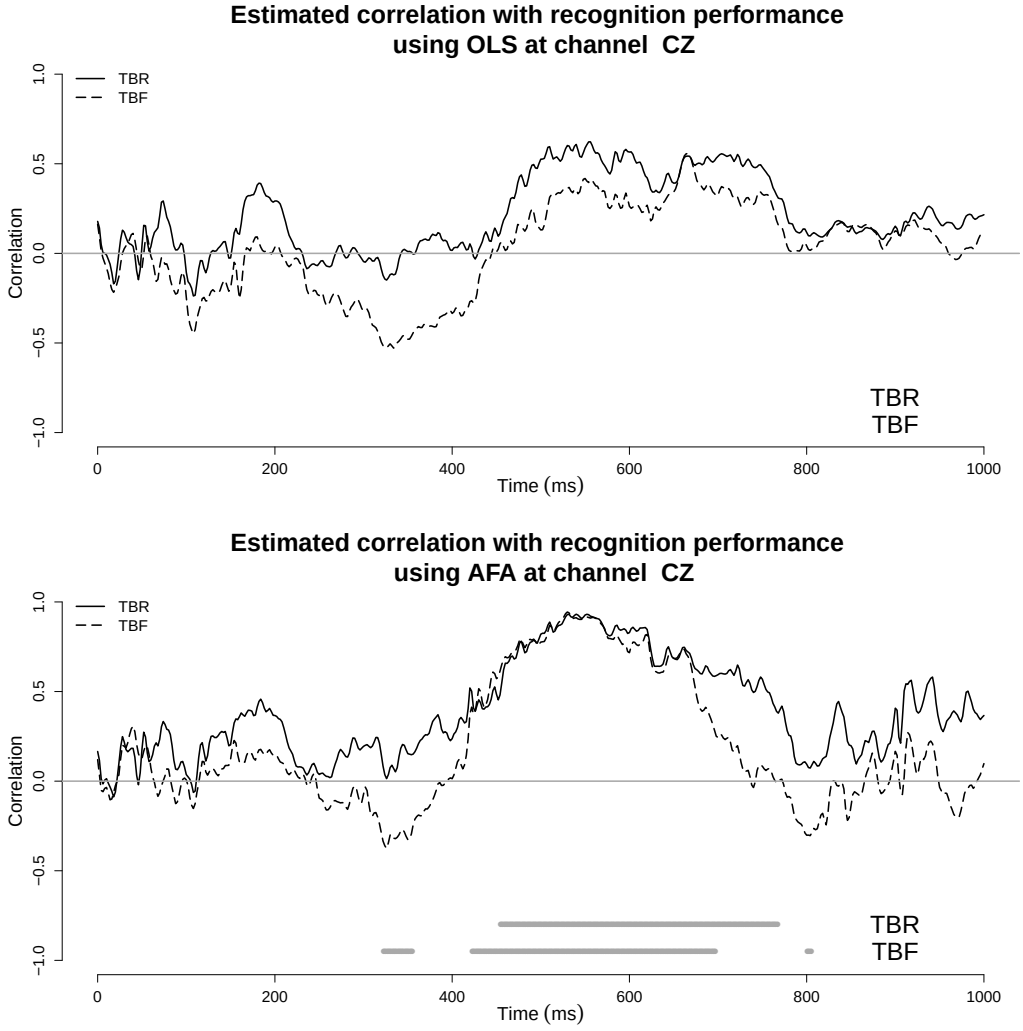


FIGURE 2.13 – Corrélations entre les ERPs et le score de reconnaissance pour les deux conditions (trait plein pour la condition TBR, pointillé pour la condition TBF) estimées par la méthode des moindres carrés (graphique du haut) et par AFA (graphique du bas). Les intervalles de temps significatifs sont indiqués par des points gris sous l'axe des abscisses.

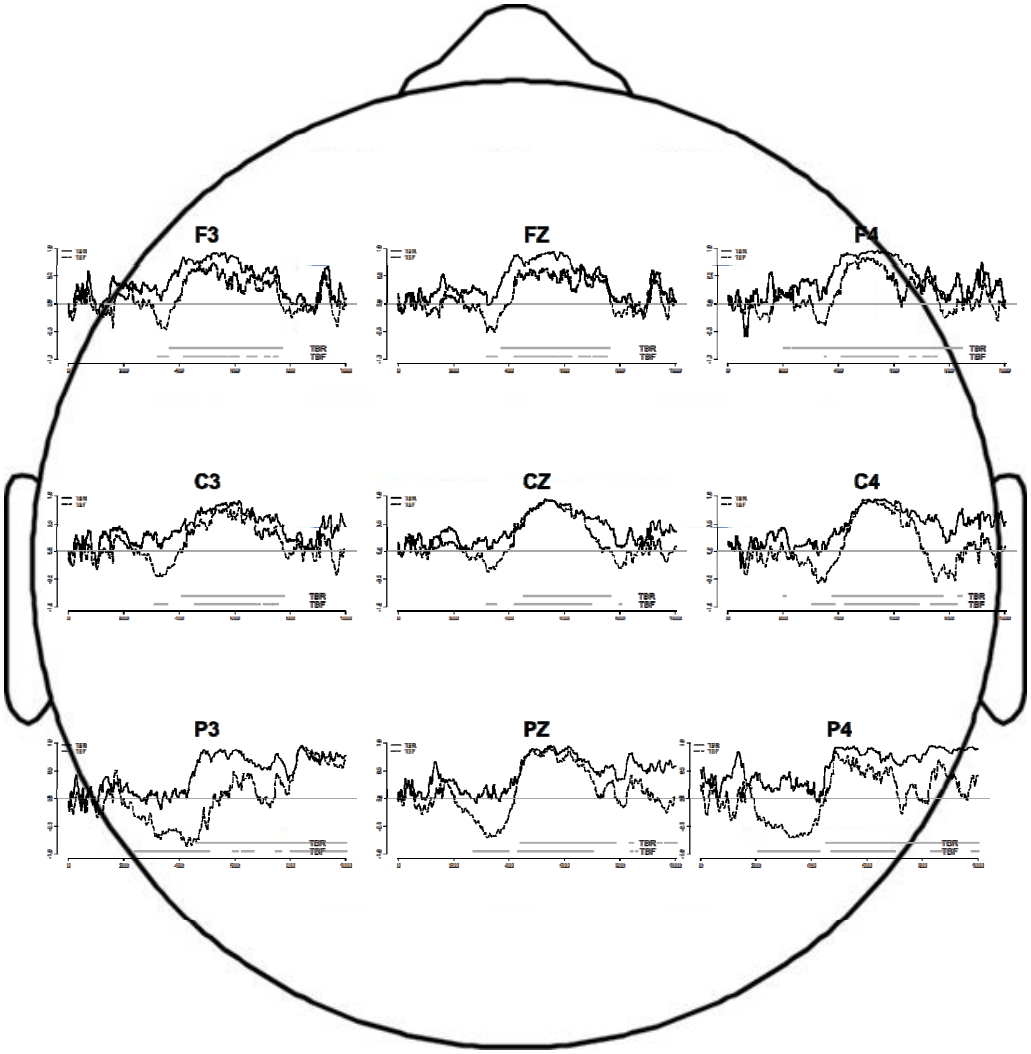


FIGURE 2.14 – Corrélations (trait plein pour TBR, pointillé pour TBF) estimées par la méthode AFA et intervalles de temps déclarés significatifs, pour les 9 électrodes.

de maintenir une puissance de détections raisonnable et les méthodes favorisant l'un ou l'autre sont sujettes à controverses (Vul et al. (2009)).

La forte dépendance temporelle caractéristique des ERPs ajoute une difficulté supplémentaire au traitement de ces données car les résultats des procédures de tests multiples habituellement observées sous l'indépendance deviennent instables et échouent à localiser et à estimer l'amplitude du signal, même après correction de la dépendance. La méthode d'ajustement sur les facteurs proposée traite la problématique de l'analyse de tests univariés dans le contexte d'un modèle linéaire multivarié par une modélisation de la dépendance via des facteurs latents et une estimation jointe du signal et du bruit sous-jacent, sachant une connaissance a priori d'intervalles de temps pour lesquels le signal est absent. Une procédure itérative est proposée pour estimer les paramètres du modèle. Enfin, la méthode est implémentée dans un package `R` intitulé `ERP` et disponible sur le `CRAN` (voir Causeur and Sheu (2014)).

Les tests par permutation (voir Blair and Karniski (1993), Westfall and Young (1993)) sont couramment utilisés pour l'analyse de données ERPs, ils ne sont pas évoqués dans ce chapitre car une étude récente (voir Lage-Castellanos et al. (2010)) montre que la méthode de Benjamini-Hochberg (voir Benjamini and Hochberg (1995)) et la méthode du FDR local (voir Efron (2007)) donnent le même compromis entre l'erreur de type I et de type II (comparé aux tests par permutation) dans des situations pour lesquelles aucune information a priori sur les instants et les électrodes où des différences d'ERPs ont lieu n'est disponible. D'après leur conclusion, les résultats observés dans l'étude comparative de la Section 5 sont encourageants : la méthode proposée surpasse les autres méthodes tout en contrôlant le FDR et en maintenant une bonne puissance de détections. De plus, l'analyse exploratoire des données issues de l'expérience d'oubli direct illustre le fait que la méthode AFA est adaptée à la détection de pics ERPs faibles, mêlés à un bruit complexe et fortement dépendant.

Cette procédure d'estimation est vraisemblablement applicable à d'autres problèmes de tests multiples où la dépendance est forte : par exemple pour l'analyse de longueurs d'onde issues de mesures en spectroscopie infra-rouge (NIRS) ou distribuées dans l'espace lors de l'analyse de mesures par imagerie à résonance magnétique (fMRI).

CHAPITRE 3

STABILITÉ DE LA SÉLECTION DE VARIABLES EN CLASSIFICATION SUPERVISÉE POUR DES DONNÉES DÉPENDANTES DE GRANDE DIMENSION

Résumé : la prise en compte de la dépendance dans les procédures de sélection de variables n'est pas une approche systématique dans le contexte de la classification supervisée en grande dimension. En effet, certains articles récents montrent la supériorité des approches de type *naive Bayes* fondées sur une hypothèse peu réaliste d'indépendance entre les prédicteurs alors que d'autres auteurs recommandent de modéliser la structure de dépendance afin de décorréler les statistiques de sélection. Dans le contexte de l'analyse linéaire discriminante (LDA), ce chapitre illustre dans un premier temps l'impact de la dépendance sur la stabilité des résultats en terme de sélection. Le second objectif est de proposer une méthode améliorant cette stabilité en supposant une décomposition en facteurs pour la structure de covariance. Les variables latentes introduites par le modèle à facteurs permettent de définir une nouvelle règle de Bayes conditionnelle. Une procédure d'estimation jointe de l'espérance et de la variance du modèle est ensuite proposée et comparée à des approches récentes d'analyse diagonale discriminante régularisée, supposant l'indépendance entre prédicteurs, et à des procédures de LDA régularisée. La comparaison est faite en terme de performance de classification et de stabilité de l'étape de sélection. La méthode proposée est implémentée dans un package `R` intitulé `FADA` et disponible sur le CRAN.

Sommaire

Le chapitre précédent présente une façon de tirer profit de la particularité des données ERPs (à savoir la dépendance temporelle élevée liant les mesures entre elles) afin de décorréler efficacement les statistiques d'association entre ERP et condition expérimentale. Une procédure de tests multiples atteignant de bonnes performances de sélection d'intervalles de temps pertinents est ainsi proposée.

Dans le présent chapitre, on revient à un cadre plus général de dépendance, qu'on autorise à être faible ou forte, voire absente. Ce chapitre est motivé par des applications en génomique notamment, où beaucoup d'applications consistent à classer des patients en groupes (stades d'un cancer, sain/malade, tumeur maligne/bénigne) et à mesurer, par exemple, les profils d'expression de certains gènes à mettre en relation avec ces différents groupes. On se place donc dans un contexte de classification supervisée en grande dimension, dont le but est d'établir un modèle parcimonieux

de prédiction de la classe d'un individu. Le but n'est donc plus seulement d'identifier le support du signal. Comme le Chapitre 2, ce Chapitre 3 repose sur le principe commun de tirer profit de la structure de dépendance pour améliorer l'estimation du signal (ici, les moyennes par classes) et permettre une décorrélation des données par ajustement sur les facteurs latents performante.

Ce chapitre est associé à un article intitulé "Stability of feature selection in classification issues for high-dimensional correlated data", publié dans la revue *Statistics and Computing* (Perthame et al. (2015)).

1 Introduction

Les procédures de classification supervisée sont utilisées pour prédire la classe d'un individu à partir de son profil biologique. Dans ce contexte, l'impact de la dépendance entre prédicteurs sur ces procédures pose encore un problème statistique. De récentes études sur l'effet de la dépendance sur les performances des procédures de classification dans des situations où le nombre de variables excède le nombre d'individus ont menées à des conclusions contradictoires.

En effet, la supériorité de certaines méthodes fondées sur une hypothèse peu réaliste d'indépendance entre les covariables est constatée (Dudoit et al. (2002), Levina (2002), Bickel and Levina (2004)) alors que de plus en plus de méthodes prennent en compte la structure de dépendance (voir entre autres Guo et al. (2007), Dabney and Storey (2007), Xu et al. (2009) et Zuber and Strimmer (2009)). Plus récemment, Ahdesmäki and Strimmer (2010) approfondit ce sujet en revisitant l'approche *naive Bayes* proposée par Efron (2008), aussi connue sous le nom d'analyse diagonale discriminante (DDA), en définissant des statistiques de tests décorrélées. Les variables sont dans ce cas supposées indépendantes et estimer le support du signal (c'est-à-dire le sous-ensemble de prédicteurs discriminants) revient à étudier l'ordre de statistiques de test, ce qui correspond à une situation de comparaisons multiples. Cependant, Ahdesmäki and Strimmer (2010) précise que les procédures de tests multiples permettent le contrôle du nombre de faux positifs alors que les procédures de sélection cherchent à contrôler le nombre de variables prédictives non sélectionnées à tort dont l'optique est la prédiction. Comme cités dans l'introduction de ce manuscrit, plusieurs auteurs reportent l'impact négatif de la forte dépendance entre covariables sur la consistance des statistiques d'ordre des probabilités critiques associées aux tests d'hypothèses (voir notamment Leek and Storey (2007), Leek and Storey (2008), Friguet et al. (2009), Sun et al. (2012)). Ces auteurs proposent de gérer les corrélations par une modélisation jointe des relations entre les covariables et des variances résiduelles par un modèle supposant l'existence de facteurs latents captant linéairement la dépendance entre les variables. Le but de ce chapitre est de proposer une méthode de modélisation de la dépendance adaptée au contexte de la classification supervisée.

Le premier objectif de ce chapitre est d'illustrer l'instabilité de la sélection de variables dans un contexte d'analyse linéaire discriminante (LDA) lorsque le nombre

de variables dépasse le nombre d'observations. Dans ce contexte de grande dimension, les procédures régularisées fondées sur la pénalisation ℓ_1 ou ℓ_2 des fonctions de perte usuelles sont connues pour atteindre un équilibre entre le compromis biais-variance dans l'estimation des scores discriminants (voir Tibshirani et al. (2002) et Tibshirani et al. (2003) pour une DDA régularisée, Hastie et al. (1995) pour une LDA pénalisée et Friedman et al. (2010) pour une pénalisation *elastic net* de la déviance d'un modèle). La stabilité et les performances de classification, ainsi que l'impact de la dépendance sur la répétabilité des résultats d'une de ces procédures classiques sont étudiés.

Les principales méthodes de sélection de variables en classification supervisée sont rappelées dans la Section 2 de ce chapitre. Une étude par simulations illustrant l'impact de la dépendance sur le sous-ensemble des variables sélectionnées par le Lasso est présentée dans la Section 3. Le modèle à facteurs et la méthode de décorrélation des données par ajustement sur l'effet de variables latentes captant la dépendance sont décrits en Section 4. La Section 5 étudie l'apport de la méthode proposée sur des données réelles et la Section 6 illustre les propriétés de la méthode sur des simulations. Enfin, la méthode proposée est implémentée dans un package `R` intitulé `FADA` dont le fonctionnement est expliqué dans la Section 7. Enfin, la Section 8 conclut ce chapitre.

2 Sélection de variables pour la classification en grande dimension

Ce chapitre s'inscrit dans un contexte de classification supervisée, où la variable de groupe notée Y prend ses valeurs dans $[0; K-1]$. On note p_y la probabilité de l'événement $Y = y$ pour tout $y \in [0; K-1]$. Un individu est caractérisé par son groupe, connu pour un échantillon d'apprentissage. Le label d'un individu test est inconnu mais l'on suppose qu'il appartient à $[0, K-1]$. Chaque observation est caractérisée par un m-vecteur de covariables X que l'on suppose normalement distribué. On observe les profils de n individus et on note n_y le nombre d'individus dans la classe y. Le problème statistique est de construire un modèle permettant de prédire le label d'un individu à partir de ses covariables. Pour cela, on distingue deux méthodes standards bien connues : la LDA et le modèle linéaire généralisé (GLM). Le but de cette section est de rappeler les fondements de l'analyse linéaire discriminante et de la régression logistique en général puis en grande dimension. Quelques méthodes développées dans ce contexte, comparées lors d'une étude par simulations ou sur données réelles dans la suite du chapitre, sont présentées.

2.1 Analyse linéaire discriminante

Plusieurs approches permettent de définir l'analyse linéaire discriminante. La règle de Bayes repose sur l'hypothèse de normalité des covariables, l'analyse discriminante de Fisher est fondée sur la maximisation de la variance inter-classes contre la minimisation de la variance intra classes. Enfin, on peut définir l'analyse linéaire

discriminante comme une régression sur des variables indicatrices dites *dummy variables*.

Règle de classification de Bayes Pour établir la règle de Bayes, on suppose que les covariables suivent une loi normale :

$$X|Y=y \quad \sim \quad \mathcal{N}_m(\mu_y, \Sigma). \tag{3.1}$$

Chaque groupe est caractérisé par une moyenne différente d'un groupe à l'autre et les structures de covariances sont supposées égales entre les groupes. Sous ces conditions et en appliquant le théorème de Bayes, la probabilité d'appartenir au groupe y sachant un profil x s'écrit :

$$\mathbb{P}(y|x) \quad = \quad \frac{p_y f_X(x|Y=y)}{f_X(x)}.$$

où $f_X(\cdot)$ (resp. $f_X(\cdot|Y=y)$) est la fonction de densité des covariables X (resp. conditionnelle à Y). En supprimant les termes constants entre les groupes, calculer la log-probabilité $\log(\mathbb{P}(y|x))$ est en fait équivalent à calculer le score $d(y|x)$:

$$\log(\mathbb{P}(y|x)) \quad \propto \quad d(y|x) = \log p_y - 0.5\mu_y'\Sigma^{-1}\mu_y + x'\Sigma^{-1}\mu_y. \tag{3.2}$$

Ce score est appelé *règle de classification de Bayes* ou *classifieur de Bayes*. On remarque que ce score est linéaire en x. Enfin, la classe attribuée à un individu de profil $X=x$ est calculée en maximisant ce score :

$$\hat{y} \quad = \quad \text{argmax}_y d(y|x).$$

En pratique, les scores sont calculés en appliquant la règle de Bayes aux estimations de la matrice de covariance $\hat{\Sigma}$, des moyennes $\hat{\mu}_y$ et des probabilités de base $\hat{p}_y$. On peut montrer que cette règle de classification est la meilleure parmi toutes les règles de classification linéaires, c'est-à-dire qu'elle minimise l'erreur théorique de mauvais classement d'un individu. Cette règle de classification, certes simple, a de bonnes propriétés en pratique.

Afin d'étudier l'optimalité de la règle de Bayes, on s'intéresse au cas simple où le nombre de classes K est égal à 2. Sous cette condition, on considère les règles de classement linéaires de la forme :

$$\log \frac{\mathbb{P}(Y=1|x)}{\mathbb{P}(Y=0|x)} \quad = \quad \beta_0 + \beta' x,$$

où $\beta_0 \in \mathbb{R}$ et β est un vecteur de taille m. Ainsi, la prédiction pour un individu de profil x est :

$$\begin{aligned} \hat{Y} \quad &= \quad 1 \text{ si } \beta_0 + \beta' x > 0, \\ &= \quad 0 \text{ sinon.} \end{aligned}$$

En toute généralité, l'erreur théorique de mauvais classement d'un tel classifieur s'écrit :

$$\begin{aligned} \pi(\beta_0, \beta) \quad &= \quad \mathbb{P}(\hat{Y} \neq Y) \\ &= \quad p_1 \mathbb{P}(\beta_0 + \beta' x < 0 | Y=1) + p_0 \mathbb{P}(\beta_0 + \beta' x > 0 | Y=0). \end{aligned}$$

Compte tenu de l'hypothèse de normalité des covariables, cette erreur s'écrit, en fonction de β et β_0 :

$$\pi(\beta_0, \beta) = p_1 \left[1 - \Phi\left(\frac{\mu_1'\beta + \beta_0}{(\beta'\Sigma\beta)^{1/2}}\right)\right] + p_0\Phi\left(\frac{\mu_0'\beta + \beta_0}{(\beta'\Sigma\beta)^{1/2}}\right),$$

où Φ est la fonction de répartition d'une loi normale centrée réduite. En optimisant cette fonction en β et β_0, on obtient les coefficients β^* et β_0^* minimisant l'erreur théorique de mauvais classement. Ils s'expriment en fonction de μ_0, μ_1 et Σ :

$$\beta^* = \Sigma^{-1}(\mu_1 - \mu_0) \tag{3.3}$$

$$\beta_0^* = \log\frac{p_1}{p_0} - 0.5\left(\mu_1'\Sigma^{-1}\mu_1 - \mu_0'\Sigma^{-1}\mu_0\right). \tag{3.4}$$

Si l'on considère le log ratio des probabilités introduit Expression (3.2), on retrouve bien l'expression des coefficients de la règle de Bayes. Dans ce cas, l'erreur théorique de mauvais classement d'une telle règle de classification est minimale et peut aussi s'écrire :

$$\pi^* = p_1 \left[1 - \Phi\left(\log\frac{p_1}{p_0}\frac{1}{\Delta_\Sigma} + \frac{\Delta_\Sigma}{2}\right)\right] + p_0\Phi\left(\log\frac{p_1}{p_0}\frac{1}{\Delta_\Sigma} - \frac{\Delta_\Sigma}{2}\right)$$

où $\Delta_\Sigma = \sqrt{(\mu_1 - \mu_0)'\Sigma^{-1}(\mu_1 - \mu_0)}$ est la distance de Mahalanobis entre les groupes 0 et 1 pour la métrique Σ. Cette expression permet d'introduire la fonction d'erreur π suivante :

$$\pi(\mu_0, \mu_1, \Sigma) = p_1 \left[1 - \Phi\left(\log\frac{p_1}{p_0}\frac{1}{\Delta} + \frac{\Delta}{2}\right)\right] + p_0\Phi\left(\log\frac{p_1}{p_0}\frac{1}{\Delta} - \frac{\Delta}{2}\right)$$

où $\Delta = \sqrt{(\mu_1 - \mu_0)'\Sigma^{-1}(\mu_1 - \mu_0)}$. C'est l'erreur de classement théorique d'une règle de classification de Bayes construite à partir des paramètres μ_0, μ_1 et Σ. La Figure 3.1 montre les valeurs que prend cette fonction π pour plusieurs distances de Mahalanobis entre deux groupes et pour plusieurs probabilités de base dans les populations. On voit naturellement que plus la distance entre les groupes augmente, plus l'erreur de classement du classifieur de Bayes associée à cette situation est faible, plus il est facile de classer les individus sans erreur. On peut donc se dire qu'on souhaite se ramener à une situation où les parametres (μ_0, μ_1, Σ) sont tels que cette erreur est la plus faible possible. Cette fonction sera utile par la suite pour comparer le classifieur de Bayes et le classifieur de Bayes conditionnel, défini plus loin dans ce chapitre.

L'analyse discriminante quadratique (QDA) étend la LDA au cas où les matrices de covariances diffèrent d'un groupe à l'autre. La QDA ne sera pas détaillée dans ce manuscrit car elle est rarement mise en avant en grande dimension. En effet, le nombre d'individus étant souvent faible, de l'ordre de quelques dizaines, il devient alors difficile d'estimer la matrice de covariance dans chaque groupe.

Analyse discriminante de Fisher Sans hypothèse sur la loi des variables X, l'analyse discriminante de Fisher (Fisher (1936)) peut être vue comme la recherche d'une projection des observations X afin d'atteindre une bonne séparation des

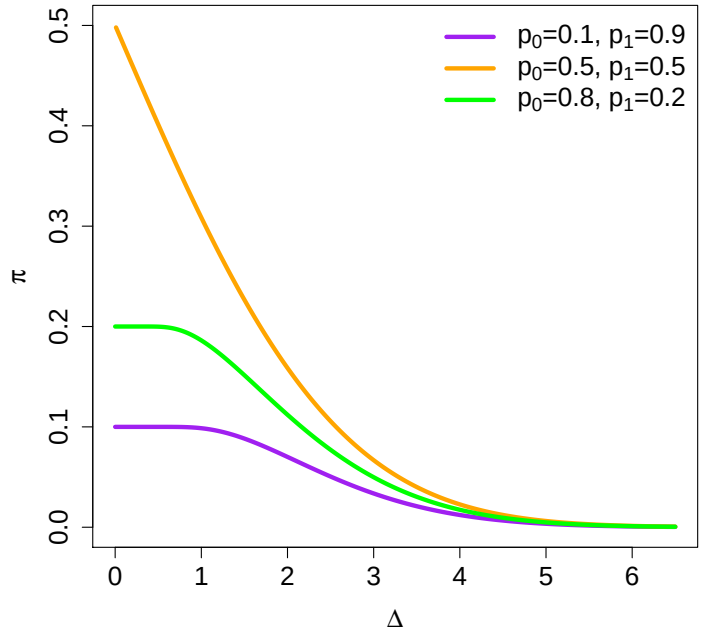


FIGURE 3.1 – Erreur de classement théorique en fonction de la distance de Mahalanobis entre les groupes pour différentes valeurs de p_0 et p_1. Cette fonction est naturellement décroissante en Δ : l'erreur de théorique de mauvais classement s'amenuise au fur et à mesure que la distance séparant deux groupes augmente.

groupes. Si μ désigne la moyenne totale sur tous les groupes, la variance inter-groupes s'écrit :

$$\Sigma_b = \frac{1}{K}\sum_{y=1}^{K}(\mu_y - \mu)(\mu_y - \mu)'.$$

Fisher propose d'étudier la maximisation du ratio de la variance inter-groupes Σ_b sur la variance intra-groupes Σ. On cherche alors des vecteurs discriminants $(\beta_1, \ldots, \beta_{K-1})$ orthogonaux tels que :

$$\max_{\beta_y} \quad \frac{\beta_y' \Sigma_b \beta_y}{\beta_y' \Sigma \beta_y}.$$

Σ_b étant de rang $K-1$, les vecteurs propres de la matrice $\Sigma^{-1}\Sigma_b$ sont solution de cette maximisation (Hardle and Simar (2007)). En pratique, ces vecteurs discriminants sont utiles pour visualiser la séparation des groupes en traçant les nuages de point $(X\beta_1, X\beta_2)$ etc.

Optimal scoring Une troisième formulation de la LDA peut-être faite par analogie avec la régression (Hastie et al. (1994)). La variable catégorielle de groupe Y est transformée en variables quantitatives par des scores. On note $Y^{(d)}$ la matrice $(n \times K)$ de variables design contenant les indicatrices d'appartenance à un groupe : $Y_{iy}^{(d)} = 1$ si l'individu appartient à la classe y et 0 sinon. Le problème de classification revient à estimer les paramètres $(\theta_1, \ldots, \theta_{K-1})$ et $(\beta_1, \ldots, \beta_{K-1})$ tels que :

$$\begin{aligned} \min_{\beta_y, \theta_y} \quad & \|Y^{(d)}\theta_y - X\beta_y\| \\ s.c. \quad & \frac{1}{n}\theta_y' Y^{(d)'} Y^{(d)} \theta_y = 1, \theta_y' Y^{(d)'} Y^{(d)} \theta_{y'} = 0, y' < y. \end{aligned}$$

où $(\theta_1, \ldots, \theta_{K-1})$ sont des K-vecteurs de scores et $(\beta_1, \ldots, \beta_{K-1})$ sont les mêmes vecteurs discriminants que dans l'analyse de Fisher (Clemmensen et al. (2011)).

Finalement, la LDA est simple à implémenter et donne en général de bons résultats en terme d'erreur de classement. Cependant, elle fait appel à l'inverse de la matrice de covariance, difficile à estimer en grande dimension, et n'est pas parcimonieuse. De nombreuses extensions ont donc été proposées pour adapter la LDA au cadre de la grande dimension, dont il sera question dans la Section 2.2 de ce chapitre.

2.2 Analyse linéaire discriminante en grande dimension

Hypothèse d'indépendance Face au problème d'inversion de la matrice de variance covariance des covariables lors de la mise en œuvre d'une LDA, certains auteurs ont simplifié le problème en supposant l'indépendance entre les covariables,

ce qui revient à négliger les corrélations entre les variables. Cette hypothèse aboutit à l'analyse diagonale discriminante ou classifieur naïf de Bayes (Bickel and Levina (2004), Efron (2008)). Cette méthode consiste à remplacer la matrice de covariance dans le score calculé Expression (3.2) par sa diagonale. Pour classer un individu de profil x, il faut calculer les scores associés à chaque classe y :

$$d_{DDA}(y|x) \quad = \quad \log p_y - 0.5\mu_y' D^{-1}\mu_y + x' D^{-1}\mu_y,$$

où $D = diag(\Sigma)$, et choisir la classe qui maximise ce score. Malgré sa simplicité, cette méthode atteint étonnement de bonnes performances de classification en pratique mais elle construit un modèle à partir de toutes les variables, ce qui le rend difficile à interpréter.

Ainsi, certains auteurs proposent des versions pénalisées de l'analyse diagonale discriminante (DDA) afin de réduire la dimension des données aux variables les plus prédictives. Par exemple, les *Nearest Shrunken Centroids* (Tibshirani et al. (2002)) (NSC) annulent la différence de moyenne standardisée pour certaines covariables, lorsque cette différence est inférieure à un certain seuil. La classification se fait par DDA sur le sous ensemble de variables sélectionnées. Le seuil est estimé par validation croisée et par minimisation de l'erreur de classement.

Relâchement de l'hypothèse d'indépendance Certains auteurs proposent de prendre en compte les corrélations dans des LDA pénalisées (voir Hastie et al. (2009)). La LDA est alors revue selon chaque approche : règle de Bayes, *optimal scoring* ou analyse de Fisher, pour tenter de prendre en compte toute la matrice de covariance dans l'analyse, et pas uniquement sa diagonale.

La méthode *Shrunken Centroids Regularized Discriminant Analysis* (SCRDA) proposée par Guo et al. (2007) est présentée comme une amélioration des NSC car elle quitte le contexte de l'analyse diagonale discriminante pour s'inscrire dans le contexte général de l'analyse linéaire discriminante. En effet, comme pour les NSC, les moyennes par groupe sont ramenées à la moyenne globale pour les prédicteurs peu informatifs selon un certain seuil. De plus, la matrice de covariance empirique est biaisée en la remplaçant par une version shrinkée $\widetilde{\Sigma} = \alpha\widehat{\Sigma} + (1-\alpha)\mathbb{I}_m$ dans le calcul des scores individuels et des différences de moyennes standardisées. La matrice $\widetilde{\Sigma}$ permet de contourner le problème d'inversion de la matrice de covariance empirique $\widehat{\Sigma}$ car les valeurs propres de $\widetilde{\Sigma}$ sont toutes positives. Cette technique rend aussi l'estimation de la matrice de covariance plus stable car elle est moins sensible à un petit changement dans les données (Guo et al. (2007)). La méthode SCRDA comprend deux paramètres de régularisation : le seuil sur les moyennes par groupe et le paramètre α de biais sur la matrice de covariance. Ces paramètres sont estimés par une double validation croisée. En pratique, lorsque plusieurs couples de paramètres correspondent à la même erreur de prédiction, Guo et al. (2007) suggèrent de choisir le modèle le plus parcimonieux par la règle du "Min-Min". Cette méthode est implémentée dans le package `R rda` (Guo et al. (2012)).

D'autres auteurs proposent d'étendre la pénalisation ℓ_1 de la méthode Lasso à l'analyse discriminante. Entre autres, la LDA parcimonieuse (SparseLDA) a été proposée par Clemmensen et al. (2011). Cette méthode replace la LDA dans un contexte

de régression : SparseLDA est une pénalisation LASSO de la LDA considérée sous l'approche par *optimal scoring*. Cette méthode est implémentée dans le package `R SparseLDA` (Clemmensen and Kuhn (2012)). Simultanément, Witten and Tibshirani (2011) ont proposé la LDA pénalisée (PenalizedLDA). PenalizedLDA est une pénalisation LASSO sur les vecteurs discriminants β_y de la LDA vue sous l'approche de Fisher. Ces vecteurs parcimonieux sont estimés par un algorithme minimisation-maximisation (Lange (2004), Hunter and Lange (2004)) et le paramètre de régularisation est estimé par validation croisée. Cette méthode est implémentée dans le package `R penalizedLDA` (Witten (2011)).

2.3 Régression logistique

Un des pendants de la régression linéaire pour modéliser une variable qualitative est la régression logistique. Celle-ci ne nécessite pas la normalité des variables mais suppose une relation linéaire directe entre les covariables et le logit des probabilités individuelles a posteriori. Si la classe K est choisie comme classe de référence, alors on suppose l'existence de $K-1$ vecteurs de coefficients de régression tels que :

$$\log \frac{\mathbb{P}(y|x)}{\mathbb{P}(K|x)} = \beta_0^{(y)} + x'\beta^{(y)}.$$

Sans perte de généralité et afin d'alléger les notations, on considère dans ce paragraphe le cas binaire $K = 2$ où Y est codé 0 ou 1. Dans ce cas, le modèle s'écrit :

$$\log \frac{\mathbb{P}(Y=1|x)}{1-\mathbb{P}(Y=1|x)} = \beta_0 + x'\beta.$$

Dans ce cas, la vraisemblance du modèle s'écrit :

$$\mathcal{L}(\beta_0, \beta) = \prod_{i=1}^{n} \mathbb{P}(Y_i=1|x)^{Y_i}(1-\mathbb{P}(Y_i=1|x))^{1-Y_i}.$$

Ainsi, la déviance du modèle à minimiser s'écrit :

$$\mathcal{D}(\beta_0, \beta) = \sum_{i=1}^{n}((1-Y_i)(\beta_0 + x'\beta) - \log(1+e^{\beta_0+x'\beta})$$

Dans le cas de la régression logistique multinomiale, on obtient $K-1$ équations de déviance à minimiser. L'estimation des paramètres de régression se fait soit par maximum de vraisemblance via l'algorithme de Newton-Raphson, car il n'existe pas de solution analytique pour ces estimateurs, soit par moindres carrés pondérés si l'on suppose que la vraisemblance appartient à la famille exponentielle (Hilbe (2009)). Dans les deux cas, ce problème d'optimisation nécessite le calcul de l'inverse de la matrice de covariance empirique.

En théorie, sous l'hypothèse de normalité des covariables, lorsque le nombre d'observations tend vers l'infini et $K = 2$, la LDA et la régression logistique tendent vers le même estimateur, qui est l'estimateur du maximum de vraisemblance des coefficients β et β_0. En pratique, la LDA donne de bons résultats en terme d'erreur de classement et est recommandée pour sa simplicité d'implémentation. La régression logistique est connue pour être plus robuste à un écart à la loi normale.

2.4 Régression logistique pénalisée

La régression logistique hérite des méthodes de sélection de variables de la régression linéaire. Les régressions Ridge (Hoerl and Kennard (1970)), Lasso (Tibshirani (1996)) (aussi connu sous le nom de *basis pursuit* en traitement du signal) et Elastic-net (Zou and Hastie (2005)), fondées sur des pénalisations du critère des moindres carrés ou de la vraisemblance du modèle, sont des méthodes de référence dans ce domaine.

La régression Ridge propose une modification des moindres carrés de façon à ce que la matrice de covariance soit inversible en remplaçant la matrice de covariance empirique $X'X$ par $(X'X+\lambda\mathbb{I})$. Cette formulation est équivalente à une pénalisation ℓ_2 de la déviance du modèle :

$$(\hat{\beta}_0, \hat{\beta}) \quad = \quad arg \min_{\beta_0,\beta} \left(\mathcal{D}(\beta_0, \beta) + \lambda \|\beta\|_2^2 \right).$$

Cette méthode est formulée pour parer au problème des covariables corrélées. En pratique, cette méthode dite de *shrinkage* des coefficients aboutit à des modèles peu parcimonieux.

La méthode Lasso a été introduite pour proposer des solutions parcimonieuses aux problèmes de régression en grande dimension. Cette méthode est une pénalisation par la norme ℓ_1 de la déviance :

$$(\hat{\beta}_0, \hat{\beta}) \quad = \quad arg \min_{\beta_0,\beta} \left(\mathcal{D}(\beta_0, \beta) + \lambda \|\beta\|_1 \right).$$

L'inconvénient du Lasso est qu'il peut sélectionner au maximum $\min(n, m)$ variables. En effet, Efron et al. (2004) montre que si $\#\hat{S}_{\lambda_{CV}}$ désigne le cardinal de l'ensemble des variables sélectionnées par le Lasso lorsque le paramètre λ est choisi par validation croisée, alors $\#\hat{S}_{\lambda_{CV}} \leq \min(n, m)$. Ceci est une contrainte forte en grande dimension où le nombre d'individus peut-être très faible (quelques dizaines) au regard du nombre de variables (quelques milliers) : on risque alors d'obtenir un modèle très parcimonieux et de manquer des variables prédictives.

La méthode Elastic-net (Zou and Hastie (2005)) est un compromis entre les deux méthodes ci-dessus en introduisant les deux pénalités de la déviance :

$$(\hat{\beta}_0, \hat{\beta}) \quad = \quad arg \min_{\beta_0,\beta} \left(\mathcal{D}(\beta_0, \beta) + \lambda \|\beta\|_1 + \gamma \|\beta\|_2^2 \right).$$

La régression Elastic-net est présentée comme plus robuste à la dépendance. Elle possède aussi l'avantage de pouvoir sélectionner plus que $\min(n, m)$ variables contrairement au Lasso.

L'idée principale de ces méthodes de régularisation est de jouer sur le compromis biais-variance afin de réduire l'erreur quadratique moyenne (EQM) du modèle. Ces régressions sont dites *biaisées* mais la variance des estimateurs est plus faible que celle de l'estimateur des moindres carrés. Le paramètre de régularisation contrôle la complexité du modèle : quand le paramètre de régularisation $\lambda \to \infty$ (et $\gamma \to \infty$

pour Elastic-net) l'estimation de β est de plus en plus parcimonieuse et tend vers le vecteur nul. Inversement, si $\lambda \to 0$ (et $\gamma \to 0$ pour Elastic-net) alors la solution tend vers l'estimateur des moindres carrés. En pratique, les hyper-paramètres λ et α sont estimés par validation croisée de façon à minimiser la déviance du modèle ou l'erreur de mauvais classement. Dans le cadre du modèle linéaire généralisé, les paramètres de régression sont estimés par un algorithme de type *cyclic coordinate descent*. Ces méthodes sont entre autres implémentées dans le package `R glmnet` (Friedman et al. (2010)).

Le Lasso est connu dans la littérature pour être instable en cas de dépendance. Dans la suite de ce chapitre, on illustre que le Lasso est très instable en cas de dépendance : cela se traduit par une faible reproductibilité des résultats. Les méthodes Bolasso (Bach (2008)) et *stability selection* (Meinshausen and Bühlmann (2010)) ont donc été proposées pour améliorer la stabilité du Lasso. Elles sont toutes les deux fondées sur du ré-échantillonnage des individus. L'ensemble des variables sélectionnées est celui des variables ayant été sélectionnées un grand nombre de fois dans la procédure. Ces méthodes sont accompagnées de bonnes propriétés théoriques. Par exemple, Bach (2008) montre que, sous certaines hypothèses sur la fonction génératrice des cumulants, sous l'hypothèse de parcimonie du modèle et sous réserve d'invisibilité de la matrice de covariance des covariables, la probabilité que la méthode Bolasso sélectionne les variables prédictives tend vers 1. Ce résultat peut expliquer qu'en pratique, ces méthodes sont très conservatives et aboutissent à des modèles extrêmement parcimonieux.

2.5 Autres approches

Bien d'autres méthodes ont été proposées pour prendre en compte la dépendance dans les problèmes de régression. La régression sur composantes principales et la régression PLS (Partial Least Squares) par exemple cherchent des variables intermédiaires, combinaisons linéaires des variables initiales et orthogonales les unes aux autres sur lesquelles effectuer la régression.

Enfin, d'autres méthodes issues du domaine du *machine learning* telles que le boosting (Freund and Schapire (1996),Freund and Schapire (1997)), où l'on génère et agrège un grand nombre de modèles selon des pondérations différentes des données ou les forêts aléatoires (Breiman (2001)) sont connues pour leur efficacité. Le fonctionnement de ces méthodes n'est pas détaillé ici car elles ne font pas partie des comparaisons réalisées dans ce chapitre.

2.6 Cadre théorique

Dans ce chapitre, on choisit de se placer dans un contexte d'analyse linéaire discriminante selon l'approche de la règle de Bayes. On suppose donc que les covariables suivent une loi normale conditionnellement au groupe des individus. On reprend donc les hypothèses du modèle (3.1) introduit en Section 2.1 :

$$X|Y = y \quad \sim \quad \mathcal{N}_m(\mu_y, \Sigma),$$

où $\Sigma \neq \mathbb{I}_m$ représente une situation de forte dépendance. Le reste de ce chapitre est consacré à l'illustration de l'impact de la dépendance sur le lasso puis la méthode proposée pour prendre en compte la dépendance dans ce contexte d'analyse linéaire discriminante est détaillée. Enfin, les propriétés de cette méthode sont illustrées sur des données de méthylation de l'ADN et sur une étude par simulations.

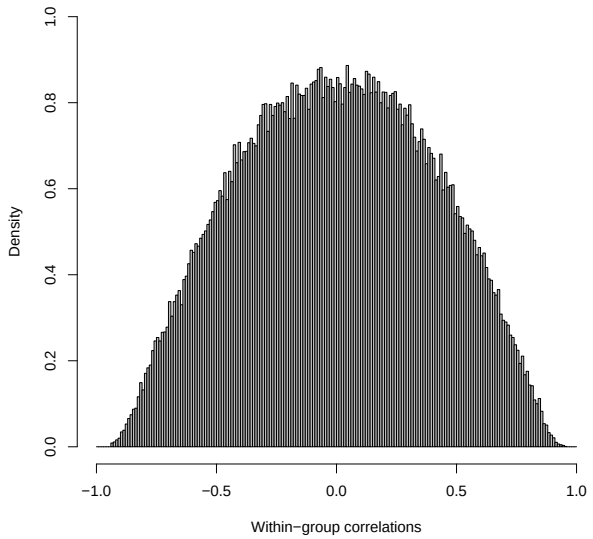
3 Impact de la dépendance

On propose dans cette section d'étudier sur des simulations la stabilité d'une procédure standard de sélection de variables : la procédure lasso (Tibshirani (1996)), en comparant le cas d'indépendance à un cas de dépendance. On considère deux classes, de $n_0 = n_1 = 30$ observations chacune. Les profils individuels sont de dimension 500. La moyenne μ_0 du groupe 0 est nulle. Toutes les composantes de μ_1, la moyenne du groupe 1, sont nulles, exceptées les 100 dernières valant $\delta = 0.74$. Cette valeur δ est la plus petite différence détectable par un test de Student comparant 2 groupes de 30 observations, de niveau 5%, de puissance 80% et d'écart-type 1. La matrice de corrélations Σ est générée selon un modèle à 5 facteurs afin de représenter un niveau de dépendance élevé ($trace(BB')/trace(\Sigma) = 0.75$). L'histogramme présenté en Figure 3.2(a) montre en effet de fortes corrélations entre les covariables. A partir de ces paramètres, on peut calculer les coefficients de régression $\beta = \Sigma^{-1}(\mu_1 - \mu_0)$, présentés en Figure 3.2(b). On peut définir un ordre entre les covariables à partir de $|\beta|$: la variable ayant le plus grand coefficient en valeur absolue aura le rang 1 etc. En effet, on peut s'attendre à ce que les variables ayant un coefficient élevé en valeur absolue soit plus souvent sélectionnée qu'une variable avec un faible signal.

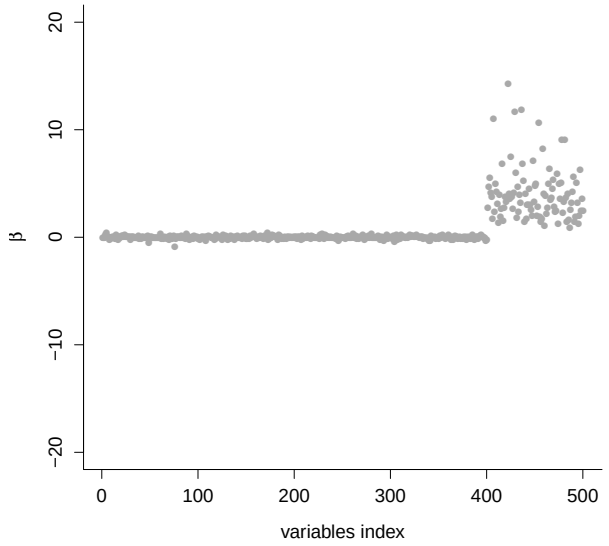
Pour le cas d'indépendance, la matrice de corrélation Σ est l'identité. Afin de conserver des situations comparables, les coefficients β sont les mêmes que dans le cas de dépendance. Ainsi, μ_0 est toujours le vecteur nul et $\mu_1 = \beta$.

La méthode Lasso est appliquée aux 1000 jeux de données simulés selon chaque scenario. La Figure 3.3 présente les histogrammes du nombre de variables sélectionnées. De plus, on calcule, à partir de $|\beta|$, la précision moyenne d'un sous-ensemble sélectionné en calculant le rang moyen des variables de ce sous-ensemble. Les histogrammes de la Figure 3.4 présentent la distribution de cette précision moyenne.

Le premier impact remarquable de la dépendance est sur le nombre de variables sélectionnées : celui-ci est plus grand lorsque les variables sont corrélées. De plus, le signal étant assez élevé dans ces simulations, on remarque qu'aucune variable non prédictive n'a été sélectionnée à tort lorsque les variables sont indépendantes (le taux de fausses découvertes est nul dans 100% des simulations). En revanche, lorsque les variables sont corrélées, le taux de fausses découvertes est non nul dans 12.1% des simulations. La dépendance altère aussi la précision de la sélection : lorsque les variables sont indépendantes, le Lasso sélectionne les variables les plus prédictives, parmi les 20 plus prédictives. En situation de dépendance, le Lasso a tendance à sélectionner des variables de rang 40 à 80, c'est-à-dire des variables ayant un coefficient plus faible. Ceci est cohérent avec le nombre de variables sélectionnées :



(a) Histogramme des corrélations - cas de dépendance



(b) Coefficients du modèle

FIGURE 3.2 – Plan de simulation. La figure du haut montre une distribution des corrélations caractéristique d'une situation de dépendance : la distribution est très étalée et présente une importante proportion de corrélations élevées, positives ou négatives. La figure du bas montre que les 400 premières variables ont un pouvoir discriminant nul. Pour les 100 dernières variables, la structure de dépendance entraîne un pouvoir discriminant différent d'une variable à l'autre, même si les différences de moyennes entre les deux groupes sont égales.

le Lasso rattrape en quelques sortes le manque de pouvoir prédictif des variables sélectionnées par des sous-ensembles sélectionnés plus grands.

4 Modèle à facteurs pour la sélection de variables

La stratégie proposée dans la suite de ce chapitre consiste à appliquer les méthodes classiques de sélection de variables aux données ajustées sur l'effet des facteurs latents. On suppose donc un modèle à facteurs pour la matrice de covariance Σ. Déjà détaillé dans l'introduction de ce manuscrit, on rappelle que sous l'hypothèse d'un modèle à facteurs pour les covariables, le modèle (3.1) devient :

$$X = \mu_y + BZ + \varepsilon; \text{ avec } y = 1 \text{ si } Y = 1 \text{ et } y = 0 \text{ sinon} \quad (3.5)$$

et que ce modèle est équivalent à la décomposition suivante pour la matrice de covariance Σ :

$$\Sigma = \Psi + BB'.$$

4.1 Définition et intérêt du classifieur de Bayes conditionnel

Sous ce modèle, on remarque que la distribution jointe des covariables et des facteurs sachant le groupe Y est une loi normale multivariée telle que :

$$\begin{pmatrix} X \\ Z \end{pmatrix} \sim \mathcal{N}\left[\begin{pmatrix} \mu_y \\ 0 \end{pmatrix}, \begin{pmatrix} \Sigma & B \\ B' & \mathbb{I}_q \end{pmatrix}\right].$$

On définit alors le classifieur de Bayes conditionnel, optimal sachant les covariables et les facteurs, à partir de l'inverse par bloc de la matrice de covariance de l'Equation (3.6) :

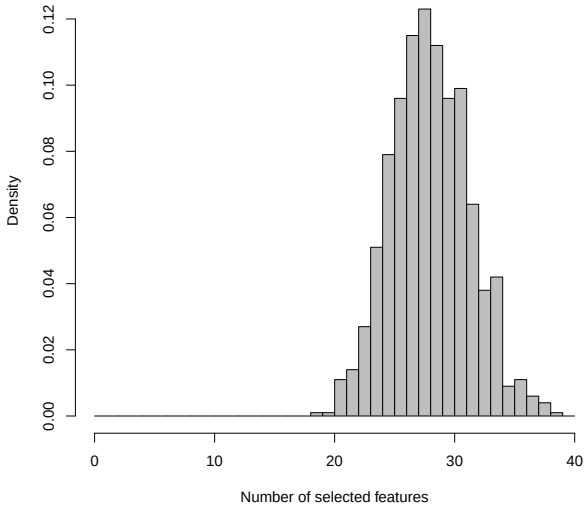
$$LR(x,z) = \log\frac{p_1}{p_0} - \frac{1}{2}\left(\mu_1'\Psi^{-1}\mu_1 - \mu_0'\Psi^{-1}\mu_0\right) + (x - Bz)'\Psi^{-1}(\mu_1 - \mu_0). \quad (3.6)$$

On remarque que la dépendance en x et z dans l'Equation (3.6) se fait par les variables ajustées sur l'effet des facteurs $x - Bz$. Ceci confirme que, si la structure en facteurs est connue, le meilleur classifieur linéaire est la règle de Bayes usuelle appliquée aux profils facteurs-ajustés.

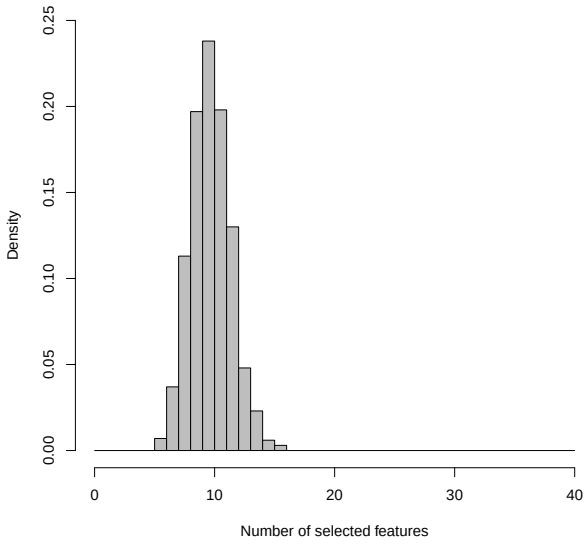
La probabilité de mauvais classement pour cette règle de classification est minimale et est notée $\pi_z^* = \pi(\mu_0, \mu_1, \Psi)$ où la fonction π est définie en (3.5). Si on définit les *loadings* standardisés par $B^* = \Psi^{-1/2}B$, on obtient l'inégalité suivante :

$$\frac{1}{1+\rho^2_{\max}} \leq \frac{\Delta^2_\Sigma}{\Delta^2_\Psi} \leq 1,$$

où $\Delta^2_\Psi = (\mu_1 - \mu_0)'\Psi^{-1}(\mu_1 - \mu_0)$ désigne la distance de Mahalanobis entre les deux groupes pour la matrice de covariance Ψ et $\rho_{\max}$ est la plus grande valeur singulière

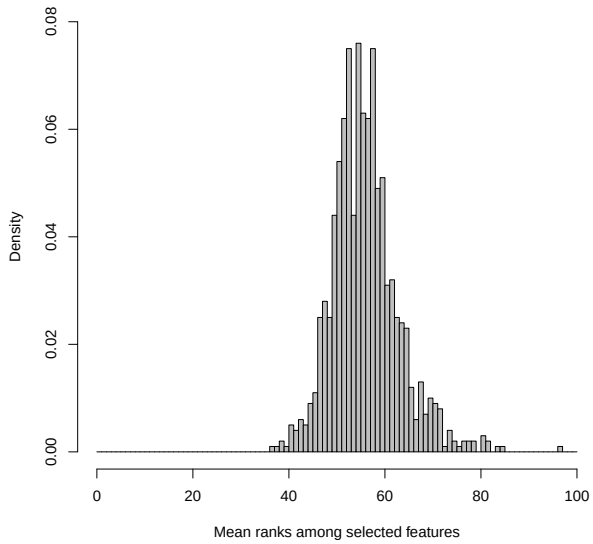


(a) Dépendance

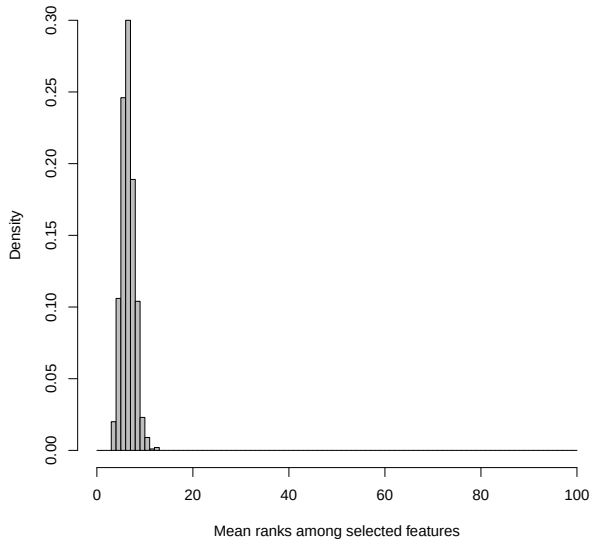


(b) Indépendance

FIGURE 3.3 – Distribution du nombre de variables sélectionnées par le lasso dans le cas dépendant et indépendant. On remarque que le lasso a tendance à sélectionner plus de variables dans le cas dépendant et que le nombre de variables sélectionnées est plus variable.



(a) Dépendance



(b) Indépendance

FIGURE 3.4 – Rang moyen des sous-ensembles sélectionnés par le lasso dans le cas dépendant et indépendant. On remarque que le lasso sélectionne des variables dont le pouvoir discriminant est faible, avec un rang compris entre 40 et 80 et ne sélectionne jamais les variables les plus prédictives en situation de dépendance. Lorsque les variables sont indépendantes, le lasso sélectionne bien les variables les plus prédictives, dont le rang est inférieur à 20.

de B^*. Comme la fonction π décroit lorsque la distance entre les groupes augmente, on peut déduire que $\pi_z^* \leq \pi^*$. De plus, le terme de gauche dans cette inégalité montre que le gain attendu par l'approche conditionnelle augmente avec $\rho^2_{\max}$, qui est aussi la plus grande valeur propre de la matrice $B'\Psi^{-1}B$. Ainsi, le gain attendu est plus important en situation de forte dépendance, c'est-à-dire lorsque les *loadings* prennent les valeurs élevées par rapport à la variance spécifique.

L'optimalité de la règle de classification de Bayes standard, établie sans hypothèse sur la matrice de covariance Σ, n'est pas remise en cause. Cependant, sous l'hypothèse d'une structure en facteurs pour Σ, le résultat ci-dessus établit la supériorité théorique d'une approche fondée sur des covariables ajustées sur les facteurs $x - Bz$. La section suivante présente un algorithme d'estimation des paramètres du modèle (3.5).

4.2 Algorithme d'estimation du modèle à facteurs

L'algorithme proposé alterne l'estimation de μ_0, μ_1, B et Ψ et le calcul des facteurs latents Z.

Initialisation L'algorithme est initialisé par les moyennes empiriques par groupe $\hat{\mu_0} = \bar{x_0}$ et $\hat{\mu_1} = \bar{x_1}$. A partir de ces estimations, les données sont centrées par groupe. Ces profils centrés $x - \hat{\mu_y}$ sont utilisés pour estimer B et Ψ avec l'algorithme EM détaillé dans Friguet et al. (2009). Les estimateurs associés sont notés $\hat{B}$ et $\hat{\Psi}$.

Etape 1 : extraction des facteurs Sous le modèle à facteurs, la méthode de Thompson permet de calculer les facteurs latents. On propose de l'adapter au cas de la classification supervisée. En effet, on déduit de la loi jointe multivariée des covariables et des facteurs latents que l'espérance conditionnelle des facteurs sachant le profil x s'écrit :

$$\mathbb{E}_x(Z) \quad = \quad (\mathbb{I}_q + B'\Psi^{-1}B)^{-1}B'\Psi^{-1}\left(x - [\mu_0\mathbb{P}_x(Y=0) + \mu_1\mathbb{P}_x(Y=1)]\right) \quad (3.7)$$

où

$$\mathbb{P}_x(Y=1) \quad = \quad 1 - \mathbb{P}_x(Y=0) = \frac{1}{1+\exp(-\beta_0 + \beta^{*\prime}x)}.$$

Remarques

1. Les estimateurs des coefficients de régression $(\hat{\beta}_0, \hat{\beta})$ peuvent être calculés explicitement en estimant les paramètres des l'Expressions (3.3) et (3.4). En effet, d'après la formule de Woodbury, l'inverse de Σ fait appel à l'inversion d'une matrice de petite dimension $(q \times q)$:

$$\Sigma^{-1} \quad = \quad \Psi^{-1} - \Psi^{-1}B(\mathbb{I}_q + B'\Psi^{-1}B)^{-1}B'\Psi^{-1}$$

2. A partir des estimations $(\hat{\beta}_0, \hat{\beta})$, on peut déduire les probabilités individuelles $\mathbb{P}_x(Y=1)$. Cependant, ces valeurs sont affectées en pratiques par le surajustement ce qui pénalise les performances de classification de la méthode.

Afin de réduire l'impact du sur-ajustement, on propose donc d'estimer ces probabilités par des méthodes parcimonieuses telle que la régression logistique régularisée par une pénalisation ℓ_1.

On peut ensuite calculer des facteurs latents par un estimateur plug-in de l'Expression (3.7) en remplaçant les paramètres par leurs estimations $\hat{\mu}_0$, $\hat{\mu}_1$, $\hat{B}$ et $\hat{\Psi}$:

$$\hat{Z} \quad = \quad (\mathbb{I}_q + \hat{B}'\hat{\Psi}^{-1}\hat{B})^{-1}\hat{B}'\hat{\Psi}^{-1}\left(x - [\hat{\mu}_0\hat{\mathbb{P}}_x(Y=0) + \hat{\mu}_1\hat{\mathbb{P}}_x(Y=1)]\right)$$

Etape 2 : estimation des paramètres L'actualisation des estimations des moyennes par groupe se fait par ajustement du modèle (3.5) et estimation par la méthode des moindres carrés, où les facteurs Z sont remplacés par leur estimation $\hat{Z}$. On peut ainsi actualiser les profils centrés $(x - \hat{\mu}_y)$ pour mettre à jour les estimateurs des paramètres de covariance $\hat{B}$ et $\hat{\Psi}$.

Itération et critère d'arrêt Les étapes 1 et sont 2 sont réitérées, alternant mise à jour des facteurs latents et actualisation du modèle à facteurs. L'algorithme s'arrête lorsque deux estimations successives des moyennes sont suffisamment proches.

Les données ajustées sur les facteurs (ou décorrélées) sont définies par $x - \hat{B}\hat{Z}$. La stratégie proposée consiste à définir la "version facteur-ajustée" d'une méthode de classification par le fait d'appliquer cette méthode sur les données décorrélées.

Dans la suite, on illustre sur des données réelles et sur des simulations que cette nouvelle méthode de décorrélation des données améliore la sélection de variables en terme d'erreur de classification et de reproductibilité de l'ensemble des variables sélectionnées.

5 Illustration sur des données réelles

Lorsque l'on analyse des données réelles, il est impossible de savoir quelles variables ont été sélectionnées à tort ou non. Néanmoins, on tente dans cette section d'illustrer l'instabilité du Lasso sur des données réelles en étudiant la répétabilité des variables sélectionnées et l'erreur de classification calculée par validation croisée. Ensuite, l'apport de la méthode proposée sera étudié sur des données de méthylation de l'ADN.

5.1 Stabilité de la sélection de variables

5.1.1 Données

Les données détaillées par Hedenfalk et al. (2001), issues de puces à ADN sont souvent utilisées dans la littérature pour étudier des procédures statistiques en grande dimension. Il s'agit de données de cancer du sein téléchargeables à l'adresse `http://research.nhgri.nih.gov/microarray/NEJM_Supplement/`. Les individus ont été initialement répartis en trois groupes : BRCA1(7 observations),

TABLE 3.1 – Sélection sur les données complètes

Données	Variables	Erreur de prédiction
Brutes	11	0.400
Décorrélées	8	0.267

BRCA2 (8 observations) et Sporadic (6 observations) et les expressions de 3226 gènes sont mesurées. Un individu étant classé dans 2 groupes à la fois, celui-ci a été supprimé. 196 gènes présentant des valeurs extrêmes (au dessus de 10 ou inférieur à 0.1) ont été supprimés et les données sont passées au $\log_2$. Finalement, le problème revient à sélectionner parmi 3030 gènes les variables discriminant au mieux les groupes BRCA1 et BRCA2 à partir de 15 observations.

5.1.2 MÉTHODES

Le but de cette étude est de comparer les résultats obtenus par le Lasso sur les données brutes à ceux obtenus par le Lasso appliqué aux données ajustées sur l'effet des facteurs. Bien que le Lasso soit connu pour être peu robuste à la corrélation (Bach (2008)), cet exemple permet d'illustrer l'apport de la décorrélation en terme de stabilité des résultats. D'abord, le Lasso est appliqué sur le jeu de données complet et l'erreur de classification est estimée par validation croisée. Afin d'étudier la reproductibilité des résultats, la même procédure est appliquée sur 15 jeux de données incomplets (14 observations), où une observation a été successivement supprimée. Ceci permet d'étudier la stabilité des résultats à la suppression d'une observation. On compare ainsi l'ensemble des variables sélectionnées sur le jeu de données complet à chaque ensemble sélectionné sur les données incomplètes.

Ensuite, la même procédure est appliquée sur les données ajustées sur l'effet des facteurs selon la méthode présentée en Section 4. Le nombre de facteurs est estimé à 1 sur les données complètes et incomplètes. Sur chaque jeu de données incomplet, un nouveau modèle à facteurs est estimé. Enfin, la taille de l'échantillon étant très faible, le paramètre de régularisation est estimé par Leave-One-Out.

5.1.3 RÉSULTATS SUR DONNÉES COMPLÈTES

Les résultats de la procédure de sélection appliquée aux données complètes (brutes et ajustées sur les facteurs) sont reportés Table 3.1. En appliquant le Lasso aux données décorrélées, on observe un plus petit sous-ensemble de variables sélectionnées et une erreur de classification plus faible. Dans la suite, on notera I_{raw} (respectivement I_{FA}) ce sous-ensemble de variables sélectionnées par le Lasso appliqué aux données brutes (resp. ajustées sur l'effet des facteurs).

5.1.4 RÉSULTATS SUR DONNÉES INCOMPLÈTES

La procédure Lasso est donc appliquée sur 15 jeux de données incomplets, en supprimant un à un chaque observation des données initiales. La Table 3.2(a) (resp. 3.2(b)) présente le nombre de variables sélectionnées, le nombre et la proportion de

TABLE 3.2 – Sélection sur les données incomplètes

(a) Données brutes

Obs. supprimée	1	2	3	4	...	12	13	14	15	Moy.	Ec.type
Variables	1	10	7	8	...	6	12	6	6	6.4	(3.6)
Inclus (N)	1	9	3	6	...	6	6	3	5	4.2	(2.5)
Inclus (%)	9.1	81.8	27.3	54.5	...	54.5	54.5	27.3	45.5	38.2	(22.3)
Erreur	0.571	0.214	0.286	0.214	...	0.214	0.357	0.214	0.357	0.300	(0.138)

(b) Données décorrélées

Obs. supprimée	1	2	3	4	...	12	13	14	15	Moy.	Ec.type
Variables	9	7	9	10	...	9	7	12	8	7.9	(2.5)
Inclus (N)	5	7	6	8	...	7	6	8	7	5.7	(2)
Inclus (%)	62.5	87.5	75.0	100.0	...	87.5	75.0	100.0	87.5	70.8	(24.9)
Erreur	0.357	0.214	0.286	0.071	...	0.357	0.357	0.214	0.214	0.229	(0.115)

Note : "Variables" représente le nombre de variables sélectionnées / "Inclus (N)" représente le nombre de variables sélectionnées appartiennant à I_{raw} ou I_{FA}, c'est le nombre de variables dites "stables" / "Inclus (%)" est la proportion de variables stables / "Erreur" représente l'erreur de prédiction calculée par validation croisée

variables appartenant à I_{raw} (resp. I_{FA}) et l'erreur de prédiction pour chaque sous-jeu de données. La table présente les résultats après suppression des 4 premières et 4 dernières observations ainsi que les moyenne (Moy.) et écart-type (Ec.type). Pour éviter de charger le tableau, les résultats pour toutes les observations ne sont pas présentés.

La Table 3.2(a) présente des situations très variées, en fonction de l'observation qui est supprimée. Certaines observations ont une influence forte sur la stabilité des résultats. Par exemple, le Lasso semble très sensible au retrait de l'observation 1, car une seule variable est sélectionnée au lieu de 11 sur le jeu de données complet. Parmi les 6 variables sélectionnées après le retrait de l'observation 14, seulement 3 appartiennent à I_{raw}. Ce phénomène est moins marqué lorsque le Lasso est appliqué aux données facteurs-ajustées (Table 3.2(b)). On remarque qu'en moyenne, la proportion de variables appartenant à I_{FA} est plus élevée (38.2% contre 70.8%) et l'erreur de prédiction moyenne est plus faible (0.300 contre 0.229).

5.1.5 CONCLUSION

Cet exemple permet d'illustrer que la stabilité des résultats d'une procédure classique de sélection de variables est vraisemblablement affectée par la dépendance. Un faible changement dans les données, comme la suppression d'un individu, introduit une grande variabilité dans les performances de classification et mène à des ensembles de variables sélectionnées très différents. L'ajustement sur des facteurs latents aide à atténuer les effets dû à l'hétérogénéité des données et améliore la stabilité de la sélection mais aussi l'erreur de prédiction.

5.2 ETUDE DE DONNÉES DE MÉTHYLATION DE L'ADN

Les données sur la méthylation de l'ADN intéressent les biologistes car elles permettent de mettre en évidence de nouveaux processus biologiques. Ces données sont caractérisées par une très forte hétérogénéité et il est intéressant d'étudier

l'apport de l'ajustement sur les facteurs sur ce type de données (Houseman et al. (2015)).

5.2.1 Données

Cancer de l'estomac Les données sont issues d'une étude sur le cancer de l'estomac. Les données ont été initialement publiées par Zouridis et al. (2012) et contiennent 27 578 mesures de méthylation de l'ADN et 297 observations. 2 573 variables sont supprimées pour cause de données manquantes ce qui laisse 25 005 variables. La variable réponse est binaire et on compte 203 cas (cancer gastrique) et 94 échantillons de tissus de l'estomac non malades.

Carcinome Les données sont issues d'une étude sur le carcinome cellulaire squameux de la tête et du cou (Langevin et al. (2012) et Houseman et al. (2015)) et contiennent initialement 27 578 mesures de méthylation de l'ADN dans le sang de 92 individus atteints de carcinome et 92 témoins. La variable réponse est binaire : cas/contrôle. Après suppression des données manquantes, le jeu de données contient 26 482 colonnes.

5.2.2 Méthodes

D'après l'étude par simulations de la section suivante, SDA semble être la méthode issue de la littérature la plus performante en terme d'erreur de prédiction et de précision de sélection. Dans un premier temps, la méthode SDA est appliquée au jeu de données complet. Les résultats sont comparés à ceux obtenus par SDA après décorrélation par l'algorithme proposé. L'erreur de prédiction est estimée par validation croisée avec 10 folds et 20 répétitions. Ainsi, le modèle est estimé sur 200 sous-échantillons.

5.2.3 Résultats

Cancer de l'estomac Le critère d'inflation de la variance (Friguet et al. (2009)) suggère d'extraire 10 facteurs. La Table 3.3(a) contient l'erreur de prédiction et le nombre de variables sélectionnées par les deux procédures comparées.

Carcinome Le critère d'inflation de la variance (Friguet et al. (2009)) suggère d'extraire 8 facteurs. La Table 3.3(b) présente les résultats obtenus par les deux méthodes. On remarque que l'ajustement sur les facteurs latents mène à une erreur de classification équivalente à celle atteinte par SDA mais pour un plus petit nombre de variables sélectionnées.

6 Simulations

Afin d'étudier l'apport de l'ajustement sur les facteurs en sélection de variables et en classification, on réalise une étude par simulations. Plusieurs méthodes de classification tirées de la littérature sont appliquées sur des données simulées selon plusieurs scenario de dépendance sont considérés : indépendance, dépendance en

TABLE 3.3 – Nombre de variables sélectionnées et taux d'erreurs pour le jeu de données sur le cancer de l'estomac et le carcinome.

(a) Cancer de l'estomac

Méthode	Nombre de variables	Erreur de prédiction
SDA	2638	0.0301
Factor-adjusted SDA	305	0.0217

(b) Carcinome

Méthode	Nombre de variables	Erreur de prédiction
SDA	2915	0.2719
Factor-adjusted SDA	619	0.2626

blocs, structure en facteurs et matrice de Toeplitz, à la manière de Meinshausen and Bühlmann (2006)). Les performances des méthodes originales sont comparées à celles des méthodes appliquées sur les données décorrélées.

6.1 Plan de simulations

Les jeux de données sont simulés selon une loi normale multivariée. Chaque jeu de données contient $m = 1\ 000$ variables et $n = 30$ observations. On considère une variable réponse binaire Y telle que les jeux de données soient séparés en deux groupes de taille $n_0 = n_1 = n/2$. Les m-profils individuels X sont distribués selon une loi normale de moyenne $\mu_0 = 0_m$ dans le premier groupe ($Y = 0$), où $0_m \in \mathbb{R}^m$ est le vecteur nul. La moyenne dans le second groupe ($Y = 1$) est μ_1. Un sous-ensemble I de 50 variables prédictives est tiré au hasard. Concernant ces variables, le vecteur μ_1 possède des composantes non nulles : $\mu_{1j} = \delta$ si $j \in I$ et $\mu_{1j} = 0$ sinon. La valeur de δ est 0.55 ou 0.47, ce qui correspond à un signal fort ou modéré (d'après la définition de la force d'un signal introduite par Donoho and Jin (2008)). 1 000 jeux de données sont simulés selon chaque structure de covariance Σ décrit ci-dessous :

(A) Les m variables sont indépendantes et distribuées selon une loi normale de variance 1. La matrice de covariance est alors la matrice identité $\mathbb{I}_m$. Cette structure de dépendance est utilisée comme un cas "contrôle" et permet de vérifier que la méthode proposée ne détecte pas à tort de la corrélation ;

(B) Σ est une matrice composée de deux blocs. La corrélation entre les 100 premières variables est égale à 0.7 et la corrélation entre les 900 dernières est 0.3. Cette matrice de covariance est utilisée pour étudier l'impact de la dépendance sur les procédures de tests multiples dans le contexte de l'analyse d'expression de gènes par Zuber and Strimmer (2009).

(C) Σ se décompose en une partie de variance spécifique et de variance commune,

comme dans le modèle à facteurs : $\Sigma = \Psi + BB'$. Ψ est une matrice diagonale de variance spécifique et B est une matrice $m \times q$ générée de façon à ce que la proportion de dépendance partagée par les covariables $trace(BB')/trace(\Sigma)$ soit élevée (78%). Dans ces simulations, le nombre de facteurs est $q = 5$. Ce cas est favorable car la matrice de covariance est générée selon un modèle à facteurs. On impose donc un signal plus faible $\delta = 0.47$.

(D) Σ est une matrice de Toeplitz. Cette structure de dépendance temporelle correspond à la matrice de covariance d'un processus auto-régressif tel que la covariance entre une variable i et une variable j soit $\sigma\rho^{|i-j|}$. Dans ces simulations, $\sigma = 1$ et $\rho = 0.99$.

6.2 Méthodes

4 procédures de classification sont appliquées à chaque jeu de données simulé. Ces méthodes sont présentées plus en détails dans l'introduction de ce manuscrit Chapitre 1.

(LASSO) régression logistique régularisée par une pénalisation ℓ_1 implémentée dans le package `R glmnet` (Friedman et al. (2010))

(SLDA) Sparse Linear Discriminant Analysis, qui est une LDA régularisée par une pénalisation ℓ_1 (Clemmensen et al. (2011)) implémentée dans le package `R sparseLDA`. Le nombre de variables à inclure dans le modèle est arbitrairement fixé à 10 dans ces simulations.

(SDA) Shrinkage Discriminant Analysis, qui est une LDA shrinkée par des estimations de type James-Stein des paramètres, implémentée dans le package `R sda` (Ahdesmäki and Strimmer (2010)). On peut toutefois noter que SDA consiste finalement en un ajustement sur la corrélation des scores utilisés pour la sélection de variables dans la DDA.

(DDA) Shrinkage Diagonal Discriminant Analysis, qui suppose l'indépendance entre les données, implémentée dans le package `R sda`. L'estimation du modèle de DDA se fait par une pénalisation de type ridge.

Plusieurs seuils sont implémentés dans le package `R sda` pour réaliser la DDA et SDA tels que le contrôle du FNDR (False Non Discovery Rate) ou le Higher Criticism (Donoho and Jin (2008)). Ces deux méthodes donnent des résultats comparables et les résultats présentés ici ne concernent que la sélection par contrôle du FNDR.

Chaque procédure est appliquée sur les données brutes et sur les données ajustées sur les facteurs, en utilisant la méthode présentée Section 4.2 : pour chaque jeu de données, les paramètres de covariance Ψ et B et les facteurs latents Z sont estimés sur les jeux de données d'apprentissage. Les données facteurs-ajustées (étape de décorrélation) sont calculées suivant la formule $x - Bz$. Les estimations $\hat{\Psi}$ et $\hat{B}$ sont utilisées pour estimer les facteurs latents des données test, décorrélées ensuite de la même manière. Les méthodes de classification sont ensuite appliquées sur les données d'apprentissage décorrélées et testées sur les jeux de données test décorrélés.

Les erreurs de prédiction sont calculées sur un jeu de données indépendant composé de 10 000 observations équilibrées dans chaque groupe et généré selon

chaque scenario de dépendance. Les performances des méthodes sont mesurées par l'erreur de prédiction sur les données test, le nombre de variables sélectionnées et la proportion de variables prédictives sélectionnées (et noté ensuite "précision") .

6.3 Résultats

Validation croisée La Table 3.4 reporte les erreurs de prédiction dans une situation sans signal ($\mu_0 = \mu_1 = 0_m$, structure de covariance en 2 blocs). Les résultats ne sont pas "trop" optimistes dans le sens où les erreurs de prédiction sont très proches de 0. Ceci assure que tous les paramètres sont bien étudiés indépendamment des données tests et que les étapes de sélection de variables et de classification sont renouvelées pour chaque jeu de données. Comme le mentionne Hornung et al. (2014), cette vérification n'est pas triviale et peut affecter la mesure de l'erreur de prédiction.

Table 3.4 – Vérification des taux d'erreur par validation croisée (erreurs de prédiction) dans une situation sans signal. On constate que l'erreur de classement est proche de 50%, qui est la valeur attendue lorsqu'aucune différence n'est introduite entre les groupes.

	Données brutes	Données facteurs-ajustées
LASSO	0.4989	0.4990
SLDA	0.4989	0.4992
SDA	0.5000	0.5004
DDA	0.4999	0.4996

Cas de l'indépendance Le scenario **(A)** permet de confirmer que la méthode proposée ne détecte pas de la dépendance à tort. En effet, aucun facteur n'a été extrait sur les 1000 jeux de données simulés selon une structure d'indépendance : les méthodes facteurs-ajustées donnent donc exactement les mêmes résultats que leur version originale (voir Table 3.5).

Table 3.5 – Aucun facteur n'est extrait pour les jeux de données générés sous l'indépendance **(A)**. La version facteurs-ajustée des méthodes est équivalente à la version brute.

	Erreur de prédiction	Variables	Précision (%) mean (sd)
LASSO	0.3858	12.82	40.32 (20.96)
SLDA	0.3873	10.00	39.50 (15.33)
SDA	0.3868	35.09	35.52 (21.77)
DDA	0.3489	32.90	38.44 (23.68)

Structures avec dépendance On s'intéresse maintenant aux résultats obtenus sur les structures **(B)**, **(C)** et **(D)**. D'après la Table 3.6 et la Figure 6.3, on remarque

que les quatre méthodes de sélection comparées (LASSO, SparseLDA, DDA et SDA) sont améliorées par l'ajustement sur les facteurs. En effet, les erreurs de classement sont plus faibles et les précisions de sélection sont globalement meilleures.

Plus précisément, concernant la structure en blocs **(B)**, les taux d'erreur sont plus faibles pour toutes les méthodes et les variables prédictives sont plus souvent sélectionnées, excepté pour la méthode SDA où les précisions sont similaires.

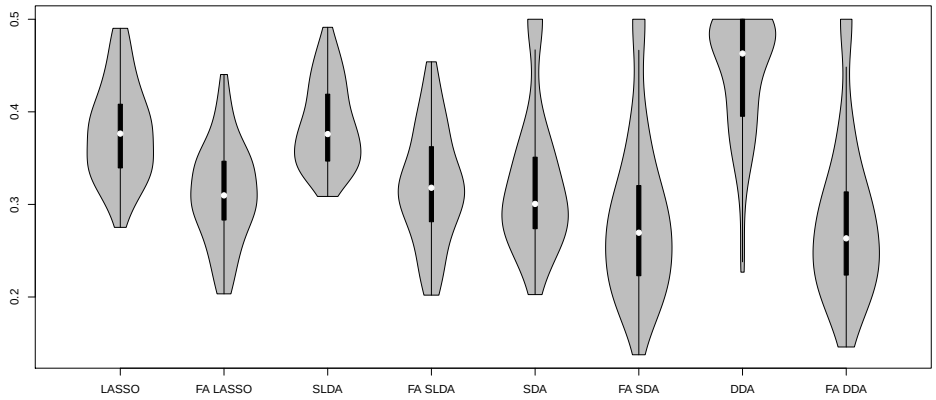
Le scenario **(C)** mène sans surprise aux résultats les plus marqués car les données sont simulées d'après un modèle à facteurs.

La méthode DDA est celle qui donne les taux d'erreur les plus élevés, lorsqu'elle est appliquée aux données brutes. L'étape de sélection est très instable pour la structure **(C)**. En effet, aucune variable n'est sélectionnée dans 15% des cas ce qui explique la moyenne de 4.18 dans la Table 3.6. Concernant les deux autres scenario, le nombre de variables sélectionnées est élevé sans pour autant que les variables prédictives le soient. Enfin, la DDA, qui suppose l'indépendance entre les covariables, est plus adaptée aux données décorrélées et donne ainsi de meilleurs résultats en terme de sélection et de classification.

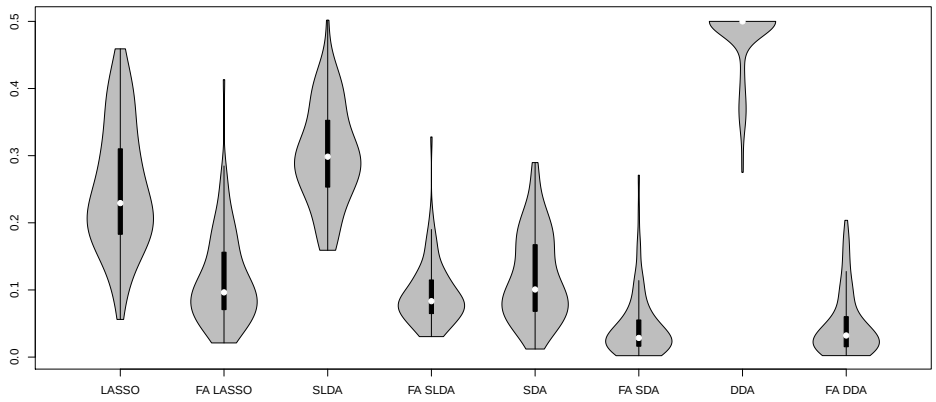
L'ajustement sur les facteurs bénéficie le plus aux méthodes LASSO et Sparse LDA. On peut remarquer que ces deux méthodes donnent des résultats similaires, ce qu'on peut attribuer au fait qu'elles sont toutes fondées sur une pénalisation ℓ_1 et que Clemmensen et al. (2011) utilise une approche par régression sur variables indicatrices de la LDA pour estimer les paramètres de régression. Par ailleurs, la méthode SDA est celle qui est le moins améliorée par la décorrélation. En effet, SDA est une méthode comparable à FADA car elle est aussi fondée sur une étape de décorrélation. Néanmoins, les performances de SDA sont légèrement améliorées par FADA. On peut l'expliquer par la flexibilité du modèle à facteurs pour capter les structures complexes de covariance (notamment la dépendance temporelle, voir Sheu et al. (2015)), peut-être plus performante que l'approche par shrinkage par un estimateur de type James-Stein.

7 Package FADA

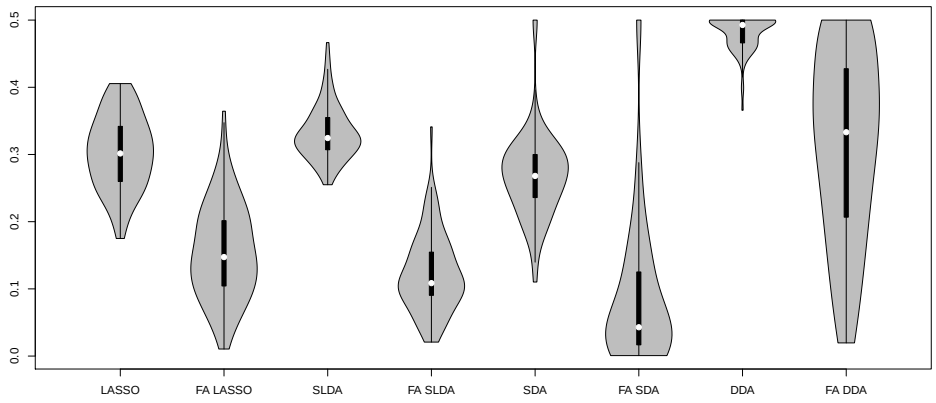
Cette méthode est implémentée dans un package R disponible sur le CRAN, intitulé FADA (Factor Adjusted Discriminant Analysis, voir Perthame et al. (2014)). Ce package est composé de trois fonctions principales et fonctionne en deux étapes. La première étape est la décorrélation d'un jeu de données d'apprentissage via la fonction `decorrelate.train` selon l'algorithme décrit ci-dessus. La fonction `decorrelate.test` permet de décorréler un jeu de données test à partir des estimations obtenues lors de l'appel de la fonction `decorrelate.train` sur des données d'apprentissage. Ensuite, la seconde étape est une étape de classification via la fonction `FADA`. Cette fonction propose plusieurs méthodes de classification supervisée : SDA, SparseLDA et Lasso. La fonction hérite de toutes les options disponibles dans les packages implémentant ces méthodes. Le calcul de l'erreur de classification peut se faire sur des données test ou par validation-croisée si aucun jeu de données test n'est fourni. L'ensemble des variables sélectionnées ainsi que les coefficients de régression sont renvoyés.



(a) 2-blocks structure **(B)**



(b) Factor structure **(C)**



(c) Temporal dependence **(D)**

TABLE 3.6 – Simulation results for several designs of dependence

Method	Prediction error	Features	Precision (%) mean (sd)
		Block structure **(B)**	
LASSO	0.3780	12.64	40.05 (23.85)
Factor-adjusted LASSO	0.3118	15.44	49.16 (21.30)
SLDA	0.3872	10.00	39.80 (15.50)
Factor-adjusted SLDA	0.3426	10.00	50.80 (16.00)
SDA	0.3244	41.63	42.12 (17.77)
Factor-adjusted SDA	0.2863	44.19	42.46 (18.08)
DDA	0.4393	165.10	28.31 (24.46)
Factor-adjusted DDA	0.2820	48.44	42.13 (19.14)
		Factor structure **(C)**	
LASSO	0.2660	14.025	62.67 (14.94)
Factor-adjusted LASSO	0.1038	8.477	90.43 (12.35)
SLDA	0.3000	10.00	68.80 (17.25)
Factor-adjusted SLDA	0.0926	10.00	87.50 (11.67)
SDA	0.1258	70.00	50.29 (14.84)
Factor-adjusted SDA	0.0452	53.17	65.17 (19.00)
DDA	0.4772	4.18	69.75 (18.30)
Factor-adjusted DDA	0.0474	55.26	65.04 (20.61)
		Temporal dependence **(D)**	
LASSO	0.3020	13.10	62.36 (20.63)
Factor-adjusted LASSO	0.1510	8.03	93.02 (9.69)
SLDA	0.3314	10.00	62.50 (17.08)
Facto-adjusted SLDA	0.1222	10.00	90.90 (10.83)
SDA	0.2695	57.20	75.07 (23.94)
Factor-adjusted SDA	0.0893	68.22	67.93 (25.66)
DDA	0.4813	149.42	15.58 (15.27)
Factor-adjusted DDA	0.3146	97.65	48.76 (29.91)

TABLE 3.7 – Violin plots of error rates

8 Conclusion

Dans ce chapitre, l'impact de l'hétérogénéité des données sur le rang et la stabilité des variables sélectionnées est illustré sur un exemple de modèle de classification supervisée. La plupart des méthodes classiques en classification supervisée supposent l'indépendance ou des corrélations faibles entre les variables. Cependant, l'hétérogénéité des données remet en cause cette hypothèse. La méthode décrite dans ce chapitre propose un cadre générale de modélisation de la dépendance. Un modèle à facteurs est utilisé pour capter la dépendance sur un petit nombre de variables latentes et la règle de Bayes conditionnelle à la structure de dépendance est définie. Ensuite, un algorithme prenant en compte la structure de covariance pour estimer simultanément la matrice de covariance, le signal et les probabilités individuelles est proposé afin de décorréler les données. Enfin, on montre que l'optimalité de la règle de classification de Bayes conditionnelle revient à appliquer la règle de Bayes usuelle aux données ajustées sur l'effet des facteurs latents.

On remarque que l'ajustement sur les facteurs améliore la stabilité de certaines procédures usuelles de classification. Un apport de cette étape de décorrélation est que les performances de classification sont nettement améliorées dans les situations où la structure de dépendance est élevée et peut être modélisée par un modèle à facteurs.

L'étude par simulations donne de bons résultats sur des structures associées à des données génomiques, selon plusieurs auteurs, ce qui est un des domaines d'intérêt de ce manuscrit. Néanmoins, cette approche convient aussi à d'autres domaines scientifiques. En effet, l'illustration sur la matrice de Toeplitz laisse penser que cette méthode est intéressante aussi dans ces cas de dépendance temporelle.

Enfin, dans ce manuscrit, on se place dans un contexte de LDA. On considère donc que la structure de covariance est la même pour tous les groupes. On peut cependant extraire des facteurs latents liés à la variable de groupe en considérant un modèle à facteurs par groupe. Cependant, dans un contexte de grande dimension, le nombre d'observations est faible au regard du nombre de variables. Estimer plusieurs modèles indépendamment pour chaque sous-échantillon risque de réduire la puissance de la méthode à détecter le signal biologique des données (la différence entre les groupes).

CHAPITRE 4

IDENTIFICATION D'UN SIGNAL PAR HIGHER CRITICISM THRESHOLDING DÉCORRÉLÉ POUR DES DONNÉES ERP

RÉSUMÉ : les potentiels évoqués (ERP) sont des mesures permettant de relier l'activité électrique cérébrale à des événements moteurs, sensoriels ou cognitifs au cours du temps. Un des principaux objectifs des études ERP est de sélectionner des instants (souvent rares) durant lesquels des variations (faibles) d'ERP sont significativement associées à une condition expérimentale d'intérêt. Le Higher Criticism Thresholding (HCT) est une procédure de sélection performante sous les conditions du modèle *Rare-and-Weak* qui parait donc adaptée à l'analyse de données ERP. Cependant, les ERPs présentent une forme de dépendance temporelle complexe qui enfreint les hypothèses sous lesquelles le signal peut être identifié efficacement par la procédure HCT. Ce chapitre illustre d'abord l'impact de la dépendance sur l'identification d'un signal par la méthode HCT. Un modèle à facteurs pour la structure de covariance est introduit dans la procédure HCT afin de décorréler les statistiques de test et de restaurer sa stabilité. Les limites de détectabilité du signal pour une structure de dépendance admettant une décomposition en facteurs sont déduites, permettant d'étendre le diagramme de phase souvent associé à la méthode HCT. A partir de simulations et d'une étude de données réelles, la méthode proposée semble estimer de manière plus précise le support du signal lorsqu'elle est comparée au HCT standard et à d'autres approches fondées sur une estimation par *shrinkage* de la matrice de covariance.

Sommaire

A la différence du Chapitre 2, le Chapitre 4 aborde une procédure de tests multiples dont le but initial n'est pas de contrôler le taux de faux positifs mais de détecter un signal de façon optimale. En effet, dans le cadre de la détection d'un signal, c'est-à-dire le test global de sa nullité, le Higher Criticism est construit pour atteindre les limites d'optimalité de détection, pour certaines conditions sur la parcimonie et l'amplitude du signal. Cette méthode est aussi connue pour être efficace en terme d'estimation du support du signal. Ce chapitre étudie donc l'apport de la décorrélation des statistiques de test sur les propriétés du Higher Criticism en situation de dépendance pour la détection et pour l'identification d'un signal.

1 Introduction

Les potentiels évoqués (ERP) sont des différences de potentiels mesurées à différents emplacements du crâne d'un sujet. Les ERP permettent de relier l'activité électrique cérébrale, mesurée par électroencéphalographie (EEG), à une condition expérimentale physique ou mentale. Comme l'imagerie par résonance magnétique fonctionnelle (fMRI), les ERPs sont des outils de mesures non invasifs enregistrant directement l'activité neurologique corticale. En revanche, contrairement à la fMRI, les ERPs possèdent une meilleure résolution temporelle pour étudier l'évolution temporelle des processus mentaux et sont moins couteux. En recherche fondamentale, les ERPs offrent une méthode psychophysiologique pour étudier les processus attentionnels, le langage et les fonctions de la mémoire, ce qui permet d'obtenir des informations non disponibles à partir d'études comportementales seules. En recherche clinique, les ERPs font partie des nombreux biomarqueurs non invasifs proposés pour évaluer des désordres neurologiques ou psychiatriques tels que la

maladie d'Alzheimer, la déficience cognitive légère amnésique, les troubles de l'attention ou encore l'hyperactivité.

Durant d'une expérience, les ERPs sont habituellement mesurés en millisecondes (ms) durant une à quelques secondes, à partir du début d'un événement extérieur (stimulus). Les courbes d'ERP sont connues pour être bruitées et très variables, à la fois au sein d'un sujet et entre les sujets, ce qui explique que les courbes sont habituellement moyennées pour une même condition expérimentale et pour le même sujet. Afin d'identifier des intervalles de temps durant lesquels les ERPs sont reliés à un stimulus (réponse), les chercheurs doivent étudier simultanément la significativité de milliers de tests d'hypothèse. Un équilibre doit alors être établi entre le maintien d'un taux de faux positifs suffisamment faible et l'assurance d'une puissance de détection du signal correcte. Le but de ce chapitre est donc d'atteindre cet objectif pour des données ERP présentant une structure de dépendance temporelle forte.

La recherche d'instants durant lesquels les ERP sont significativement associés à une variable réponse peut être vue comme un problème d'identification d'un signal sous le modèle *Rare and Weak* introduit par Donoho and Jin (2004). L'analyse à grande échelle de courbes ERP est en effet fondée sur une collection de statistiques de test $\mathcal{T} = (\mathcal{T}_1, \ldots, \mathcal{T}_T)$ associées aux tests de non association de la mesure à l'instant t et de la variable cible, d'hypothèse nulle $H_{0,t}$. Le modèle *Rare and Weak* est certes simple mais il fournit un cadre théorique utile pour étudier les propriétés des procédures de détection d'un signal. Dans le contexte des tests multiples, Donoho and Jin (2004) proposent la procédure *Higher Criticism Thresholding* (HCT) en s'inspirant de l'idée de Tukey (1976) d'un critère de significativité globale d'un ensemble de tests. La procédure HCT est connue pour être optimale pour détecter un signal sous l'indépendance. En effet, les limites de détection peuvent être exprimées et Donoho and Jin (2004) montrent que la procédure HCT atteint les limites optimales théoriques de décision. Pour le problème plus difficile d'identification de variables sous l'alternative dans une optique de classification ou de prédiction, Donoho and Jin (2008) montrent la supériorité de la procédure HCT par rapport aux procédures de tests multiples visant à contrôler le taux de faux positifs (FDR).

Comme mentionné dans le Chapitre 2 et dans Causeur et al. (2012), la structure de dépendance forte observée entre les courbes ERP induit une régularité temporelle des statistiques de test : des p-valeurs faibles sont observées en dehors du support du signal ce qui peut mener à des décisions erronées. Cette instabilité du rang des probabilités critiques due à la dépendance est aussi rapportée dans le contexte de l'analyse de données génomique (voir par exemple Ahdesmäki and Strimmer (2010), Friguet et al. (2009) et dans le Chapitre 3). La procédure HCT est connue pour être performante lorsque les tests sont faiblement corrélés (voir Hall and Jin (2008)) mais ses performances peuvent être améliorées en prenant en compte la dépendance (voir Ahdesmäki and Strimmer (2010) et Hall and Jin (2010)). Par exemple, Hall and Jin (2010) montrent que les limites théoriques de détection du signal établies sous le modèle *Rare and Weak* sont affectées par une forte dépendance entre les tests et les auteurs introduisent en conséquence la procédure *innovated* HCT (iHCT). Hall and Jin (2010) démontrent alors qu'en situation de dépendance, les propriétés de détection de la procédure HCT sont rétablies grâce à la procédure iHCT.

Par ailleurs, dans un contexte de sélection de variables en analyse linéaire discriminante, Ahdesmäki and Strimmer (2010) appliquent la procédure HCT aux CAT-scores, des statistiques de tests ajustées sur la corrélation par une estimation *shrinkée* de la matrice de covariance (voir Zuber and Strimmer (2009)). Les auteurs montrent que les performances de la procédure HCT sont améliorées par cette décorrélation par une racine de l'inverse de la matrice de covariance obtenue par une estimation de type James-Stein de la matrice de covariance. Dans ce contexte de classification, dans le Chapitre 3, la notion de décorrélation à la manière de Friguet et al. (2009) est adaptée au problème de la sélection de variables pour la classification supervisée. L'approche proposée repose sur une hypothèse assez générale de structure en facteurs de la matrice de covariance et un algorithme d'estimation simultanée du signal et de la matrice de covariance est déduit, afin de décorréler efficacement les données avant l'étape de classification.

Comme illustré dans le Chapitre 2 et dans Causeur et al. (2012), la structure de dépendance observée sur les statistiques de tests calculées sur les données ERP peut être approchée par une décomposition de la matrice de covariance par un petit nombre de facteurs. La procédure proposée dans ce chapitre a pour but de sélectionner efficacement les instants pour lesquels l'activité cérébrale est associée à la variable de traitement. Cette approche possède aussi de bonnes propriétés algébriques permettant l'expression explicite d'une racine de la matrice de covariance inverse. Ce chapitre est motivé par l'analyse de données ERP collectées durant une tâche dite de *oddball* dont le déroulement de l'expérience est décrit dans la Section 2. Cette section introduit aussi quelques méthodes de détection d'un signal dans le cadre du modèle linéaire multivarié. Une rapide revue de la procédure HCT en situation d'indépendance et de dépendance est présentée dans la Section 3. Le modèle à facteurs est rappelé en Section 4 puis une extension des limites de détection sous dépendance et la méthode *Factor-innovated* HCT (F-iHCT) sont proposées. Les propriétés de cette procédure sont étudiées sur des simulations et sur les données récoltées lors de l'expérience de *oddball* dans la Section 5. Enfin, la Section 6 conclut ce chapitre par une discussion.

2 Détection d'un signal lors d'expériences ERP

On considère dans ce chapitre le problème statistique d'un test de comparaison des moyennes de I groupes de courbes ERP où le nombre de courbes dans le groupe i est noté n_i. Les $n = n_1 + \ldots + n_I$ courbes ERP sont observées aux instants $\{t_1, \ldots, t_T\}$ pour J sujets. Ce problème de détection d'un signal est motivé par une expérience auditive de *oddball*.

2.1 Expérience auditive de oddball

La tâche dite de *oddball* est une des tâches expérimentales les plus couramment utilisées dans les études ERPs (voir Picton (1992)). Le principe de cette expérience est de présenter deux classes de stimuli, l'un se produisant fréquemment (standard),

l'autre plus rarement (cible). Il est demandé au sujet de faire la distinction entre les deux stimuli et de répondre en fonction du stimulus désigné comme cible.

L'exemple traité dans ce chapitre concerne des données récoltées lors d'une expérience d'ERP auditive réalisée à Kaohshung Medical University à Taiwan. Les tâches consistent en deux tonalités pures de 500 Hz et 1000 Hz. La première tonalité est présentée 120 fois parmi les 150 répétitions alors que la seconde (la cible) n'est présentée que 30 fois. L'ordre de présentation des tonalités est aléatoire et on demande au sujet de compter silencieusement le nombre de cibles. Pour chacune des 4 électrodes (FZ, C3, C4 et O1) et pour chacun des $J = 15$ sujets, une courbe ERP est obtenue pour les deux conditions de tonalité. Chaque courbe commence à -100 millisecondes (ms) et se termine à 399.5 ms, avec une mesure toutes les demi-millisecondes. Le stimulus commence à 0 ms. Dans la suite de ce chapitre, seuls les résultats sur l'électrode FZ sont présentés.

Parmi une littérature vaste sur le sujet, Williams et al. (2005) observent un signal au niveau de la zone parieto-centrale du crâne autour de 300 ms (appelé composante P300), plus marqué après l'occurence de l'événement cible. La problématique est de sélectionner des temps, possiblement autour de 300 ms, pour lesquels les mesures ERP permettent de détecter quelle tonalité a été présentée au sujet. Le but final est de vérifier si la composante P300 peut être considérée comme un marqueur électrophysiologique pour étudier des troubles psychiatriques ou neurologiques.

On s'attend à ce que l'activation cérébrale soit différente au cours du temps, en fonction de la fréquence de tonalité écoutée par les participants : 500 ou 1000 Hz. Les données sont constituées de $T = 799$ instants mesurés pour $J = 15$ sujets et pour $I = 2$ conditions. En guise d'exemple, la Figure 4.1 montre les courbes ERP pour les sujets 1 (ligne bleue) et 2 (ligne orange) pour la condition 500 Hz (trait plein) et 1000 Hz (trait pointillé). Cette figure illustre la grande variabilité observée entre les sujets.

2.2 Modèle linéaire multivarié

Analyse de variance On désigne par Y_{ijt} la mesure ERP au temps t, pour tout $t \in \{t_1, \ldots, t_T\}$, pour la condition i, $i \in [1; I]$ et pour le sujet j, $j \in [1; n_i]$. On observe au total $n = n_1 + \ldots + n_I$ courbes d'ERP. On considère dans un premier temps le modèle d'analyse de variance suivant :

$$Y_{ijt} \quad = \quad x'_{0ij}\mu_t + a_{it} + \varepsilon_{ijt}, \tag{4.1}$$

où x_{0ij} est un r-vecteur de covariables d'ajustement pour le sujet j ne dépendant pas de la condition i, a_{it} est l'effet de la condition i au temps t sur la mesure Y_{ijt}. Pour l'expérience auditive de *oddball*, $x_{0ij} = (1, \delta_j)$ où δ_j est une variable binaire prenant la valeur 1 pour le sujet j. Le vecteur $\varepsilon_{ij} = (\varepsilon_{ijt_1}, \ldots, \varepsilon_{ijt_T})$ est un terme d'erreur distribué selon une loi normale centrée, d'écart-type $\sigma = (\sigma_{t_1}, \ldots, \sigma_{t_T})'$ et de matrice de corrélation R. La Figure 4.2 indique que les écart-types intra-condition $s = (s_{t_1}, \ldots, s_{t_T})'$ varient fortement au cours de l'expérience, où s_t^2 est l'erreur quadratique moyenne corrigée des degrés de liberté au temps t. De plus, la

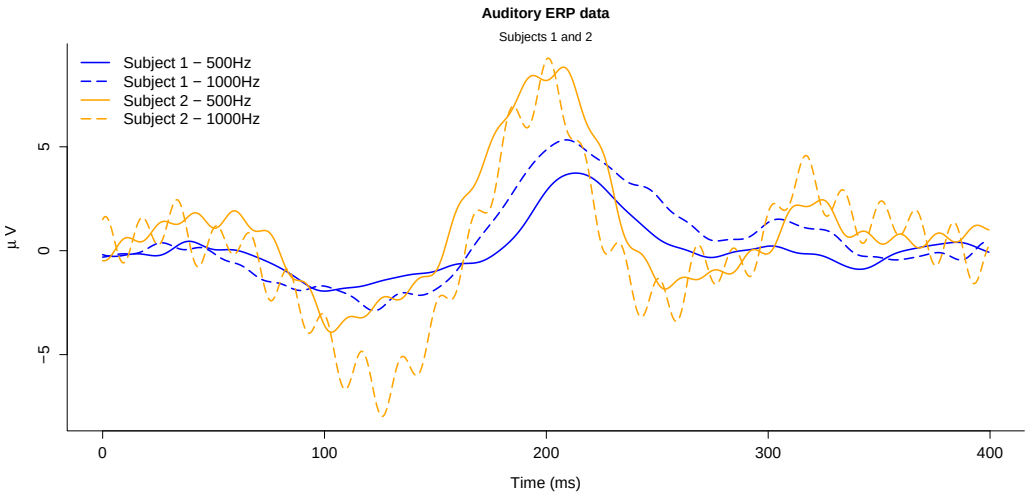


FIGURE 4.1 – Représentation des courbes ERP pour les sujets 1 (bleu) et 2 (orange) observées lors de l'expérience auditive pour la condition Hz500 (trait plein) et Hz1000 (trait pointillé)

Figure 4.3 présente une image des corrélations intra-condition entre mesures ERP et montre que la structure de dépendance est caractérisée par des auto-corrélations élevées et une structure en blocs.

Le modèle (4.1) peut s'exprimer de manière équivalente sous forme matricielle :

$$Y_t \quad = \quad X_0\mu_t + Xa_t + \varepsilon_t, \tag{4.2}$$

où $Y_t = (Y_{11t}, \ldots, Y_{1,n_1,t}, Y_{21t}, \ldots, Y_{2,n_2,t}, \ldots, Y_{I1t}, \ldots, Y_{I,n_I,t})'$, X_0 est une matrice $n \times r$ dont les lignes sont les vecteurs x'_{0ij}, $a_t = (a_{2t}, \ldots, a_{It})'$ est le $(I-1)$-vecteur des effets associés à la condition expérimentale, X est une matrice de design $n \times (I-1)$ dont la i-ème colonne X_i est nulle excepté pour les composantes comprises entre $n_1 + \ldots + n_i + 1$ et $n_1 + \ldots + n_{i+1}$ prenant la valeur 1. Enfin, le terme d'erreur ε_t est tel que $\varepsilon_t = (\varepsilon_{11t}, \ldots, \varepsilon_{1,n_1,t}, \varepsilon_{21t}, \ldots, \varepsilon_{2,n_2,t}, \ldots, \varepsilon_{I1t}, \ldots, \varepsilon_{I,n_I,t})'$.

Sous les hypothèses énoncées ci-dessus, notamment celle de normalité, et pour une structure de variance et de corrélation (σ, R), l'estimateur des moindres carrés généralisés (GLS) de a_t coïncide avec l'estimateur du maximum de vraisemblance (ML) :

$$\hat{a}_t \quad = \quad S_{xx}^{-1} S_{xy_t}, \tag{4.3}$$

où S_{xx} est la matrice $(I-1) \times (I-1)$ de variance empirique déduite de la matrice de design du modèle (4.2) :

$$S_{xx} \quad = \quad \frac{1}{n} X' P_0 X,$$

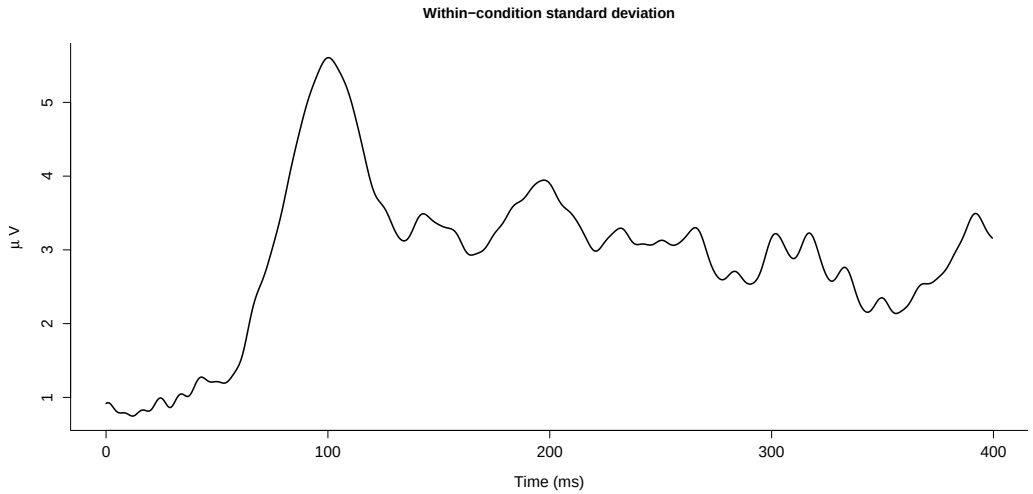


FIGURE 4.2 – Ecart-types intra-condition estimés sur les données d'ERP auditives

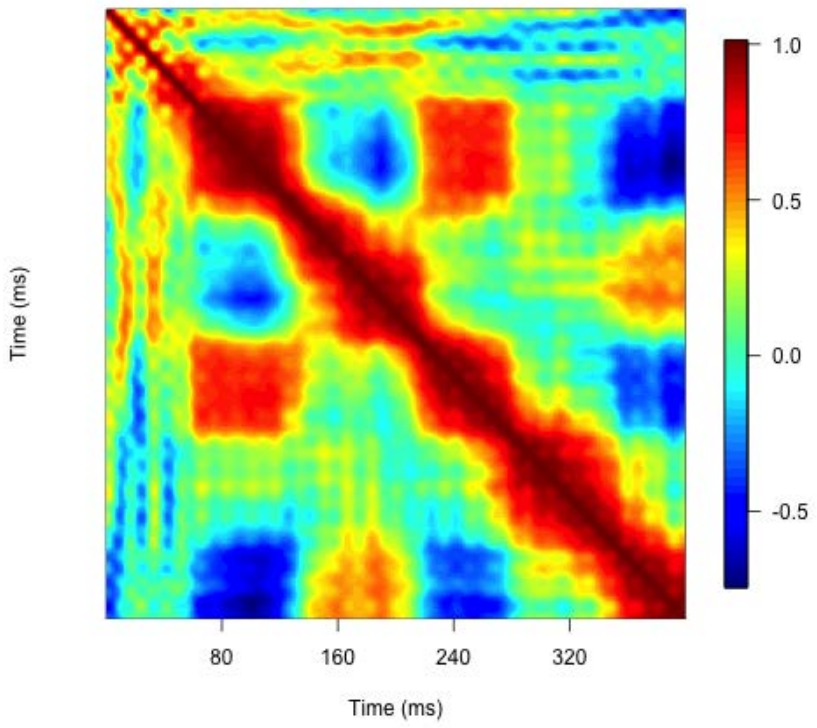


FIGURE 4.3 – Image de la matrice des corrélations intra-condition des données d'ERP auditives

où $P_0 = \mathbb{I}_n - X_0(X_0'X_0)^{-1}X_0'$. De la même manière, S_{xy_t} est le $(I-1)$-vecteur de covariance entre la variable réponse Y_t et la matrice de design :

$$S_{xy_t} = \frac{1}{n}X'P_0Y_t.$$

De plus, la variance résiduelle σ_t^2 est estimée par la moyenne quadratique des erreurs résiduelles corrigée des degrés de liberté des résidus :

$$s_t^2 = \frac{Y_t'(P_0 - P)Y_t}{n-(r+I-1)}, \tag{4.4}$$

où P est la matrice $n \times n$ de projection orthogonale suivante :

$$P = XP_0\Big[X'P_0X\Big]^{-1}P_0X'.$$

Détection d'un signal La détection statistique d'un signal revient à tester l'hypothèse nulle suivante :

$$\mathrm{H}_0 : \text{ pout tout } t,\ a_t = 0.$$

L'erreur de type I de ce test peut s'écrire comme la probabilité de déclarer à tort qu'il existe au moins un instant t pour lequel $a_t \neq 0$. Le contrôle de l'erreur de type I du test global sur toute la durée de l'expérience revient de manière équivalente à contrôler le Family-Wise Error Rate (FWER), c'est-à-dire la probabilité de rejet à tort de l'hypothèse nulle au moins une fois lors du test simultané de la collection d'hypothèses suivante :

$$\mathrm{H}_{0t} : \ a_t = 0.$$

Pour tester chacune de ces hypothèses, il est classique de construire la statistique de Fisher F_t suivante :

$$F_t = \frac{n}{I-1}\frac{\hat{a}_t'S_{xx}\hat{a}_t}{s_t^2}.$$

Les probabilités critiques associées à ces statistiques de test sont notées :

$$p_t = 1 - G_{I-1,n-(r+I-1)}(F_t),$$

où $G_{I-1,n-(r+I-1)}(.)$ est la fonction de répartition d'une loi de Fisher à $(I-1)$ et $(n-(r+I-1))$ degrés de liberté. Plusieurs procédures de tests multiples permettent de déterminer un seuil de rejet p^* sur la collection des p-valeurs associées aux tests de l'hypothèse H_{0t}, $t = \{t_1, \ldots, t_T\}$. Lorsque les tests sont indépendants, une des méthodes les plus répandues est la correction de Bonferroni, où le choix du seuil $p^* = \alpha/T$ garantit un contrôle du FWER inférieur à α :

$$\mathbb{P}_{H_0}\Big(\bigcup_t \big[\frac{n}{I-1}\frac{\hat{a}_t'S_{xx}\hat{a}_t}{s_t^2} \geq f^*\big]\Big) \leq \alpha,$$

où $f^* = G^{-1}_{I-1,n-(r+I-1)}(1-p^*)$. Cependant, les procédures de tests multiples assurant un contrôle du FWER sont connues pour être conservatives et la correction de Bonferroni n'est pas adaptée aux tests corrélés.

Analyse de variance multivariée Une approche alternative est de construire une statistique de test unique pour le test de l'hypothèse H_0 en concaténant les courbes observées en un vecteur $Y = (Y'_{t_1}, Y'_{t_2}, \ldots, Y_{t_T})'$. Le modèle (4.1) peut alors s'écrire sous la forme d'un modèle d'analyse de variance multivariée :

$$\begin{aligned} Y &= [\mathbb{I}_T \otimes X_0]\mu + [\mathbb{I}_T \otimes X]a + \varepsilon, \\ &= \tilde{X}_0\mu + \tilde{X}a + \varepsilon, \end{aligned} \tag{4.5}$$

où $\otimes$ est le produit de Kronecker et les paramètres du modèle sont tels que $\mu = (\mu'_{t_1}, \ldots, \mu'_{t_T})'$ et $a = (a'_{t_1}, \ldots, a'_{t_T})'$. Les résidus de ce modèle linéaire ne sont pas homoscédastiques. En effet,

$$\mathrm{Var}(\varepsilon) = [D_\sigma R D_\sigma] \otimes \mathbb{I}_n = V_\varepsilon,$$

où D_σ est une matrice diagonale $T \times T$ dont les termes diagonaux sont $\sigma_{t_1}, \ldots, \sigma_{t_T}$. Ainsi la variance de l'estimateur des moindres carrés généralisés $\hat{a}_{\mathrm{gls}} = (\hat{a}_{t_1}, \ldots, \hat{a}_{t_T})$ de a s'écrit :

$$\mathrm{Var}(\hat{a}_{\mathrm{gls}}) = \frac{1}{n}[D_\sigma R D_\sigma] \otimes S_{xx}^{-1}. \tag{4.6}$$

Dans le contexte des ERP, de récents articles (voir Bugli and Lambert (2006); Smith and Kutas (2015a,b)) proposent des tests de Fisher en analyse de variance fondés sur cette présentation des données. De plus, Bugli and Lambert (2006); Smith and Kutas (2015a,b) expliquent que ce contexte permet aussi d'introduire un modèle non-paramétrique de lissage pour μ et a à partir de B-splines ou d'ondelettes. Cette approche modifie la matrice de design du modèle et réduit ainsi le nombre de coefficients de régression. En effet, si φ désigne la matrice $T \times S$ associée à une base de fonctions, par exemple des B-splines (tels que $\varphi_{is} = \phi_s(t_i)$, $s = 1, \ldots, S$; $i = 1, \ldots, T$), alors $\tilde{X}_0 = \varphi \otimes X_0$ et $\tilde{X} = \varphi \otimes X$ dans l'Expression (4.5) et les paramètres μ et a sont des vecteurs de coefficients de régression sur la base de fonctions de dimension Sr et $S(I-1)$ respectivement.

Dans ce contexte, à (σ, R) connus, le meilleur estimateur linéaire sans biais des coefficients de régression est donné par la méthode des moindres carrés généralisés. La statistique de test de Fisher associée au test d'hypothèse $H_0 : \ a = 0$ s'écrit alors :

$$F_{\mathrm{gls}} = n\hat{a}'_{\mathrm{gls}}([D_{1/\sigma}R^{-1}D_{1/\sigma}] \otimes S_{xx})\hat{a}_{\mathrm{gls}}, \tag{4.7}$$

dont la distribution sous l'hypothèse nulle est une loi du χ^2 à $(I-1)T$ degrés de liberté. Cependant, la statistique de test F_{gls} n'est pas calculable en pratique car elle dépend des paramètres inconnus de variance. Pour gérer ce problème, les tests d'analyse de variance proposés dans Bugli and Lambert (2006); Smith and Kutas (2015a,b) sont fondés sur des hypothèses classiques d'homoscédasticité $V_\varepsilon = \sigma^2 I_{nT}$ ou sur une structure auto-régressive d'ordre 1 pour la matrice de corrélations R :

$$R_\rho = \begin{pmatrix} 1 & \rho & \rho^2 & \ldots & \rho^{T-1} \\ \rho & 1 & \rho & \ldots & \rho^{T-2} \\ \vdots & \vdots & & & \vdots \\ \rho^{T-1} & \rho^{T-2} & \rho^{T-3} & \ldots & 1 \end{pmatrix},$$

où ρ est le paramètre d'auto-corrélation d'ordre 1.

Hypothèse d'homoscédasticité On désigne par $\hat{Y}$ le vecteur des valeurs de Y ajustées sur le modèle (4.5) estimé par la méthode des moindres carrés généralisés. La statistique associée au test de l'hypothèse nulle $H_0 : \ a = 0$ sous l'hypothèse d'indépendance et d'homoscédasticité s'écrit :

$$F_{\text{ols}} \quad = \quad n\frac{\hat{a}'_{\text{gls}}(I_T \otimes S_{xx})\hat{a}_{\text{gls}}}{\hat{\sigma}^2}, \tag{4.8}$$

où les variances résiduelles sont estimées par :

$$\hat{\sigma}^2 \quad = \quad \frac{(Y-\hat{Y})'(Y-\hat{Y})}{T(n-(r+I-1))}. \tag{4.9}$$

Hypothèse d'hétéroscédasticité On remarque que si l'hypothèse d'homoscédasticité est relâchée, alors $V_\varepsilon = D_{\sigma^2} \otimes I_n$ et la statistique de test s'écrit :

$$F_s \quad = \quad n\hat{a}'_{\text{gls}}(D_{1/s^2} \otimes S_{xx})\hat{a}_{\text{gls}}, \tag{4.10}$$

où $s^2 = (s^2_{t_1}, \ldots, s^2_{t_T})$ est déduit de l'expression (4.4). Finalement, on remarque que F_s est proportionnel à la somme des statistiques de test individuelles F_t :

$$F_s \quad = \quad (I-1)\sum_{i=1}^{T} F_{t_i} \tag{4.11}$$

Dans ce cas, l'hétéroscédasticité peut être prise en compte simplement en considérant que la distribution de F_s est la même que celle d'une somme de T variables de loi de Fisher $\mathcal{F}_{I-1,n-(r+I-1)}$ indépendantes. Dans la suite, cette distribution sous l'hypothèse nulle est notée $\bar{\mathcal{F}}_{I-1,n-(r+I-1)}$.

Hypothèse de covariance auto-régressive Sous l'hypothèse d'une structure de covariance auto-régressive où la matrice de corrélations $R = R_\rho$, on introduit la matrice $L_{\rho,\sigma}$ suivante :

$$L_{\rho,\sigma} \quad = \quad \begin{pmatrix} -\frac{\rho}{\sigma_{t_1}} & \frac{1}{\sigma_{t_2}} & 0 & \ldots & 0 \\ 0 & -\frac{\rho}{\sigma_{t_2}} & \frac{1}{\sigma_{t_3}} & \ldots & 0 \\ \vdots & \vdots & & & \vdots \\ 0 & 0 & \ldots & -\frac{\rho}{\sigma_{t_{T-1}}} & \frac{1}{\sigma_{t_T}} \end{pmatrix}.$$

Si $\varepsilon^* = (L_{\rho,\sigma} \otimes I_n)\varepsilon$ désigne le $(n-1)T-$vecteur des innovations, alors on en déduit que $\text{Var}(\varepsilon^*) = (1-\rho^2)I_{(n-1)T}$. Ainsi, à partir d'une estimation initiale non nulle $\hat{\rho}_0$ de ρ permettant d'estimer la matrice $L_{\rho,\sigma}$, une nouvelle estimation de ρ peut être calculée via la variance de ε^* :

$$\hat{\rho}_1^2 = 1 - \frac{\sum_{i=1}^{n}\sum_{j=2}^{T} \varepsilon^{*2}_{it_j}}{(n-(r+I-1))(T-1)}.$$

A partir de cette estimation de ρ, on peut calculer un nouvel estimateur plug-in de $L_{\rho,\sigma}$ et déduire de nouveaux résidus décorrélés ε^*. Ceci définit un algorithme

itératif convergeant vers un estimateur $\hat{\rho}$ de ρ appelé estimateur Feasible GLS. De la même manière, une version *feasible* $F_{s,\hat{\rho}}$ de la statistique de tests F_{gls} est obtenue par plug-in des estimateurs de σ_t et R_ρ dans l'Expression (4.7) :

$$F_{s,\hat{\rho}} \quad = \quad n\hat{a}'_{\text{gls}}(D_{1/s}R_{\hat{\rho}}^{-1}D_{1/s} \otimes S_{xx})\hat{a}_{\text{gls}}. \tag{4.12}$$

La distribution exacte de $F_{s,\hat{\rho}}$ sous l'hypothèse nulle ne s'écrit pas de manière analytique. Les tests F d'analyse de variance sont généralement fondés sur une approximation asymptotique de la distribution sous l'hypothèse nulle par une loi de Fisher à $(I-1)T$ et $(n-(r+I-1))T$ degrés de liberté.

On considère ici les statistiques test F_s (voir Expression (4.10)). On peut remarquer que :

$$\text{Cor}(F_{t_i}, F_{t_j}) \quad \approx_{n\to+\infty} \quad \rho^{2|i-j|}.$$

La distribution sous l'hypothèse nulle de F_s dans le cas d'une structure de covariance auto-régressive peut être approchée par la distribution d'une somme de variables autocorrélées de loi marginale $\mathcal{F}_{I-1,n-(r+I-1)}$ et de paramètre d'auto-corrélation ρ^2. Dans la suite, cette distribution sous l'hypothèse nulle est notée $\bar{\mathcal{F}}_{I-1,n-(r+I-1)}(\rho)$.

2.3 Impact d'une erreur de spécification du modèle sur la détection d'un signal

Afin de comparer les différentes stratégies de détection d'un signal décrites dans la section précédente, des jeux de donnés sont simulés, dont les dimensions et la distribution tentent de reproduire celles des données réelles d'ERP auditives. Les données observées consistent en 30 courbes ERP observées dans deux conditions (15 dans chacune) sur l'intervalle de temps [0 ;400ms], mesurées toutes les demi-millisecondes. On s'intéresse ici au test de la significativité de la différence de courbes entre les deux conditions. Dans ces simulations, la courbe ERP moyenne dans la condition 1 est connue et constante, de valeur nulle. Le signal décrit une onde dans la condition 2, présentée Figure 4.4. Dans la suite, on constatera dans le cadre du modèle *Rare and Weak* de Donoho and Jin (2004), que cette situation tombe dans la région estimable du diagramme de phase d'un signal, en terme de parcimonie et de force du signal, en situation d'indépendance.

1000 jeux de données sont générés sous H_0 et sous H_1, chacun constitués de $n = 30$ lignes et $T = 799$ colonnes, d'écart-type s. La structure de corrélation est auto-régressive d'ordre 1 de paramètre $\rho = 0.99$. On remarque que l'auto-corrélation estimée sur les données de *oddball* est 0.997. Sur chaque jeu de données, les tests suivants sont réalisés :

1. Calcul de la statistique de test F_{ols} (voir Expression (4.8)) sous l'hypothèse d'homoscédasticité et d'indépendance, dont la distribution sous l'hypothèse nulle est $\mathcal{F}_{T,(n-2)T}$;
2. Calcul de la même statistique de test qu'au point précédent en incluant un lissage par B-splines des coefficients de régression. Ce test est implémenté

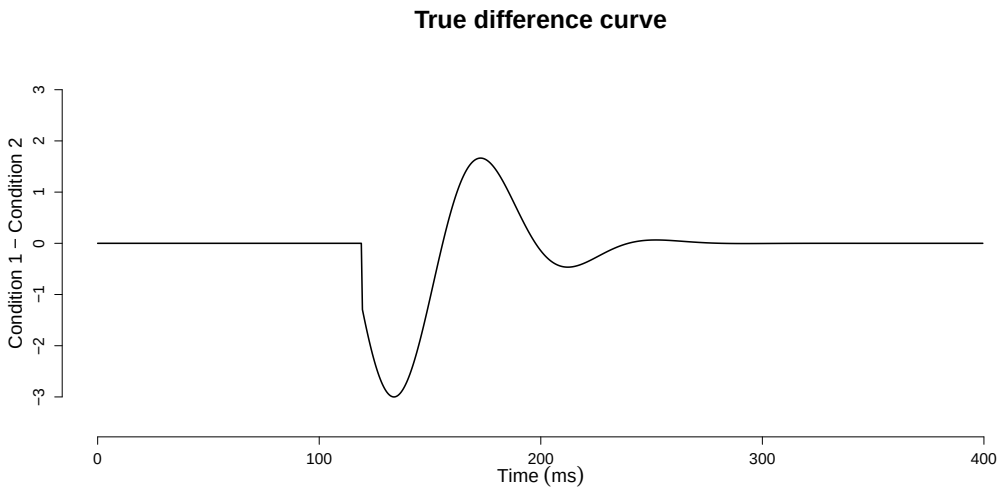


FIGURE 4.4 – Différence théorique entre les deux conditions dans l'étude par simulations

dans les fonctions `bam` et `anova.gam` du package `R mgcv`. Les paramètres de lissage sont fixés par défaut (minimisation d'un critère par validation croisée généralisée). On remarque que la distribution sous l'hypothèse nulle de la statistique de test est une loi de Fisher, dont les degrés de libertés prennent en compte le lissage en les ajustant sur la moyenne de la trace de la matrice de lissage ;

3. Calcul de la statistique de test $F_{s,\hat{\rho}}$ (voir Expression (4.12)) dont la distribution sous l'hypothèse nulle est approchée par $\mathcal{F}_{T,(n-2)T}$;
4. Calcul de la même statistique de test qu'au point précédent en incluant un lissage par B-splines des coefficients de régression. Ce test est aussi implémenté dans la fonction `bam` en spécifiant des arguments permettant d'introduire de l'auto-corrélation d'ordre 1 ;
5. Calcul de la statistique de test F_s (voir Expression (4.10)) dont la distribution sous l'hypothèse nulle est $\bar{\mathcal{F}}_{I-1,n-(r+I-1)}(\hat{\rho})$. Le calcul des probabilités critiques est fondé sur une estimation par Monte-Carlo de la distribution sous l'hypothèse nulle.

La Figure 4.5 présente la fonction de répartition empirique sous l'hypothèse nulle des probabilités critiques pour les cinq tests réalisés. Ce graphique confirme que l'impact principal d'une mauvaise spécification du modèle est sur le contrôle de l'erreur de type I, qui n'est pas contrôlée par le test F_{ols}. Ce phénomène est d'autant plus visible lorsque les coefficients de régression sont lissés. Ceci est cohérent avec les observations d'autres auteurs, comme Bugli and Lambert (2006), observant que les bandes de confiance calculées sous une hypothèse d'indépendance et d'homoscédasticité sont trop étroites. On remarque aussi que l'approximation de la distribution sous l'hypothèse nulle pour le test $F_{s,\hat{\rho}}$ est erronée et mène à un test très conservatif, avec et sans lissage des coefficients de régression. Le test global correspondant à une somme de statistiques de tests individuelles, couplé à une correction

de la distribution sous l'hypothèse nulle, contrôle l'erreur de type I au niveau choisi. En revanche, cette statistique F_s échoue à détecter le signal, comme le montre la Figure 4.6, présentant la distribution empirique sous l'hypothèse alternative des cinq tests réalisés.

Enfin, cette étude par simulations montre que si la distribution sous l'hypothèse nulle des statistiques de test en cas de dépendance est accessible, une stratégie de détection d'un signal fondée sur les statistiques individuelles F_t peut permettre d'atteindre un contrôle correct de l'erreur de type I. La méthode du Higher Criticism Thresholding présentée ci-après utilise le vecteur de ces statistiques de test individuelles. De plus, la structure de dépendance des données ERP est supposée décrite par un processus auto-régressif d'ordre 1. Cette hypothèse est peu réaliste (voir la Figure 4.3 présentant une image de la matrice de corrélation calculée sur les données de *oddball*).

Dans la suite, par cohérence avec le contexte introduit par Donoho and Jin (2004) pour introduire HCT, on s'intéresse au cas $I = 2$. Ainsi, on considère plutôt les statistique de test T, $\mathcal{T} = (\mathcal{T}_{t_1}, \ldots, \mathcal{T}_{t_T})'$ déduites en calculant la racine carrée signée des tests $\mathcal{F} = (\mathcal{F}_{t_1}, \ldots, \mathcal{F}_{t_T})'$.

3 Higher Criticism Thresholding pour la détection d'un signal

3.1 Différentes versions de la méthode Higher Criticism

Le Higher Criticism (HC) est une méthode pour la détection de signaux. Cette procédure fournit une unique probabilité critique associée au test de significativité d'un signal, calculée à partir des tests individuels de chaque coordonnée de ce signal. Le Higher Criticism a été initialement proposé par Tukey (1976) dans ses notes de cours à l'université de Princeton et a été réintroduit par Donoho and Jin (2004). Cette méthode est connue pour être optimale sous les hypothèses du modèle *Rare and Weak* (RW), défini par Donoho and Jin (2004) comme un modèle de mélange gaussien parcimonieux pour les statistiques de test. Dans ce chapitre, on préférera les notations du modèle RW de Hall and Jin (2010), équivalentes à celles de Donoho and Jin (2004) :

$$\mathcal{T} \quad = \quad \mu + \varepsilon, \varepsilon \sim \mathcal{N}(0, R) \tag{4.13}$$

où R est la matrice identité et μ est un vecteur parcimonieux dont la proportion de coordonnées non nulles est ε_T tel que $0 \leq \varepsilon_T \leq 1$. Ces coordonnées sont d'amplitude $A_T \geq 0$. On remarque que la normalité est supposée dans la plupart des études sur les ERP, où les tests d'association entre les ERP et la variable réponse sont des tests de Student (voir Causeur et al. (2012); Guthrie and Buchwald (1991)). Le modèle RW suppose une situation difficile dans laquelle le signal est rare :

$$\varepsilon_T \quad = \quad T^{-\beta}, \text{ avec } \beta \in (\frac{1}{2}, 1),$$

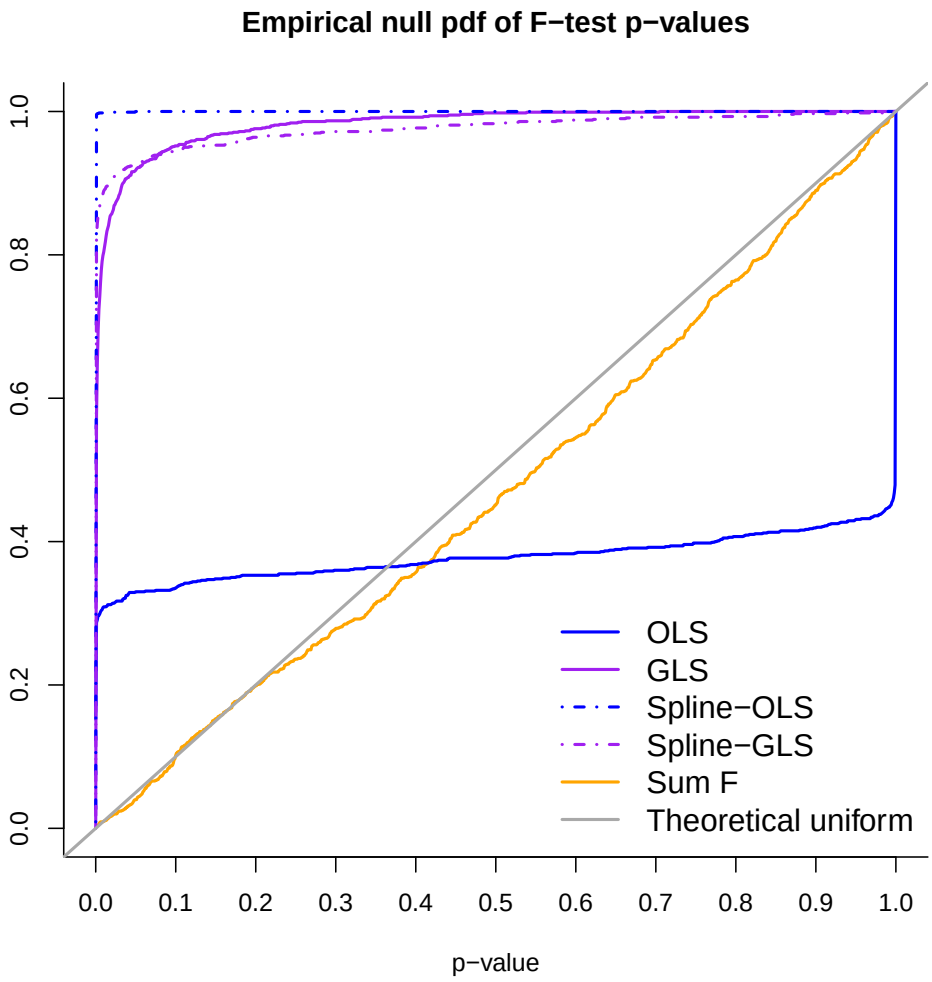


FIGURE 4.5 – Fonction de répartition empirique sous l'hypothèse nulle des probabilités critiques associées aux statistiques de test de Fisher sous l'hypothèse d'une structure de corrélation auto-régressive

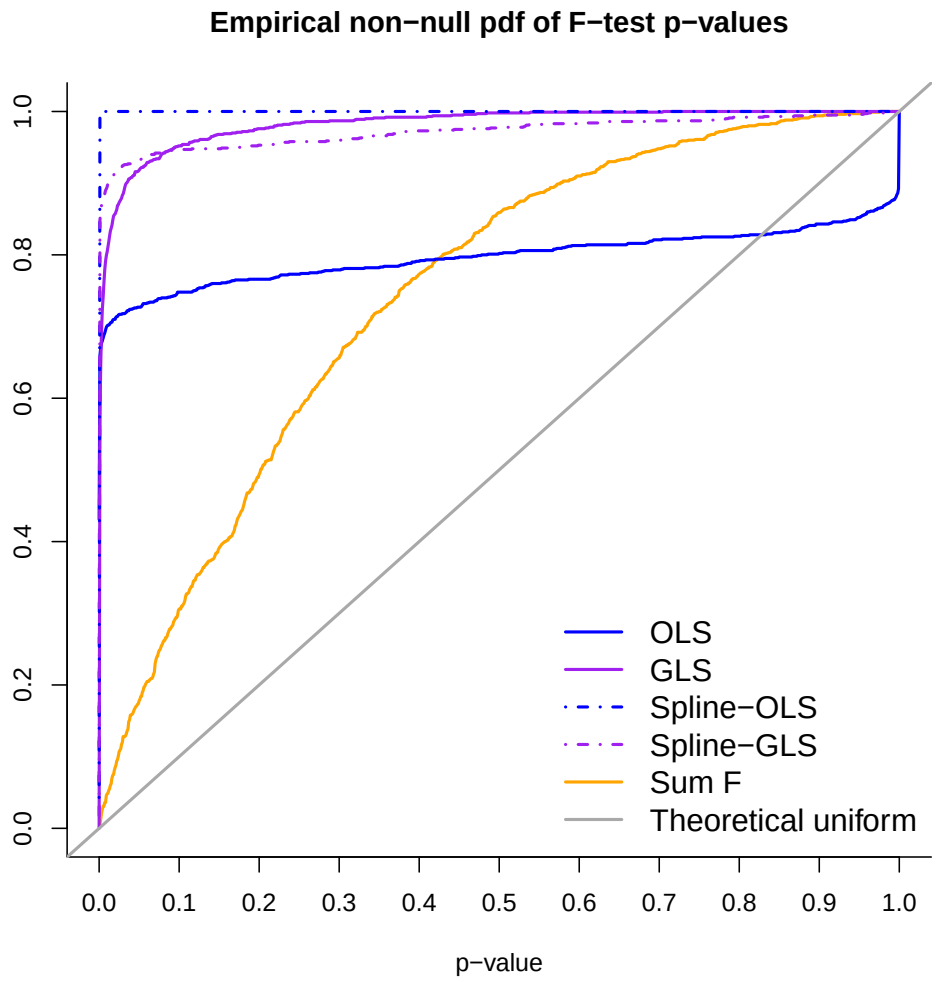


FIGURE 4.6 – Fonction de répartition empirique sous l'alternative des probabilités critiques associées aux statistiques de test de Fisher sous l'hypothèse d'une structure de corrélation auto-régressive

et faible :

$$A_T = \sqrt{2r\log(T)}, 0 < r < 1.$$

En situation d'indépendance, si les paramètres de parcimonie et de force du signal (β, r) sont connus, des limites explicites sur ces paramètres séparent l'espace (β, r) en une région indétectable, détectable (voir Ingster (1999)) et estimable (voir Donoho and Jin (2004)). Ces limites sont résumées dans un diagramme de phase (voir Figure 4.7) : si $r > \rho_D^*(\beta)$ alors le signal est détectable, ce qui signifie que la somme des erreurs de type I et type II pour le test du rapport de vraisemblance de Neymann-Pearson tend vers 0 quand le nombre de tests tend vers l'infini. Si $r > \rho_E^*(\beta)$ alors le signal n'est pas seulement détectable mais le support des coordonnées non nulles est identifiable. Si (β, r) sortent de ces limites, le signal est indétectable et, pour tout test, la somme des erreurs de type I et type II tend vers 1 lorsque le nombre de tests tend vers l'infini. Les limites de détectabilité $\rho_D^*(\beta)$ et d'estimabilité $\rho_E^*(\beta)$ ont la forme suivante :

$$\begin{aligned} \rho_D^*(\beta) &= \begin{cases} \beta - \frac{1}{2} & \text{si } \frac{1}{2} < \beta \leq \frac{3}{4} \\ (1 - \sqrt{1-\beta})^2 & \text{si } \frac{3}{4} < \beta < 1 \end{cases}, \\ \rho_E^*(\beta) &= \beta. \end{aligned}$$

Dans l'exemple introduit dans l'étude par simulations de la section précédente, la courbe de différence entre les deux conditions a la forme d'une onde. Ainsi, comme le montre la Figure 4.8, la courbe des coefficients r correspondante n'est pas constante sur l'intervalle où le signal est non nul, ce qui est supposé par le modèle RW. Cependant, on remarque que 3% ($\beta = 0.52$) du signal possède un coefficient r supérieur à 0.7. Le modèle RW correspondant à ces valeurs de paramètres $(\beta = 0.52, r = 0.7)$ appartient à la région estimable du diagramme de phase de la Figure 4.7.

La détection d'un signal par la méthode HC repose sur le principe que la présence d'un signal génère une différence entre la fonction de répartition empirique des p-valeurs et la fonction de répartition théorique uniforme $\mathcal{U}[0;1]$ sous l'hypothèse nulle. En effet, pour les coordonnées d'amplitude non nulle, on s'attend à ce que la statistique d'ordre des p-valeurs associée $p_{(t)}$ soit telle que $p_{(t)} \ll t/T$. Cette différence est mesurée par une fonction Higher Criticism notée $HC(p_{(t)}, t)$. Les variantes de cette méthode sont fondées sur différentes définitions de la fonction objectif HC : Donoho and Jin (2008) et Klaus and Strimmer (2013) standardisent la différence $t/T - p_{(t)}$ par l'écart-type des p-valeurs ordonnées sous l'hypothèse nulle, $\mathrm{Var}(p_{(t)}) = \frac{t}{T}(1 - \frac{t}{T})$ et définissent la fonction objectif HC par :

$$HC(p_{(t)}, t) = \sqrt{T}\frac{t/T - p_{(t)}}{\sqrt{t/T(1 - t/T)}}.$$

D'autre part, Donoho and Jin (2004) et Hall and Jin (2010) suggèrent de standardiser la différence $t/T - p_{(t)}$ par $p_{(t)}(1 - p_{(t)})$ et définissent la fonction objectif HC

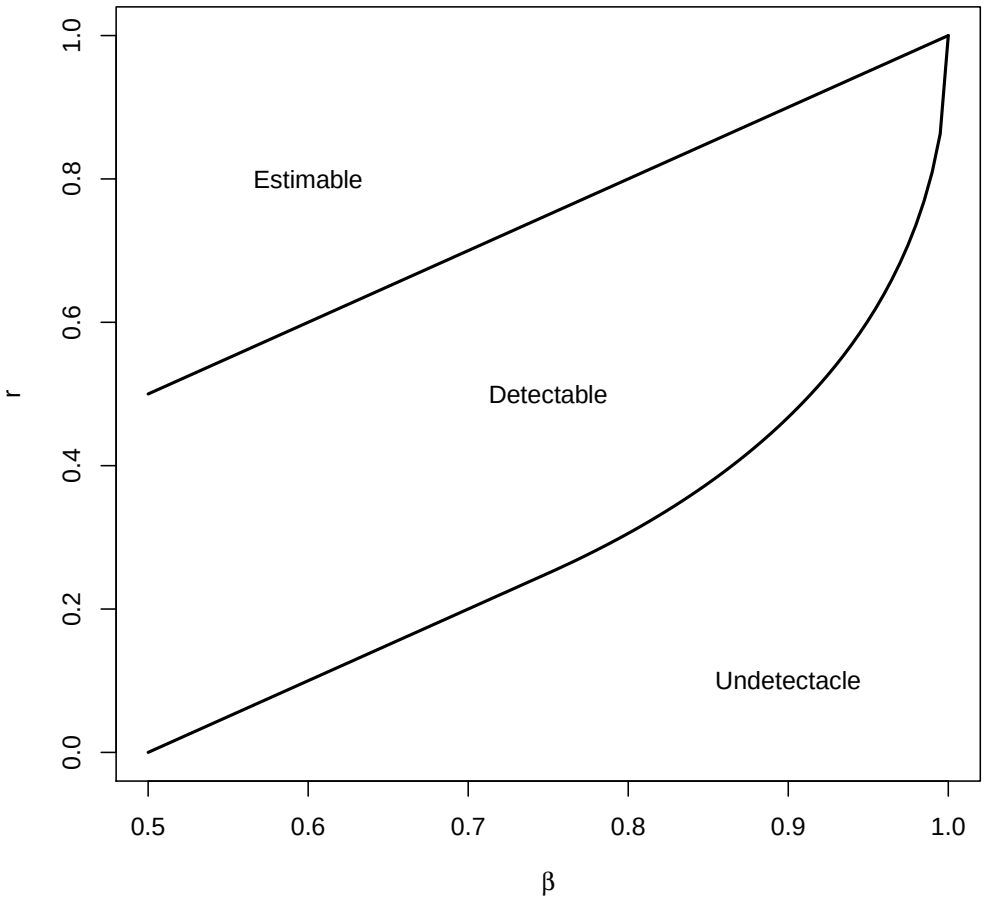


FIGURE 4.7 – Diagramme de phase des régions d'indétectabilité, de détectabilité et d'estimabilité du signal sous indépendance

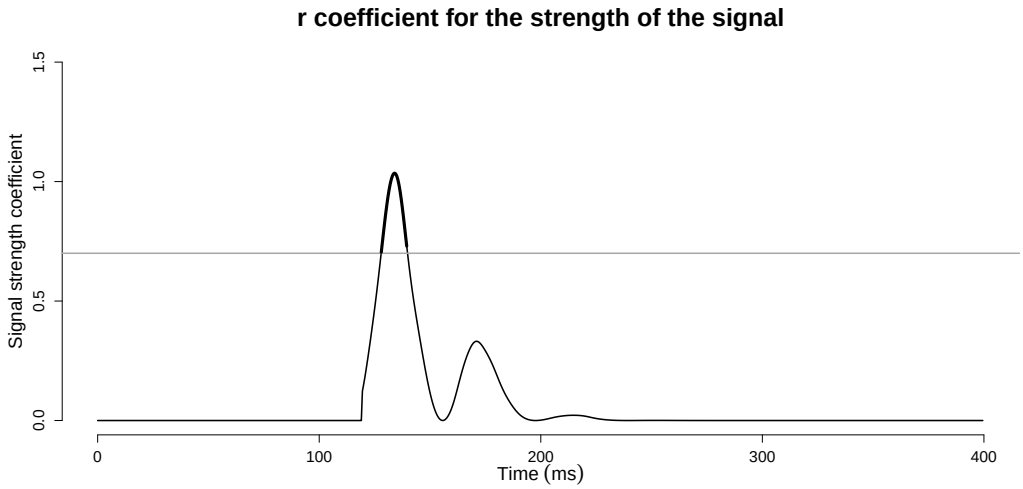


FIGURE 4.8 – Courbe des coefficients r dans l'étude par simulations de la Section 2.3. 3% ($\beta = 0.52$) du signal possède une amplitude supérieure à 0.7.

par :

$$HC(p_{(t)}, t) \quad = \quad \sqrt{T}\frac{t/T - p_{(t)}}{\sqrt{p_{(t)}(1 - p_{(t)})}}.$$

Par des arguments issus de la théorie des processus empiriques, Donoho and Jin (2004) montrent que :

$$\frac{HC^*}{\sqrt{2\log\log T}} \to 1, \text{ en probabilité,}$$

où HC^* est le maximum de $HC(p_{(t)}, t)$. On peut en déduire que l'erreur de type I d'une règle de détection d'un signal de la forme $HC^* \geq (1 + a)\sqrt{2\log\log T}$, pour $a > 0$, tend vers 0 quand T tend vers l'infini. De plus, Donoho and Jin (2004) montrent que les tests de cette forme sont optimaux dans le sens où, pour toute situation RW (β, r) dans la région détectable, l'erreur de type II de ces tests basés sur HC^* tend aussi vers 0 quand T tend vers l'infini. Cette propriété est connue comme l'adaptativité optimale du Higher Criticism. Afin d'obtenir une stratégie de test contrôlant l'erreur de type I, Cai et al. (2011) suggèrent d'estimer par Monte-Carlo la valeur de a à partir de la distribution sous l'hypothèse nulle de HC^*. On applique cette méthode au dispositif de simulations présenté dans la section précédente. Les Figures 4.9 et 4.10 présentent la fonction de répartition empirique sous l'hypothèse nulle (resp. sous l'alternative) des p-valeurs associées aux statistiques HC et F_s, qui prend en compte la dépendance. Les graphiques montrent que les deux méthodes atteignent des performances similaires.

En pratique, la fonction HC est souvent maximisée sur un sous-ensembles $\mathcal{I}$ de variables dont les p-valeurs sont faibles :

$$HC^* \quad = \quad \max_{t \in \mathcal{I}} HC(p_{(t)}, t).$$

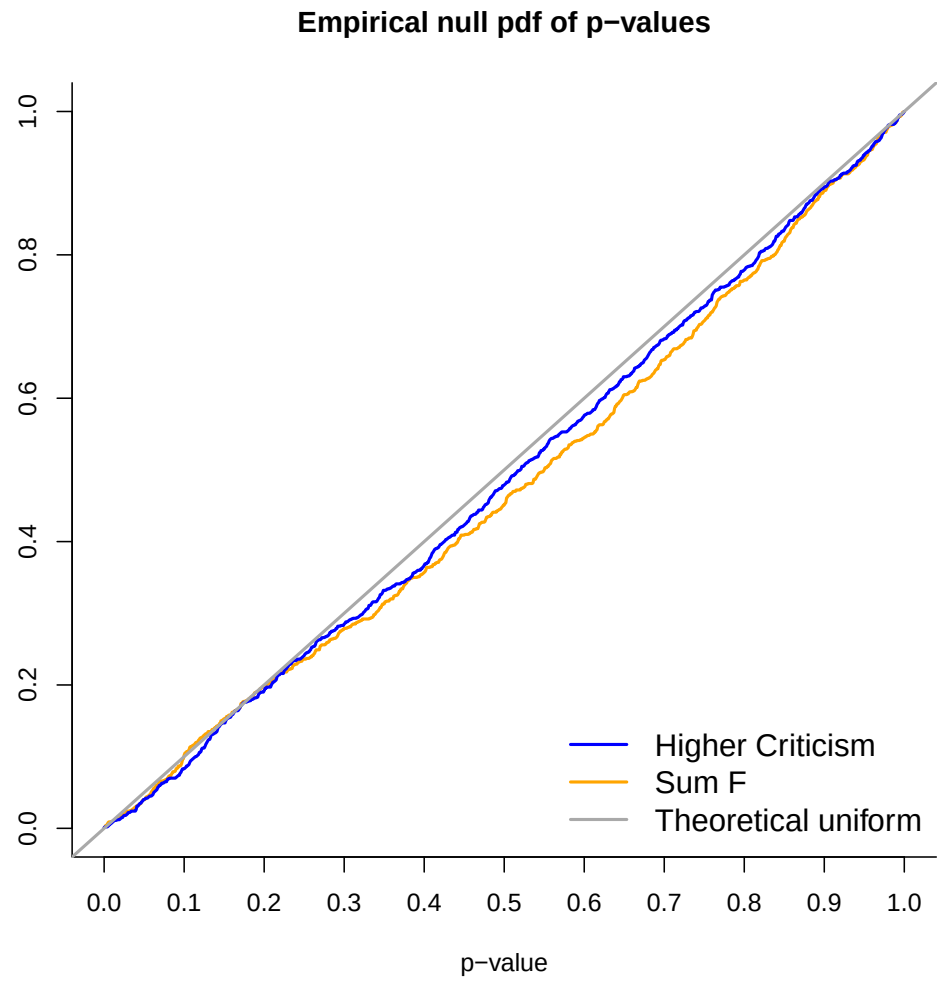


FIGURE 4.9 – Fonction de répartition empirique sous l'hypothèse nulle des probabilités critiques associées aux statistiques HC et F_s sous l'hypothèse d'une structure de corrélation auto-régressive

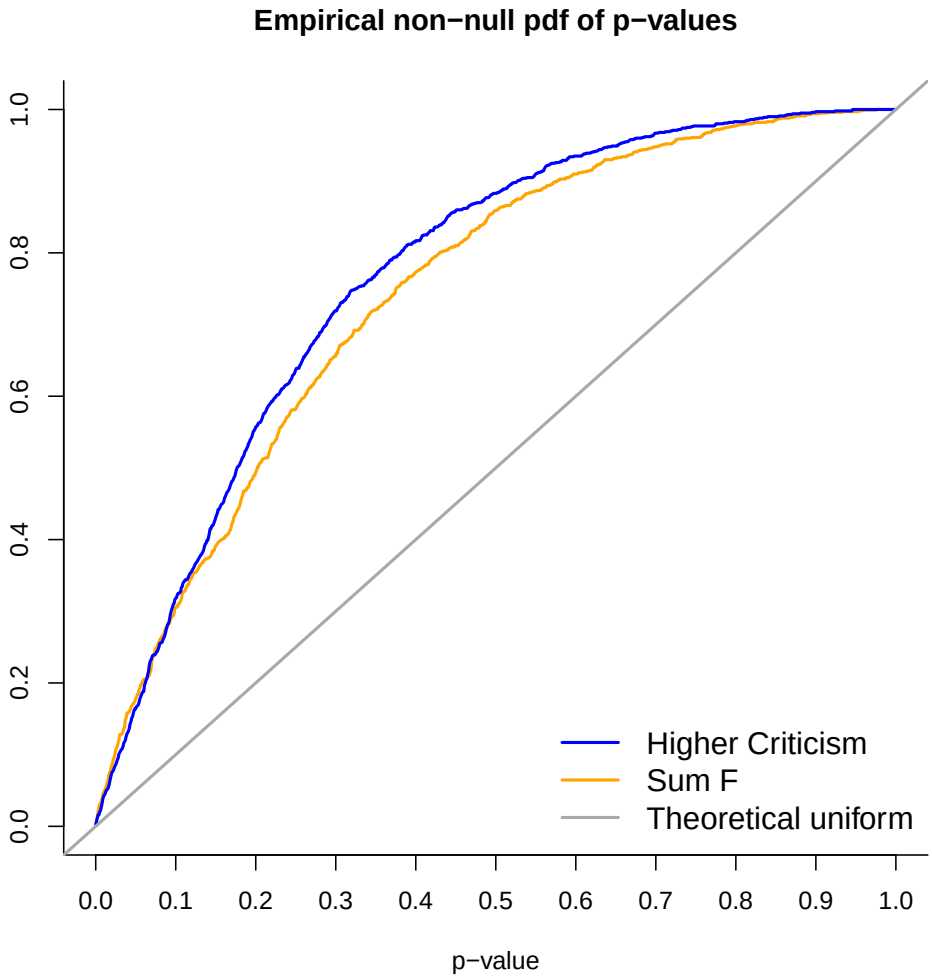


FIGURE 4.10 – Fonction de répartition empirique sous l'alternative des probabilités critiques associées aux statistiques HC et F_s sous l'hypothèse d'une structure de corrélation auto-régressive

Ainsi, les variantes de la méthode HC diffèrent aussi sur la définition du sous-ensemble $\mathcal{I}$: pour Donoho and Jin (2004) et Donoho and Jin (2008), $\mathcal{I} = \{t, 1/T \leq t/T \leq \alpha_0\}$ où $\alpha_0 \in [0,1]$ est une proportion fixée des plus petites p-valeurs. En revanche, Hall and Jin (2010) proposent de ne pas prendre en compte les p-valeurs extrêmement faibles en posant $\mathcal{I} = \{t, 1/T \leq p_{(t)} \leq \alpha_0\}$. Enfin, Klaus and Strimmer (2013) proposent une maximization globale sur $\mathcal{I} = \{1, \ldots, T\}$. Dans les présentes simulations sur la statistique HC, on pose $\alpha_0 = 0.1$ comme recommandé par Donoho and Jin (2004).

Dans la suite de ce chapitre, le terme "HCT standard" fait référence au critère suivant :

$$HC^* = \max_{1/T \leq t/T \leq \alpha_0} \sqrt{T} \frac{t/T - p_{(t)}}{\sqrt{t/T(1-t/T)}}, \tag{4.14}$$

qui semble être le plus utilisé dans la littérature.

Dans un contexte de sélection de variables pour la classification supervisée, Donoho and Jin (2008) soulignent le lien entre la maximisation de la fonction HC et la minimisation de l'erreur de classement d'une règle de classification construite à partir des variables sélectionnées de la façon suivante : si le maximum HC^* de la fonction HC est atteint à l'indice t^*, alors le sous-ensemble des variables sélectionnées est l'ensemble $\{t, p_t \leq p_{(t^*)}\}$. Donoho and Jin (2008) illustrent que le Higher Criticism Thresholding (HCT) ainsi défini surpasse les méthodes de seuillage fondées sur le contrôle du taux de faux positifs (FDR). Cependant, Klaus and Strimmer (2013) montrent l'équivalence entre la méthode HCT et la définition d'un seuil sur le FDR local associé à chaque probabilité critique.

3.2 HCT en situation de dépendance

La propriété d'optimalité du HCT est connue pour être robuste dans les cas de faible dépendance. Cependant, certains auteurs suggèrent que cette procédure peut être améliorée en prenant en compte la dépendance de manière explicite. Les variantes de HCT prenant en compte la dépendance sont principalement fondées sur une étape de décorrélation des statistiques de test. Initialement proposés par Zuber and Strimmer (2009) dans un contexte de comparaisons multiples mais aussi appliqués à la sélection de variables en classification supervisée dans Ahdesmäki and Strimmer (2010), les *Correlation-Adjusted T-scores* sont tels que :

$$\tau^{adj} = R^{-1/2}\tau,$$

où τ est le vecteur des statistiques de test standard. La matrice de corrélation R est estimée par un estimateur de type James-Stein $R_\gamma = \gamma \mathbb{I}_m + (1-\gamma)R$, proposé par Schäfer and Strimmer (2005) dont le paramètre γ est estimé analytiquement. La racine de l'inverse de la matrice R_γ est numériquement calculable en grande dimension en observant que la matrice $Z = \frac{1}{\gamma} R_\gamma$ se décompose en $Z = \mathbb{I}_m + UMU'$ où M est une matrice symétrique définie positive et U est une base orthonormale. Ainsi,

$$Z^\alpha = \mathbb{I}_m - U(\mathbb{I}_r - (\mathbb{I}_r + M)^\alpha)U' \tag{4.15}$$

où r est le rang de la matrice de corrélation empirique. Cet estimateur est efficace en pratique, en particulier pour des applications en génomique, mais semble échouer à capter la structure de dépendance de données ERP. En effet, le graphique en bas de la Figure 4.11 présente une image de la matrice de corrélation résiduelle initialement observée sur les données ERP et présentée Figure 4.3 et estimée par la méthode *shrinkage*. On remarque que les blocs de corrélation sont sous-estimés.

De même, Hall and Jin (2010) proposent la méthode *innovated* HCT (iHCT) fondée sur la décorrélation des composantes du vecteur $\mathcal{T}$. Pour cela, les auteurs introduisent une matrice U_m telle que $U_m R U_m' = \mathbb{I}_m$ et appliquent la méthode HCT standard à ces statistiques de test décorrélées, à la façon de Zuber and Strimmer (2009). Hall and Jin (2010) recommandent l'usage de la factorization de Cholesky inverse de la matrice R, instable en grande dimension. Dans un contexte de dépendance auto-régressive, le support du signal est légèrement décalé par cette transformation linéaire et les auteurs proposent de rétablir le support par lissage de la matrice U_m.

4 Factor innovated Higher Criticism Thresholding

La méthode proposée est conceptuellement similaire aux méthodes iHCT et CAT-scores mais la prise en compte de la dépendance est fondée sur une hypothèse de décomposition en facteurs de la structure de dépendance.

4.1 Décorrélation par des facteurs latents

Présenté en détails dans le Chapitre 1, on rappelle que le modèle à facteurs suppose l'indépendance conditionnelle des variables sachant un vecteur de facteurs latents. En effet, conditionnellement aux facteurs, l'Expression (4.1) s'écrit :

$$Y_{ijt} \quad = \quad x_{0ij}'\mu_t + a_{it} + b_t' z_{ij} + e_{ijt}, \tag{4.16}$$

où $z_{ij} = (z_{ij1}, \ldots, z_{ijq})$ est un vecteur gaussien de moyenne 0_q et de variance $\mathbb{I}_q$ et les termes d'erreur $e_{ij} = (e_{ij,t_1}, \ldots, e_{ij,t_T})'$ sont indépendants tels que $e_{ij} \sim \mathcal{N}_T(0_T, \Psi_\Sigma)$ avec Ψ_Σ une matrice diagonale de variances spécifiques. q est le nombre de facteurs, $q \ll T$. Les algorithmes proposés dans de nombreux articles (voir Friguet et al. (2009); Causeur et al. (2012); Sun et al. (2012); Storey et al. (2007); Perthame et al. (2015) et les Chapitres 2 et 3 de ce manuscrit) sont conçus pour l'estimation des facteurs latents afin d'ajuster les données sur leur effet. Dans ce chapitre, on s'intéresse plus particulièrement à une autre implication induite par une structure en facteurs de faible rang pour la matrice de covariance Σ. En effet, le modèle à facteurs suppose de manière équivalente que la matrice de covariance admet une décomposition de la forme :

$$\Sigma \quad = \quad \Psi_\Sigma + B_\Sigma B_\Sigma',$$

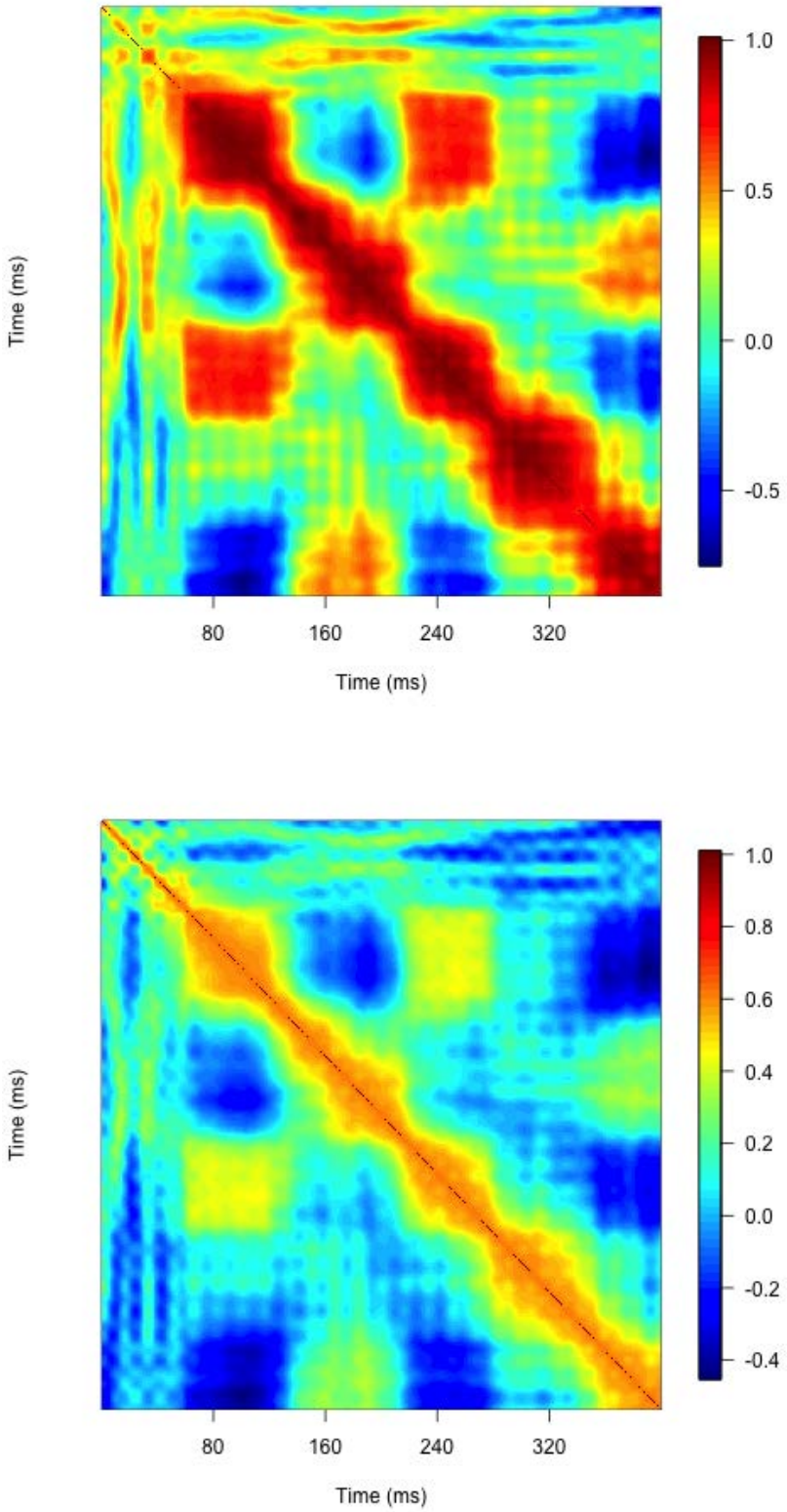


FIGURE 4.11 – Estimation de la matrice de corrélation par l'algorithme AFA (Sheu et al. (2015)) sous l'hypothèse d'un modèle à 6 facteurs (haut) et par la méthode *shrinkage* proposée par Schäfer and Strimmer (2005) (bas)

où B_Σ est une matrice $T \times q$ de *loadings* décrivant la dépendance commune partagée par les covariables et $\Psi_\Sigma = \text{diag}(\Psi_{\Sigma,1}, \ldots, \Psi_{\Sigma,T})$ est la matrice diagonale des variances spécifiques. La matrice de corrélation résiduelle R admet une décomposition similaire :

$$R \quad = \quad \Psi + BB',$$

où $\Psi_\Sigma = D_{\sigma^2}\Psi$, $\Psi = \text{diag}(\Psi_1, \ldots, \Psi_\text{T})$ et $B_\Sigma = D_\sigma B$.

Les paramètres (Ψ, B) sont estimés par l'algorithme AFA proposé dans le Chapitre 2. On rappelle que cet algorithme fournit une estimation adaptative de la structure de dépendance d'un modèle à facteurs en se fondant sur une estimation jointe du signal et de la matrice de covariance. De plus, cette méthode est construite pour tirer avantage de la connaissance a priori d'intervalles de temps durant lesquels le signal est nul, afin de débruiter le signal estimé par la méthode des moindres carrés. Le nombre de facteurs q est estimé par minimisation du critère d'inflation de la variance proposé par Friguet et al. (2009).

Le graphique en faut de la Figure 4.11 montre que la modélisation par une structure en facteurs est suffisamment flexible pour reproduire la structure de dépendance observée sur les données de *oddball*. Cette figure montre une image de l'estimation de la matrice de corrélation par un modèle à 6 facteurs.

4.2 Limites de détection

Le calcul de bornes de détection exactes pour une structure de dépendance quelconque est impossible car la dépendance temporelle peut être non homogène et générer des conditions de détectabilité et d'estimabilité du signal différentes d'une variable à l'autre. A ce propos, Hall and Jin (2010) ne donnent pas d'expression directe d'un diagramme de phase en cas de dépendance mais ils explicitent des bornes inférieures et supérieures pour les limites de détection sous certains scenario de dépendance. On propose dans cette section d'étendre le diagramme de phase présenté dans la Section 3 au cas où la dépendance possède une structure en facteurs.

Si les statistiques de test sont des transformations linéaires des covariables, ce qui est généralement le cas pour un test de Student par exemple, la structure de corrélation des statistiques de test est directement héritée de celle des corrélations résiduelles des covariables. Si $\mathcal{Z}$ désigne la même transformation linéaire appliquée à la matrice Z $(n \times q)$ donc les lignes sont les composantes des facteurs latents $z_{i,j}$, alors le modèle (4.13) devient :

$$\mathcal{T} \quad = \quad \mu + B\mathcal{Z}' + E \ , \ E \sim \mathcal{N}(0, \Psi).$$

Et de manière équivalente :

$$\Psi^{-1/2}(\mathcal{T} - B\mathcal{Z}') \quad = \quad \Psi^{-1/2}\mu + E^* \ , \ E^* \sim \mathcal{N}(0, \mathbb{I}_T). \tag{4.17}$$

L'Equation (4.17) offre un cadre RW similaire à celui du modèle (4.13), dans lequel les statistiques de test décorrélées $\Psi^{-1/2}(\mathcal{T} - B\mathcal{Z}')$ sont indépendantes de variance

1, conditionnellement aux facteurs $\mathcal{Z}$. Des bornes inférieure et supérieure pour la limite de détection d'un signal d'amplitude $\Psi^{-1/2}\mu$ et une proportion ε_T de coordonnées non nulles sont déduites. La répartition des variances spécifiques Ψ_t varie entre les variables. Ainsi, l'amplitude du signal transformé $\Psi^{-1/2}\mu$ prend des valeurs entre $\underline{\gamma_0} = A_T/\sqrt{\Psi_{\max}} = \sqrt{2\underline{r}\log(T)}$ et $\overline{\gamma_0} = A_T/\sqrt{\Psi_{\min}} = \sqrt{2\overline{r}\log(T)}$, où $\Psi_{\min} = \min_t(\Psi_t)$, $\Psi_{\max} = \max_t(\Psi_t)$, $\underline{r} = r/\Psi_{\min}$ et $\overline{r} = r/\Psi_{\max}$. Les conditions de détectabilité sur (r, β) établies sous l'indépendance peuvent être appliquées à deux valeurs différentes $\underline{r}$ et $\overline{r}$ du paramètre de force du signal r, ce qui permet de déduire des bornes inférieure et supérieure de détectabilité. Si r vérifie la condition suivante :

$$r > \Psi_{\max}\rho_D^*(\beta), \tag{4.18}$$

alors tout le signal est détectable. De plus, si r ne vérifie pas la condition suivante :

$$r > \Psi_{\min}\rho_D^*(\beta), \tag{4.19}$$

alors le signal ne peut pas être détecté. Entre ces deux bornes, la détection du signal est incertaine et dépend de la correspondance entre le profil des variances spécifiques et le support du signal.

Le même raisonnement sur l'estimabilité donne des résultats de forme similaire pour les limites d'estimabilité. Enfin, on propose de définir un modèle *Factor Rare and Weak* dans lequel la parcimonie du signal est caractérisée par :

$$\varepsilon_T \quad = \quad T^{-\beta} \text{ avec } \beta \in (1/2, 1),$$

et la force du signal par :

$$A_T \quad = \quad \sqrt{2rlog(T)} \text{ avec } 0 < r < \Psi_{max}.$$

Sous l'indépendance, la force du signal est fixée à $\sqrt{2rlog(T)}$ avec $0 < r < 1$ afin de rendre le problème de détection difficile. En effet, $\sqrt{2log(T)}$ est l'espérance du maximum de T variables aléatoires normales centrées réduites indépendantes. On note que les signaux d'amplitude $\sqrt{2rlog(T)}$ avec $\Psi_{max} < r \leq 1$ sont considérés comme faibles sous l'indépendance mais pas en situation de dépendance. Cette observation est cohérente avec celle faite par Hall and Jin (2010). Les auteurs expliquent que les designs corrélés sont en fait des situations de détection d'un signal favorables pour les méthodes prenant en compte la dépendance. Ceci est aussi illustré par la Figure 4.12, présentant les diagrammes de phase de détectabilité et d'estimabilité calculés sur les variances spécifiques estimées sur les données de *oddball*. En effet, ces graphiques confirment que les régions de détection et d'estimabilité sont plus larges en cas de dépendance.

Les bornes des Expressions (4.18) et (4.19) sont cohérentes avec celles établies par Hall and Jin (2010). En effet, les Théorèmes 3.1 et 4.2 de cet article donnent les mêmes expressions pour les limites de détection, où $\underline{\gamma}_0$ et $\overline{\gamma}_0$ sont tels que :

$$\underline{\gamma_0} = \lim \inf_{T\to+\infty} \max_{\sqrt{T}\leq k\leq T-\sqrt{T}} R_{kk}^{-1},$$
$$\overline{\gamma_0} = \lim \sup_{T\to+\infty} \max_{\sqrt{T}\leq k\leq T-\sqrt{T}} R_{kk}^{-1},$$

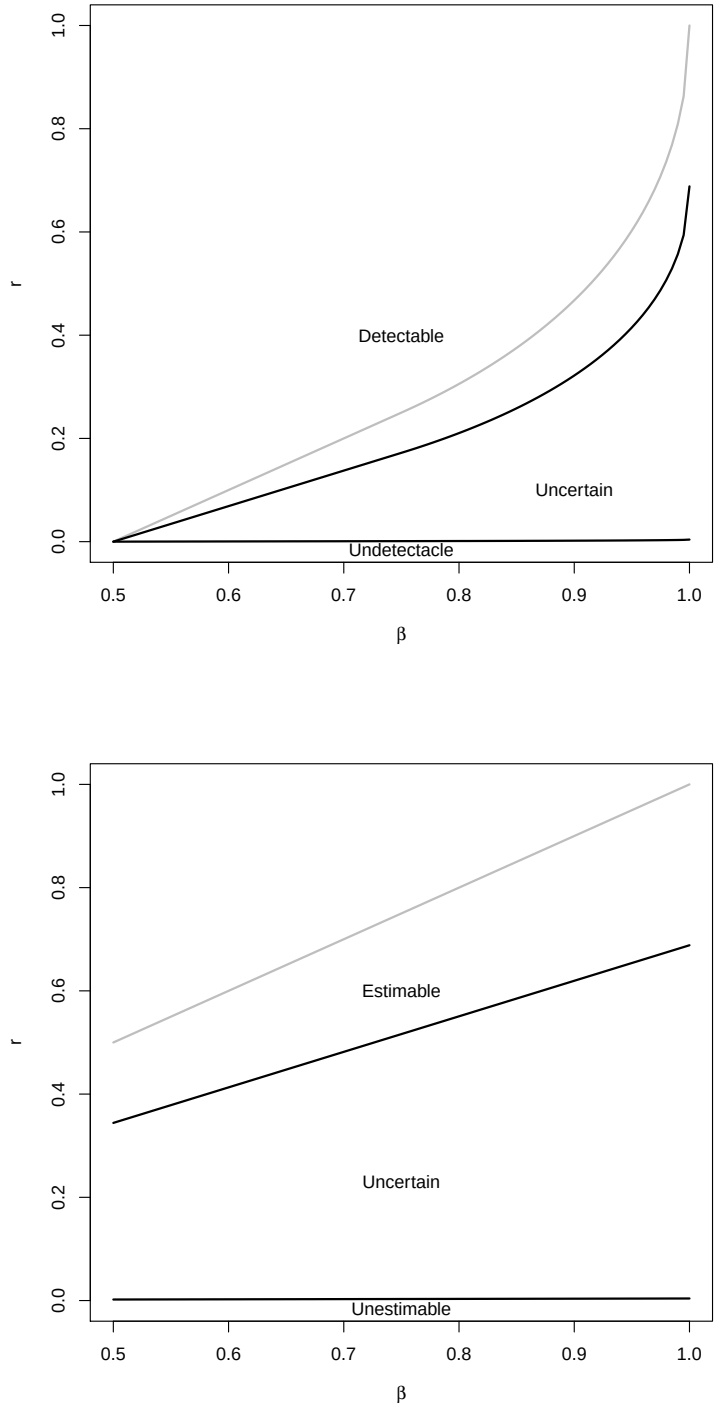


FIGURE 4.12 – Limites de détection (haut) et d'identification (bas) pour les données de *oddball*. Les lignes grises représentent les limites sous l'indépendance et les lignes noires représentent les bornes supérieure et inférieure de détectabilité (haut) et d'estimabilité (bas) pour un signal en situation de dépendance.

où $(R_{11}, \ldots, R_{kk}, \ldots, R_{TT})$ désigne la diagonale de la matrice R. Sous un modèle à facteurs, on obtient les bornes suivantes :

$$\underline{\gamma_0} = \frac{1}{\max(\Psi)} \quad \leq \quad \mathrm{diag}(R^{-1}) \leq \frac{1}{\min(\Psi)} = \overline{\gamma_0},$$

ce qui donne les Expressions (4.18) et (4.19).

4.3 Factor innovated HCT

De manière similaire à Zuber and Strimmer (2009) et Hall and Jin (2010), on propose un HCT décorrélé (F-iHCT) en appliquant la procédure HCT standard aux statistiques de test décorrélées. Sous l'hypothèse d'un modèle à facteurs RW (défini ci-dessus), on définit la racine L de la matrice de corrélation R telle que $R = LL'$:

$$L \quad = \quad (\mathbb{I}_m - U[(\mathbb{I}_q + [\mathbb{I}_q + D^2]^{1/2})^{-1} + \mathbb{I}_q]^{-1}U')\Psi^{-1/2},$$

où U et D sont déduits de la décomposition en valeurs singulières des loadings standardisés $\Psi^{-1/2}B = UDV$. On remarque que cette formule ne fait appel qu'à l'inversion et au calcul de racine de matrices diagonales. On définit finalement les statistiques de test décorrélées :

$$\mathcal{T}^* \quad = \quad L'\mathcal{T}$$

et les probabilités critiques associées p_t^*. La procédure HCT est ensuite appliquée à la collection des p-valeurs $p^* = (p_1^*, \ldots, p_T^*)$.

5 Etude par simulations et analyse de données réelles

Les performances de la méthode proposée sont maintenant comparées à celles du HCT standard et à d'autres méthodes de décorrélation par les innovations ou par ajustement sur l'effet de facteurs latents. La comparaison est faite au travers d'une étude par simulations et d'une application aux données issues de l'expérience de *oddball*.

5.1 Etude par simulations

Plan de simulations Les performances de F-iHCT sont étudiées sur des simulations. 1000 jeux de données de dimensions 30×799 sont générés selon une loi normale multivariée. La structure de corrélation et les variances sont estimées sur les données d'ERP auditives présentées dans la Section 2 (voir Figure 4.3 et Figure 4.2). Ce plan de simulation imite les dimensions et la structure de covariance des données observées durant l'expérience de *oddball*, excepté que le vrai signal est connu. Chaque jeu de données est divisé en deux groupes équilibrés. La loi normale est d'espérance nulle pour les 15 premiers sujets (groupe 1). L'espérance pour les 15 derniers sujets (groupe 2) est représentée Figure 4.13. La courbe de différence

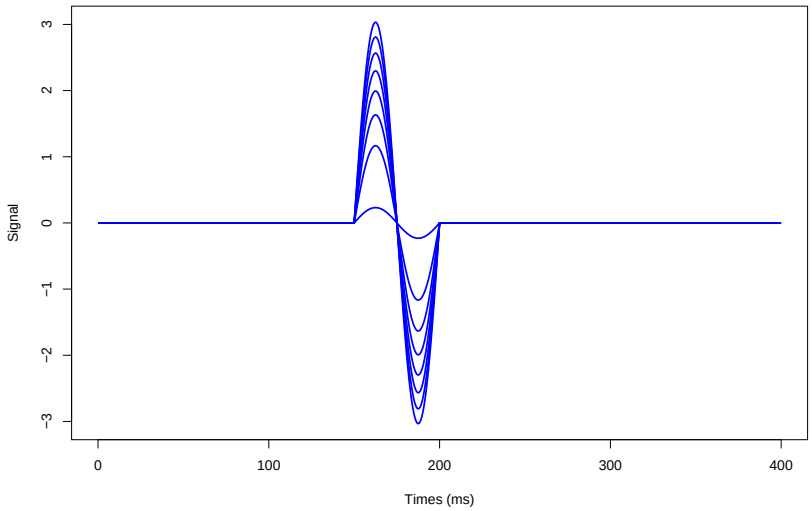


FIGURE 4.13 – Etude par simulations - Amplitude du signal au cours du temps

de moyennes entre les deux groupes a donc la forme d'une onde dont l'amplitude varie et les indices des variables sous l'alternative appartiennent à $[150\text{ms}, 200\text{ms}]$. 1 000 jeux de données d'apprentissage sont générés pour chaque force de signal décrites ci-après. Respectivement, huit jeux de données test de dimension 1000×799 équilibrés en 2 groupes sont générés selon le même plan de simulation dans un but de prédiction de la classe (groupe 1 ou 2). Les paramètres du modèle RW pour ce plan de simulation sont tels que $\varepsilon_T = 12\%$ et l'amplitude maximum des 8 signaux générés s'écrit $A_T = \sqrt{2r\log(T)}$ avec r prenant 8 valeurs uniformément réparties dans $[0.004, 0.688]$. Selon le contexte du modèle RW, cette combinaison de r et β caractérise un signal peu parcimonieux, de force faible (limite théorique de détectabilité d'une partie du signal) à forte (limite de détectabilité de la totalité du signal).

Méthodes On compare, dans cette étude par simulations, quatre méthodes décrites dans ce paragraphe. A la manière de Donoho and Jin (2008) et Donoho and Jin (2009), l'étape de sélection de variables par différentes versions de la méthode HCT est suivie d'une étape de classification supervisée par Analyse Diagonale Discriminante (DDA) réalisée sur le sous-ensemble des variables sélectionnées. On compare les méthodes suivantes :

1. Sélection de variables par HCT standard sur les probabilités critiques brutes, classification par la règle de classification naïve de Bayes (voir Bickel and Levina (2004)) et désigné par standard HCT ;
2. Sélection de variables par HCT appliqué aux CAT-scores (Zuber and Strimmer (2009); Ahdesmäki and Strimmer (2010)), classification par Shrinkage Discriminant Analysis (SDA, voir Ahdesmäki and Strimmer (2010)) diagonale et désigné par *CAT-scores* ;

3. Sélection de variables par Factor-innovated HCT, classification par règle de classification de Bayes conditionnelle (proposée dans le Chapitre 3) et désigné par *F-iHCT*;
4. Sélection de variables par HCT standard appliqué aux probabilités critiques ajustées sur l'effet des facteurs latents retournées par la procédure AFA (proposée dans le Chapitre 2), classification par règle de classification de Bayes conditionnelle et désigné par *AFA*.

Les performances des procédures sont comparées en terme de proportion de signal découvert (précision), proportion de faux positifs (FDR), nombre de variables sélectionnées et erreur de prédiction. Pour chaque jeu de données, les étapes de sélection de variables et l'estimation de la règle de classification sont réalisées sur les données d'apprentissage (en incluant l'optimisation des méta-paramètres) et l'erreur de classification est calculée sur les données test.

Résultats Lorsque le paramètre α_0 est correctement spécifié par rapport à la proportion d'hypothèse nulle ($\alpha_0 = 0.1$), on note sur la Figure 4.14 que la sélection de variables par CAT-scores semble être la méthode la plus efficace pour identifier un signal faible. En effet, cette méthode atteint à la fois le FDR le plus faible et une précision la plus élevée pour les signaux de faible amplitude. Même si la méthodes des CAT-scores n'atteint pas les meilleures performances pour les signaux forts, le FDR, la précision et le nombre de variables sélectionnées restent très stables. La méthode HCT standard semble robuste à la dépendance car elle atteint de bonnes performances en terme de FDR mais sa précision est faible par rapport aux méthodes basées sur de la décorrélation. De plus, le nombre de variables sélectionnées est faible ce qui suggère que la méthode HCT est conservative en situation de dépendance. Enfin, la classification par le classifieur naïf de Bayes atteint les taux d'erreur de prédiction les plus élevés pour les signaux faibles à modérés. Les procédures de sélection et de classification fondées sur un modèle à facteurs (AFA et F-iHCT) fournissent les meilleurs résultats à la fois en terme de faux positifs, recouvrement du signal et erreur de classement. Le FDR est faible pour les signaux modérés à forts et une puissance de détection correcte est atteinte.

On remarque sur la Figure 4.15 que toutes les méthodes sont affectées par une mauvaise spécification de la valeur du paramètre α_0 ($\alpha_0 = 0.5$). La méthode des CAT-scores sélectionne trop de variables et atteint donc une bonne précision au prix d'un FDR élevé. Les méthodes F-iHCT et AFA sont performantes en terme de classification malgré le nombre de faux positifs. L'erreur de classement et la précision de la méthode HCT standard sont aussi altérés par la valeur de α_0.

5.2 Application aux potentiels évoqués

Les 4 méthodes comparées dans l'étude par simulations sont appliquées aux données de *oddball* présentées dans la Section 2. On rappelle que l'objectif de cette expérience est de prédire la classe d'un nouvel individu à partir de ses courbes ERP. Pour chaque méthode, le nombre de variables sélectionnées et l'erreur de prédiction sont calculés. Le nombre d'observations étant faible (30 observations), l'erreur de classement est estimée par validation-croisée (CV) par la technique leave-one-out.

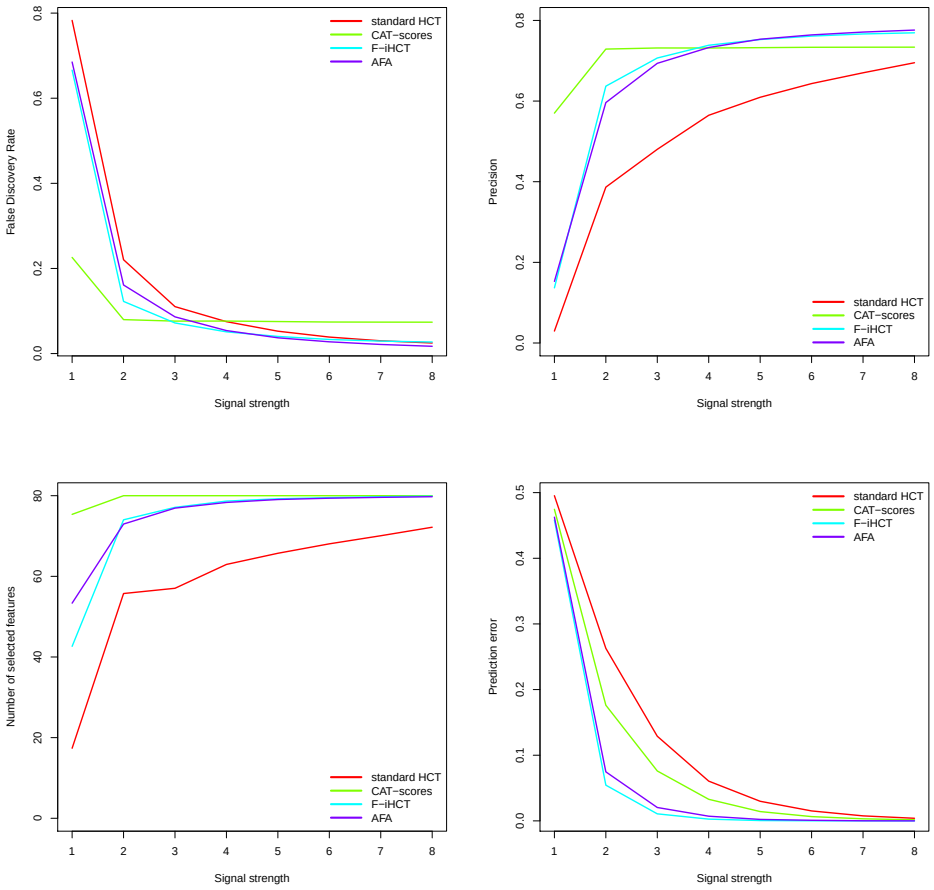


FIGURE 4.14 – Résultats de l'étude par simulations en fonction de la force du signal et pour $\alpha_0 = 0.1$: taux de faux positifs (haut/gauche), précision (haut/droit), nombre d'instants sélectionnés (bas/gauche), erreur de prédiction (bas/droit)

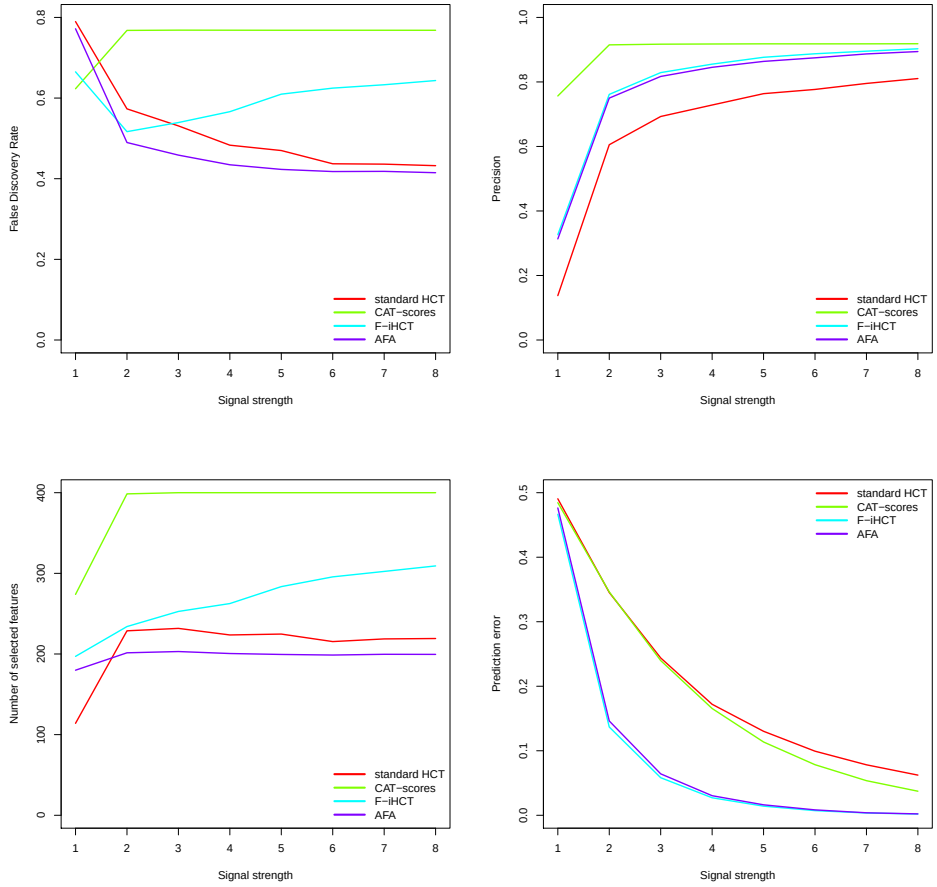


FIGURE 4.15 – Résultats de l'étude par simulations en fonction de la force du signal et pour $\alpha_0 = 0.5$: taux de faux positifs (haut/gauche), précision (haut/droit), nombre d'instants sélectionnés (bas/gauche), erreur de prédiction (bas/droit)

TABLE 4.1 – Etude de données réelles - Nombre d'instants sélectionnés pour l'expérience d'ERP auditive

α_0	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50
Nb. de variables	40	80	120	160	200	240	280	320	360	400

La Table 4.1 présente le nombre de variables sélectionnées par les 4 méthodes comparées, pour différentes valeurs du paramètre α_0. Quelle que soit la valeur de α_0, les 4 méthodes sélectionnent le même nombre de variables. La différence entre les procédures réside dans les indices des temps sélectionnés : on remarque sur la Figure 4.17 que les méthodes n'identifient pas les mêmes intervalles de temps.

La Figure 4.16 présente les taux d'erreur de prédiction calculés par validation croisée, pour différentes valeurs de α_0. Pour les valeurs de α_0 supérieures à 0.15, la méthode HCT standard est stable et plutôt performante. En ce qui concerne les modèles plus parcimonieux, la procédure HCT standard atteint des taux d'erreur plus élevés et est amélioré par les méthodes de décorrélation fondées sur une hypothèse de structure en facteurs de la dépendance (AFA et F-iHCT). Les performances de la méthode CAT-scores varient peu en fonction de α_0. Les courbes des procédures F-iHCT et AFA sont irrégulières pour les valeurs de α_0 inférieures à 0.125 mais se stabilisent lorsque α_0 augmente. Pour les valeurs de α_0 supérieures à 0.275, les procédures AFA et F-iHCT classent parfaitement les données et semblent donc être les plus efficaces. Néanmoins, on peut remarquer que pour un taux d'erreur égal, les deux méthodes ne sélectionnent pas les mêmes intervalles de temps, comme le montre le graphique en bas de la Figure 4.17.

La Figure 4.17 présente la courbe de différence de moyennes entre les deux groupes et les instants sélectionnés par les 4 méthodes comparées pour $\alpha_0 = 0.125$ (haut) et $\alpha_0 = 0.275$ (bas). Le choix de ces valeurs de α_0 repose sur le fait qu'elles fournissent des modèles de niveau de parcimonie différent pour des erreurs de classement comparables. Comme attendu, les temps après 300ms sont sélectionnés par toutes les méthodes, ce qui est cohérent avec la littérature. De plus, des instants autour de 100ms apparaissent aussi significatifs.

Enfin, cette application aux données réelles de potentiels évoqués souligne l'importance du choix de α_0 dans la procédure HCT. En effet, on remarque que ce paramètre a un impact important sur la parcimonie et les erreurs de classification des modèles, pour toutes les méthodes comparées. La procédure AFA, fondée sur la décorrélation par ajustement des covariables sur l'effet de facteurs latents semble être la méthode la mieux adaptée à cet exemple, même si la méthode F-iHCT atteint aussi d'intéressants taux de classification.

6 DISCUSSION ET CONCLUSION

Ce chapitre aborde le problème de la détection et de l'identification d'un signal dans le cadre de données de potentiels évoqués (ERP). Ce travail est motivé par

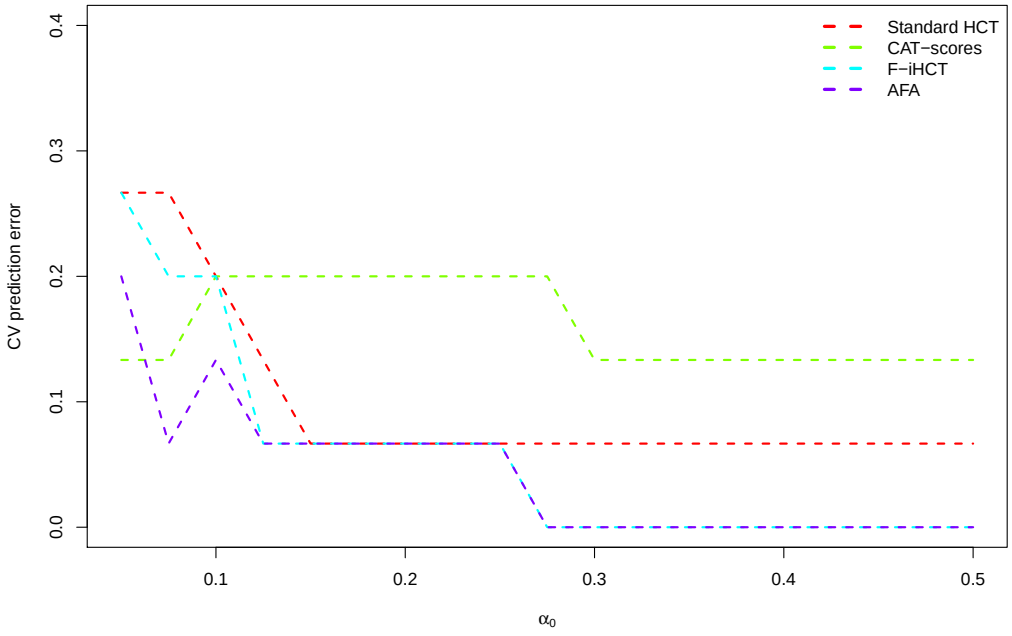


FIGURE 4.16 – Etude de données réelles - Erreur de prédiction calculée par validation croisée pour les méthodes HCT standard, HCT appliqué aux CAT-scores, factor innovated HCT et HCT appliqué aux probabilités critiques renvoyées par la procédure AFA pour l'expérience d'ERP et pour différentes valeurs de α_0

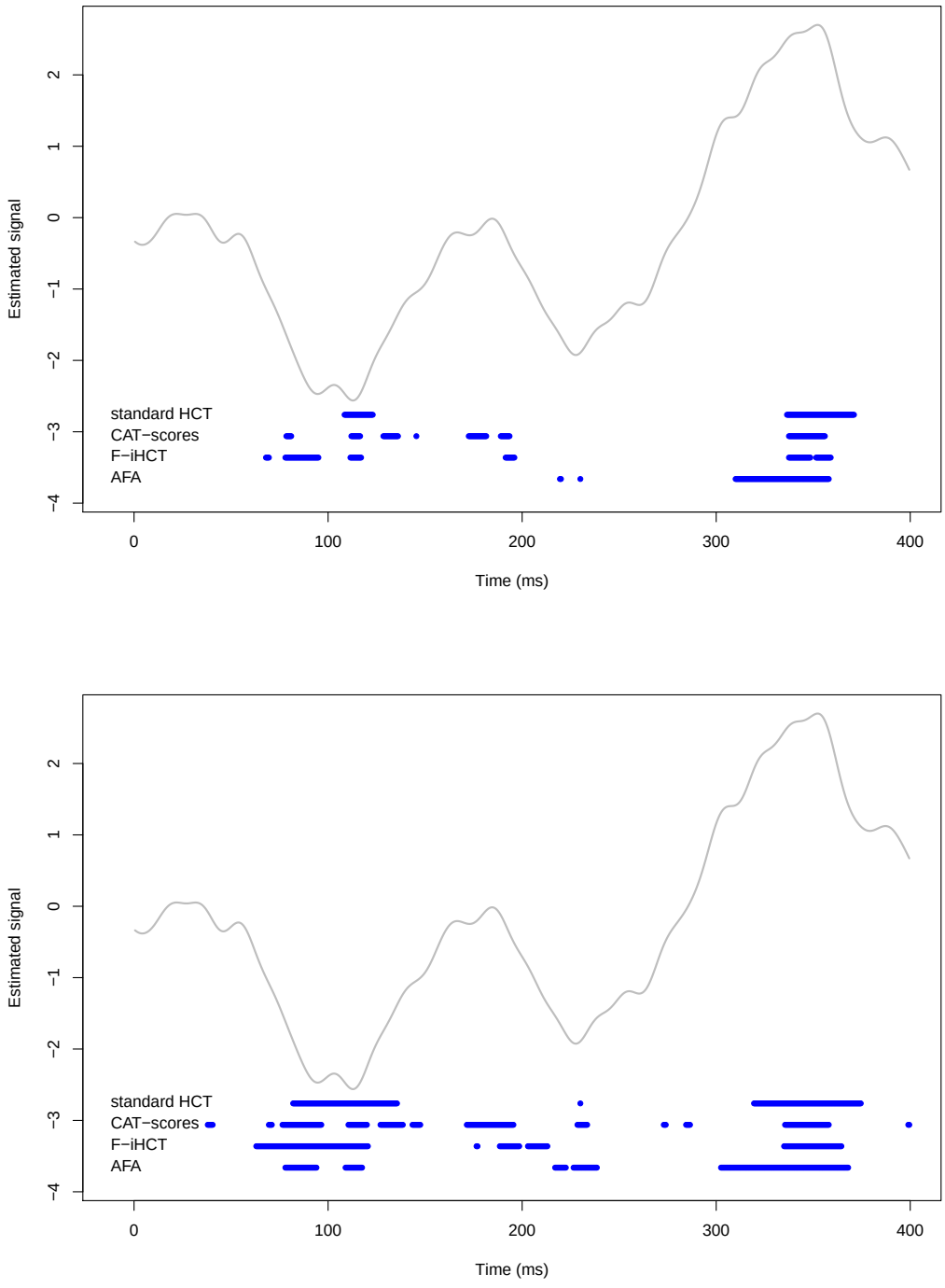


FIGURE 4.17 – Etude de données réelles - Estimation du signal (ligne grise) et instants déclarés significatifs (points bleus) par HCT standard, HCT appliqué aux CAT-scores, factor innovated HCT et HCT appliqué aux probabilités critiques renvoyées par la procédure AFA pour $\alpha_0 = 0.125$ (haut) et $\alpha_0 = 0.275$ (bas)

une étude sur des courbes ERP mesurées dans le cadre d'une expérience de *oddball* durant laquelle deux classes de stimuli sont présentées au sujet, l'une survenant fréquemment (standard) et l'autre survenant rarement (cible). Le but est d'identifier des intervalles de temps sur les courbes ERP susceptibles d'être des marqueurs de la différence entre l'occurrence rare et l'occurence fréquente et de proposer une règle de prédiction de la fréquence entendue. Lorsque les statistiques de sélection sont indépendantes, la méthode du Higher Criticism est connue pour être une procédure optimale de détection d'un signal, sous les hypothèses d'un modèle *Rare-and-Weak* (RW), initialement introduit par Donoho and Jin (2004). De plus, la procédure Higher Criticism Thresholding (HCT) est performante pour estimer le support du signal. Le modèle RW est théoriquement adapté à un problème de détection d'un signal ERP. Cependant, les statistiques de sélection déduites d'un problème d'identification d'un signal ERP sont caractérisées par une structure de dépendance forte et complexe.

Lors de l'analyse de données ERP, la détection d'un signal est habituellement traitée par un test global de Fisher de la nullité du signal sur toute la durée de l'expérience. Sous les hypothèses d'indépendance et d'hétéroscédasticité, on montre que la statistique de Fisher calculée par la méthode des moindres carrés généralisés (GLS) s'exprime comme une somme de tests de Fisher individuels pour chaque variable. Ceci permet d'obtenir une expression explicite de la distribution de la statistique sous l'hypothèse nulle. De même, si la dépendance est structurée par de l'auto-corrélation d'ordre 1, on montre que la test de Fisher fondé sur les GLS peut aussi être implémenté, à condition que la distribution sous l'hypothèse nulle de la statistique de test prenne en compte l'auto-corrélation entre les statistiques de test individuelles. Dans une étude par simulations, où la dépendance entre les variables est structurée par une auto-corrélation élevée, on illustre que ce test contrôle l'erreur de type I, ce qui n'est pas le cas pour les autres statistiques de Fisher telles que celles obtenues par des approches de type analyse de la variance fonctionnelle. De plus, la méthode de détection par Higher Criticism semble robuste à cette forte auto-corrélation.

La variante de la procédure HCT proposée dans ce chapitre tire parti de l'hypothèse d'un modèle à facteurs pour les covariables pour décorréler les statistiques de test. En effet, ce cadre théorique fournit des outils algébriques permettant de calculer une racine de l'inverse de la matrice de corrélation, impliquée dans le calcul des innovations. De plus, à la manière de Ingster (1997) et Donoho and Jin (2004), un diagramme de phase sous une structure de dépendance générale est déduit et la procédure Factor innovated HCT, un HCT décorrélé fondé sur les innovations, est proposée.

Les performances de cette méthode sont évaluées par une étude sur simulations plus intensive, dans laquelle la structure de dépendance reproduit les corrélations observées sur les données de *oddball*. Cette étude montre que les méthodes fondées sur une décorrélation de la procédure HCT sous l'hypothèse d'un modèle à facteurs pour les covariables atteignent de bonnes performances en terme de sélection et de classification. En effet, les procédures F-iHCT et AFA, fondée sur un ajustement des données sur l'effet de facteurs latents, obtiennent des résultats similaires. Enfin, les

méthodes comparées (standard HCT, CAT-scores, F-iHCT et AFA) sont appliquées aux données de *oddball* et sélectionnent des instants autour de 300ms, comme attendu par les psychologues. Les erreurs de classement par validation croisée les plus faibles sont atteintes par les deux méthodes fondées sur une hypothèse de décomposition en facteurs de la dépendance. L'application sur des simulations et sur des données réelles révèlent la sensibilité de la procédure HCT au choix de la valeur du paramètre α_0, déterminant la parcimonie du modèle.

Il serait intéressant d'exploiter ces premiers résultats dans le cadre de l'identification de signal pour établir une stratégie de détection d'un signal optimale pour les données ERP. En effet, si la distribution sous l'hypothèse nulle des statistiques de test F-iHC peut être estimée, ce nouveau test pourrait être comparé à la classe des tests de Fisher.

CHAPITRE 5

CONCLUSION

L'objectif de ce travail de thèse est de contribuer à l'amélioration des méthodes d'analyse de données à haut débit, en particulier celles dédiées à la sélection de variables pour la régression et la classification supervisée, par une meilleure prise en compte de la dépendance entre les statistiques de sélection. Pour l'essentiel, les approches développées dans cette thèse ont été motivées par le souci d'apporter des réponses à des problématiques de détection et d'identification d'un signal dans des données d'activité cérébrale mesurée par électroencéphalogramme (EEG), en réaction à un stimulus cognitif ; on parle de potentiels évoqués cognitifs ou Event-Related Potentials (ERP). Pour tenir compte de la grande dimension des données et de l'objectif de maitriser la complexité calculatoire des algorithmes d'estimation, les méthodes proposées s'appuient le plus souvent sur l'hypothèse d'un modèle à facteurs de la covariance, ce qui offre une bonne flexibilité pour s'adapter à la diversité des structures de dépendance rencontrées et permet une réduction de la dimensionnalité des problèmes. A travers les différentes situations abordées, ce travail tend à montrer qu'ignorer la dépendance peut conduire à des analyses erronées et qu'au contraire, sa prise en compte est un atout pour mieux détecter ou identifier un signal.

Le Chapitre 1 présente le contexte de la thèse et une revue de méthodes intégrant la question de la dépendance dans la sélection de variables. La grande diversité des structures de dépendance de données à haut débit est illustrée par des exemples issus d'applications en neuroscience et en génomique. Essentiellement, les approches intégrant la dépendance entre les statistiques de sélection visent à une décorrélation de ces statistiques afin de rétablir les propriétés de méthodes conçues, et le plus souvent évaluées de manière analytique, sous une hypothèse d'indépendance. Le Chapitre 1 décrit différentes méthodes, fondées sur deux techniques de décorrélation : par la transformation linéaire des données pour estimer des innovations ou par l'ajustement de l'effet de variables latentes expliquant la dépendance. Enfin, un lien est établi entre les bonnes performances de ces méthodes, même en situation de dépendance forte, et de nombreux résultats plus théoriques faisant état, de manière paradoxale, de la réduction des bornes de détectabilité et d'estimabilité en situation de dépendance.

Le Chapitre 2 présente une méthode de tests multiples adaptée à la problématique d'identification d'un signal dans des données d'ERP. Dans cette situation, les statistiques de tests d'association avec une condition expérimentale présentent une forme de dépendance temporelle structurée à la fois par des blocs d'intervalles de temps très corrélés et une forte composante auto-régressive (auto-corrélation estimée à 0.99). Par ailleurs, le signal recherché est parfois très faible au regard de la grande variabilité des courbes d'ERP entre les sujets. Pour s'affranchir de la forte régularité de la courbe des statistiques de tests induite par leur forte dépendance, on propose de s'appuyer sur un modèle à facteurs latents de la structure de covariance pour intégrer à la méthode d'estimation du signal une étape de reconstruction par régression du bruit résiduel à partir de son observation supposée sur des intervalles de temps de signal nul. Cette méthode permet de tirer profit de la connaissance a priori de ces intervalles, liée à une expertise des données d'ERP, ou de définir une méthode adaptative partant d'une estimation libérale de ces intervalles de temps et affinant ces estimations par une meilleure séparation du bruit et du signal. Une étude par simulation montre que cette méthode donne de bons résultats en matière d'identification du support du signal : le FDR est contrôlé au niveau fixé, contrairement à d'autres méthodes de prise en compte de la dépendance par un modèle à facteurs latents (SVA, LEAPP) et le taux de détection du signal est plus important. Enfin, une interprétation des résultats obtenus sur données réelles est proposée. La méthode est implémentée dans le package `R` intitulé `ERP` (Causeur and Sheu (2014)).

Le Chapitre 3 propose une adaptation des méthodes de décorrélation présentées dans le Chapitre 1 à l'objectif de la sélection de variables pour des modèles linéaires de classification supervisée. L'impact de la dépendance sur les procédures de sélection de variables ignorant la dépendance, telle la méthode LASSO, se traduit essentiellement par un défaut de précision de la sélection, à savoir une sélection en partie non-conforme au réel pouvoir prédictif des variables explicatives. Par ailleurs, l'étude des propriétés de ces méthodes de sélection dans une situation d'analyse de données d'expression de gènes met en évidence un fort défaut de reproductibilité de la sélection. Dans le cadre général de variables explicatives distribuées selon une loi normale multivariée homoscédastique, sous-jacent à l'Analyse Discriminante Linéaire, on propose d'introduire des facteurs latents conditionnellement auxquels les variables explicatives sont indépendantes. On démontre que la règle linéaire de classification optimale conditionnellement à ces facteurs est plus performante que la règle non-conditionnelle, et qu'elle est équivalente à la règle de Bayes appliquée aux variables explicatives ajustées de l'effet des facteurs.

Un algorithme de type Expectation-Maximisation (EM) pour l'estimation des paramètres du modèle est proposé, dans lequel l'appartenance aux classes définies par les valeurs de la variable à expliquer sont également considérées comme manquantes. Dans ce contexte, l'étape E consiste en une estimation jointe des facteurs latents et des probabilités individuelles d'appartenance aux classes. La méthode d'estimation ainsi définie est compatible avec une problématique de prédiction dans laquelle les classes définies par les valeurs de la variable réponse sont inconnues. L'analyse comparative des propriétés de la méthode FADA, pour Factor-Adjusted Discriminant Analysis, montre une amélioration générale des performances de prédiction,

lorsque le modèle est construit à partir d'une procédure de sélection de variables intégrant un modèle à facteurs de la dépendance. De plus, l'analyse de données sur le cancer du sein suggère que la méthode proposée permet d'obtenir un ensemble plus reproductible de variables sélectionnées. La méthode est implémentée dans le package `R` intitulé `FADA (Factor-Adjusted Discriminant Analysis` (Perthame et al. (2014)).

Les problématiques de détection et d'identification de signal en situation de données dépendantes sont abordées de manière plus formelle dans le Chapitre 4. En effet, le cadre général de ce Chapitre est celui proposé par Donoho and Jin (2004) et repris par Hall and Jin (2010) d'un modèle de mélange Gaussien pour le vecteur des statistiques de tests, dont les composantes ont des moyennes différentes si l'hypothèse nulle est vraie ou non. Donoho and Jin (2004) définit le paradigme d'un signal "Rare-and-Weak" comme la combinaison de valeurs faibles pour la proportion d'hypothèses non-nulles (parcimonie) et pour l'amplitude du signal. Sous l'hypothèse d'indépendance entre les statistiques de tests, il donne les bornes théoriques de parcimonie et d'amplitude pour la détectabilité et l'estimabilité du signal et propose une méthode, nommée Higher Criticism Thresholding (HCT), dont il démontre qu'elle atteint ces bornes théoriques. Le Chapitre 4 présente une étude des propriétés de HCT en situation de dépendance. Les bornes de détectabilité et d'estimabilité sont étendues à des situations arbitrairement complexes de dépendance. La méthode proposée est appliquée à des données d'ERP auditives et sur des simulations.

Ce travail de thèse conduit finalement à deux éléments de conclusion principaux. Le premier de ces éléments peut prendre la forme d'une incitation à la prudence lors de l'analyse de données fortement dépendantes par des méthodes qui ignorent cette dépendance. Par exemple, l'étude des propriétés des méthodes de sélection d'intervalles de temps dans des données d'ERP donne quelques illustrations spectaculaires de la déformation d'un signal par un bruit fortement auto-corrélé. Le deuxième élément est à la fois plus constructif et plus sujet à discussion. En effet, en réponse à la non-consistance des méthodes ignorant la dépendance, le propos de cette thèse est de proposer des méthodes de décorrélation des statistiques de sélection, basées sur un modèle d'analyse en facteur de la dépendance. Bien que présentant une grande flexibilité, ce modèle est particulièrement adapté à des profils de loi normale multivariée. Il est par exemple peu approprié pour des données qualitatives telles que celles de génotypage rencontrées en statistique génomique dans les études d'association à l'échelle du génome. Dans ce cas par exemple, il peut être plus intéressant de modéliser la dépendance par l'existence d'états cachés.

Par ailleurs, enrichir une méthode statistique par un modèle de la dépendance introduit un risque de sur-ajustement, pouvant se traduire par d'apparentes bonnes performances sur les données d'apprentissage, contredites par l'application à de nouvelles données. Ce travail a veillé à limiter ce risque par l'usage de méthodes de validation respectant scrupuleusement la séparation entre données d'apprentissage et données de validation. Toutefois, il n'explore pas la possibilité de limitation de la complexité des modèles par des méthodes conduisant à des estimations parcimonieuses des structures de dépendance. Ainsi, l'inverse de la matrice de covariance d'un modèle à facteurs intervient indirectement dans chacun des chapitres de cette

thèse. Or une forte dépendance se traduit aussi par un grand nombre de valeurs très faibles dans la matrice de covariance inverse.

La méthode de décorrélation par les innovations introduite dans le Chapitre 1 et reprise dans le Chapitre 4 fait appel à la racine de cette matrice. Il serait intéressant d'étudier l'apport d'une modélisation parcimonieuse de cette matrice d'innovation. En effet, notamment en situation de forte dépendance temporelle, on peut supposer que l'auto-corrélation partielle entre deux temps conditionnellement à tous les autres s'annule pour des délais courts entre ces temps, ce qui se traduit par une matrice de covariance inverse de forme bande-diagonale. C'est l'hypothèse faite par Hall and Jin (2010) pour définir la méthode *innovated* HCT. Sous une approche par un modèle à k facteurs pour une matrice de variance Σ, il est possible de donner l'expression d'une décomposition de Σ^{-1} sous la forme de la somme d'une matrice diagonale positive et du produit d'une matrice de rang k par sa transposée. On peut alors définir une méthode visant directement à estimer Σ^{-1} de manière parcimonieuse par une pénalisation ℓ_1 ou ℓ_2 de la vraisemblance. L'intérêt d'une telle approche est conforté par une récente publication (Van Wieringen and Peeters (2014)), dans laquelle les auteurs proposent une estimation Ridge de l'inverse de la matrice de covariance et démontre la supériorité de leur méthode par rapport à des approches de type Lasso (Friedman et al. (2008)).

Par ailleurs, la question de l'introduction de modèles non paramétriques pour décrire l'évolution des paramètres des modèles d'ERP au cours du temps de manière régulière s'est posée tout au long de ce travail de thèse. Par exemple, Hazarika et al. (1997) et Subasi et al. (2005) proposent des méthodes fondées sur une décomposition des courbes sur une base d'ondelettes et Bugli and Lambert (2006) proposent un modèle d'analyse de variance fonctionnelle estimé par la méthode des splines pénalisés, initialement proposés par Eilers and Marx (1996). Toutefois, comme le rapportent de nombreux auteurs (voir Wood (2012); Wiesenfarth et al. (2012)), l'estimation des variances et des covariances résiduelles dans les modèles additifs généralisés est problématique, conduisant souvent à une sous-estimation systématique. Les tests d'analyse de variance qui en découlent ne contrôlent pas le risque de 1ère espèce au niveau souhaité. Par ailleurs, cette mauvaise estimation perturbe les méthodes de décorrélation, ce qui génère une mauvaise séparation du signal et du bruit.

En statistique génomique, les données protéomiques prennent également la forme de spectres. On remarque à partir de données publiques associées à une telle expérience (voir Petricoin et al. (2002)) que ces spectres présentent une structure de dépendance similaire à celle des ERPs, avec une diagonale de forte auto-corrélations et des blocs de corrélations. Ces données sont également analysées par des modèles mixtes d'analyse de variance fonctionnelle permettant de prendre en compte la grande variabilité inter-individus des spectres de peptides et font appel à des décompositions sur base d'ondelettes (voir notamment Giacofci et al. (2013)). Dans l'une ou l'autre de ces approches (splines ou ondelettes), il pourrait être pertinent d'étudier l'apport des méthodes de décorrélation.

CHAPITRE 6

LISTE DES TRAVAUX

Publications

Perthame E., Friguet C. and Causeur D., *Stability of feature selection in classification issues for high-dimensional correlated data*, Statistics and Computing (2015).

Sheu C.-F., Perthame E., Lee Y.-S. and Causeur D., *Accounting for time dependence in large-scale multiple testing of event-related potential data*, en seconde révision (2015).

Communications orales

Le nom de l'orateur est en premier

Conférences

Perthame E., Causeur D. Variable selection by decorrelated HCT for supervised classification in high dimension VU University, Amsterdam, Netherlands, July 6-10 2015. European Meeting of Statisticians (EMS)

Perthame E., Causeur D. Variable selection by decorrelated HCT for supervised classification in high dimension Université de Lille, Lille, France, June 1-5 2015. 47e Journées de Statistique de la SFDS (JdS)

Perthame E., Friguet C., Causeur D. FADA : an R package for variable selection in supervised classification of strongly dependent data UCLA, Los Angeles, California, USA, June 30 - July 3 2014. useR !2014

Causeur D., Perthame E., Sheu, C.-F. ERP : an R package for Event-Related Potentials data analysis UCLA, Los Angeles, California, USA, June 30 - July 3 2014. useR !2014

Perthame E., Sheu C.-F., Lee Y.-S., Causeur D. Dealing with long-time range dependence in large-scale multiple testing of Event-Related Potentials data ENSAI, Rennes, France, June 2-6 2014. 46e Journées de Statistique de la SFDS (JdS)

Perthame E., Friguet C., Causeur D. Stabilité de la sélection de variables pour la classification de données en grande dimension ESC Toulouse, Toulouse, France, May 27-31 2013. 45e Journées de Statistique de la SFDS (JdS)

Perthame E., Friguet C., Causeur D. Stability of variable selection for high-dimensional data VU University, Amsterdam, Netherlands, January 24-25 2013. Statistical Methods for (post)-Genomics Data (SMPGD)

Séminaires

Perthame E., Friguet C., Causeur D. Sélection de variables pour la classification supervisée en grande dimension. ENSAI, Rennes, France, June 13 2014. Séminaire des doctorants de statistique de l'IRMAR

Perthame E., Friguet C., Sheu C.-F., Causeur D. Prise en compte de la dépendance en grande dimension. Application aux problèmes de sélection de variables et aux tests multiples. Université Pierre et Marie Curie, Paris, France, April 11 2014. Séminaire des doctorants du LSTA

BIBLIOGRAPHIE

M. Ahdesmäki and K. Strimmer. Feature selection in omics prediction problems using cat scores and false non-discovery rate control. *Annals of Applied Statistics*, 4 :503–519, 2010. (pages 14, 15, 17, 21, 22, 27, 65, 86, 95, 96, 113 et 120)

M. Ahdesmäki, V. Zuber, S. Gibb, and K. Strimmer. *sda : Shrinkage Discriminant Analysis and CAT Score Variable Selection*, 2014. URL `http://CRAN.R-project.org/package=sda`. R package version 1.3.5. (page 22)

H. Akaike. Information theory and an extension of the maximum likelihood principle. Originally published in *Proceedings of the Second International Symposium on Information Theory,* B.N. Petrov and F. Caski, eds, Akademia Kiado, Budapest, pages 267–281, 1973. (page 12)

G. I. Allen, L. Grosenick, and J. Taylor. A generalized least-square matrix decomposition. *Journal of the American Statistical Association*, 109(505) :145–159, 2014. (pages 14 et 17)

U. Alon et al. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Cell Biology*, 96 :6745–6750, 1999. (page 17)

T. W. Anderson. Asymptotic theory for principal component analysis. *Annals of Mathematical Statistics*, 34 :122–148, 1963. (page 26)

F. Bach. Bolasso : model consistent lasso estimation through the bootstrap. *Proceedings of the Twenty-fifth International Conference on Machine Learning (ICML)*, 2008. (pages 16, 74 et 82)

U. Baron et al. Dna methylation analysis as a tool for cell typing. *Epigenetics*, 1 : 55–60, 2006. (page 13)

M. S. Bartlett. Tests of significance in factor analysis. *British Journal of Psychology,*, 3 :77–85, 1950. (page 26)

M. S. Bartlett. A further note on tests of significance. *British Journal of Psychology*, 4 :1–2, 1951. (page 26)

Y. Benjamini and Y. Hochberg. Controlling the False Discovery Rate : a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57 :289–300, 1995. (pages 14, 35, 41, 54 et 62)

Y. Benjamini and D. Yekutieli. The control of the False Discovery Rate in multiple testing under dependency. *Annals of Statistics*, 29(4) :1165–1188, 2001. (pages 14, 41 et 54)

P. Bickel and E. Levina. Some theory for Fisher's Linear Discriminant function, naive Bayes, and some alternatives when there are many more variables than observations. *Bernoulli*, 10(6) :989–1010, 2004. (pages 15, 65, 71 et 120)

R.-C. Blair and W. Karniski. An alternative method for significance testing of waveform difference potentials. *Psychophysiology*, 30 :518–524, 1993. (pages 35 et 62)

Y. Blum, G. LeMignon, S. Lagarrigue, and D. Causeur. A factor model to analyze heterogeneity in gene expression. *BMC bioinformatics*, 11 :368, 2010. (page 24)

C. Bonferroni. Teoria statistica delle classi e calcolo delle probabilità. *Pubblicazioni del R Istituto Superiore si Scienze Economiche e Comerciali di Firenze*, pages 3–62, 1936. (pages 14 et 40)

L. Breiman. Heuristics of instability and stabilization in model selection. *The Annals of Statistics*, 24(6) :2350–2383, 1996. (page 13)

L. Breiman. Random forests. *Machine learning*, 45 :5–32, 2001. (page 74)

K. Broman and T. Speed. A model selection approach for the identification of quantitative trait loci in experimental crosses. *Journal of the Royal Statistical Society, Series B*, 64 :641–656, 2002. (page 13)

C. Bugli and P. Lambert. Functional anova with random functional effects : an application to event-related potentials modelling for electroencephalograms analysis. *Statistics in Medicine*, 25 :3718–3739, 2006. (pages 17, 101, 104 et 132)

A. Buja and N. Eyuboglu. Remarks on parallel analysis. *Multivariate Behavioral Research*, 27(4) :509–540, 1992. (page 26)

T. Cai, J. Jeng, and J. Jin. Optimal detection of heterogeneous and heteroscedastic mixtures. *Journal of the Royal Statistical Society, Series B*, 73, 2011. (page 110)

C. Carvalho, C. J., J. Lucas, J. Nevins, Q. Wang, and M. West. High-dimensional sparse factor modeling : applications in gene expression genomics. *Journal of the American Statistical Association : Applications and Case Studies*, 103 : 484, 2008. (pages 14, 17, 19 et 24)

G. Casella, F. J. Giron, M. L. Martinez, and E. Moreno. Consistency of bayesian procedure for variable selection. *The Annals of Statistics*, 37 :1207–1228, 2009. (page 13)

D. Causeur and C.-F. Sheu. *ERP : Significance analysis of Event-Related Potentials data*, 2014. URL `http://CRAN.R-project.org/package=ERP`. R package version 1.0.1. (pages 54, 62 et 130)

D. Causeur, C. Friguet, M. Houée, and M. Kloareg. Factor analysis for multiple testing (FAMT) : an R package for large-scale significance testing under dependence. *Journal of Statistical Software*, 40(14) :1–19, 2011. (pages 25 et 35)

D. Causeur, M.-C. Chu, S. Hsieh, and C.-F. Sheu. A factor-adjusted multiple testing procedure for erp data analysis. *Behavior Research Methods*, 44 :635–643, 2012. (pages 25, 35, 47, 54, 95, 96, 105 et 114)

D. Chung and S. Keles. Sparse partial least squares classification for high dimensional data. *Statistical Applications in Genetics and Molecular Biology*, 9(17), 2010. (page 17)

L. Clemmensen and M. Kuhn. *sparseLDA : Sparse Discriminant Analysis*, 2012. URL `http://CRAN.R-project.org/package=sparseLDA`. R package version 0.1-6. (page 72)

L. Clemmensen, T. Hastie, D. Witten, and B. Ersbøll. Sparse discriminant analysis. *Technometrics*, 53(4) :406–413, 2011. (pages 16, 70, 71, 86 et 88)

T. Cover and J. Thomas. *Elementary Information Theory*. Wiley, Hoboken, NJ, 2006. (page 28)

A. Dabney and J. Storey. Optimality driven nearest centroid classification from genomic data. *PLoS One*, 2(e1002), 2007. (page 65)

D. Donoho and J. Jin. Higher criticism for detecting sparse heterogeneous mixtures. *The Annals of Statistics*, 32 :3 :962–994, 2004. (pages 14, 19, 30, 31, 95, 103, 105, 108, 110, 113, 127 et 131)

D. Donoho and J. Jin. Higher criticism thresholding : Optimal feature selection when useful features are rare and weak. *Proceedings of the National Academy of Sciences*, 105 :39 :14790–14795, 2008. (pages 14, 28, 59, 85, 86, 95, 108, 113 et 120)

D. Donoho and J. Jin. Feature selection by higher criticism thresholding achieves the optimal phase diagram. *Philosophical Transactions of the Royal Society A*, 367 :4449–4470, 2009. (page 120)

D. Donoho and J. Jin. Higher criticism for large-scale inference, especially for rare and weak effects. *Statistical Science*, 30 :1 :1–25, 2015. (page 22)

S. Dudoit, J. Fridlyand, and T. Speed. Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American Statistical Association*, 97 :77–87, 2002. (page 65)

B. Efron. The efficiency of logistic regression compared to normal discriminant analysis. *Journal of the American Statistical Association*, 70 :892–896, 1975. (page 28)

B. Efron. Correlation and large-scale simultaneous testing. *Journal of the American Statistical Association*, 102 :93–103, 2007. (pages 35, 40 et 62)

B. Efron. *Empirical Bayes estimates for large-scale prediction problems*. Technical report, Dept. Statistics, Stanford Univ, 2008. (pages 15, 27, 65 et 71)

B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression (with discussion). *Annals of Statistics*, 32 :407–451, 2004. (page 73)

P. Eilers and B. Marx. Flexible smoothing with B-splines and penalties. *Statistical Science*, 11(2) :89–121, 1996. (page 132)

J. Fan and R. Li. Variable selection via non-concave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456) :1348–1360, 2001. (page 13)

J. Fan and J. Lv. A selective overview of variable selection in high dimensional feature space. *Statistica Sinica*, 20 :101–148, 2010. (page 16)

R. A. Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7(2) :179–188, 1936. (page 68)

J. Ford, R. MacCallum, and M. Tait. The application of exploratory factor analysis in applied psychology : a critical review and analysis. *Personnel Psychology*, 39 : 291–314, 1986. (page 26)

Y. Freund and R. Schapire. Experiments with a new boosting algorithm. *Proceedings of the Thirteenth International Conference on Machine Learning*, 1996. (page 74)

Y. Freund and R. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55 : 119–139, 1997. (page 74)

J. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9 :432–441, 2008. (page 132)

J. Friedman, T. Hastie, and R. Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33 :1–22, 2010. (pages 66, 74 et 86)

C. Friguet and D. Causeur. Estimation of the proportion of true null hypotheses in high-dimensional data under dependence. *Computational Statistics and Data Analysis*, 55 :2665–2676, 2011. (page 38)

C. Friguet, M. Kloareg, and D. Causeur. A factor model approach to multiple testing under dependence. *Journal of the American Statistical Association*, 104 :488 : 1406–1415, 2009. (pages 14, 17, 19, 21, 24, 25, 26, 27, 35, 47, 48, 51, 54, 65, 80, 84, 95, 96, 114 et 116)

J.-M. Gardiner, B. Gawlik, and A. Richardson-Klavehn. Maintenance rehearsal affects knowing, not remembering : Elaborative rehearsal affects remembering, not knowing. *Psychonomic Bulletin & Review*, 1 :107–110, 1994. (page 59)

M. Giacofci, S. Lambert-Lacroix, G. Marot, and F. Picard. Wavelet-based clustering for mixed-effects functional models in high dimension. *Biometrics*, 69(1) :31–40, 2013. (page 132)

T. Golub et al. Molecular classification of cancer : class discovery and class prediction by gene expression monitoring. *Science*, 286 :(5439 :5531, 1999. (page 17)

D.-M. Groppe, T.-P. Urbach, and M. Kutas. Mass univariate analysis of event-related brain potentials/fields : I. A critical tutorial review. *Psychophysiology*, 48 :1711–1725, 2011a. (pages 16, 35, 40 et 59)

D.-M. Groppe, T.-P. Urbach, and M. Kutas. Mass univariate analysis of event-related brain potentials/fields : II. Simulation studies. *Psychophysiology*, 48 : 1726–1737, 2011b. (pages 16, 35 et 40)

Y. Guo, T. Hastie, and R. Tibshirani. Regularized discriminant analysis and its application in microarrays. *Biostatistics*, 8 :86–100, 2007. (pages 15, 65 et 71)

Y. Guo, T. Hastie, and R. Tibshirani. *rda : Shrunken Centroids Regularized Discriminant Analysis*, 2012. URL `http://CRAN.R-project.org/package=rda`. R package version 1.0.2-2. (page 71)

D. Guthrie and J.-S. Buchwald. Significance testing of difference potentials. *Psychophysiology*, 28 :240–244, 1991. (pages 17, 35, 41, 54 et 105)

P. Hall and J. Jin. Properties of higher criticism under strong dependence. *The Annals of Statistics*, 36 :1 :381–402, 2008. (pages 14, 27 et 95)

P. Hall and J. Jin. Innovated higher criticism for detecting sparse signals in correlated noise. *The Annals of Statistics*, 38 :3 :1686–1732, 2010. (pages 14, 21, 23, 24, 27, 28, 29, 95, 105, 108, 113, 114, 116, 117, 119, 131 et 132)

D. Hand. Classifier technology and the illusion of progress. *Statistical Science*, 21 : 1–14, 2006. (page 28)

T. Handy. *Event-Related Potentials*. The MIT Press : Cambridge, 2004. (pages 13, 16, 34 et 39)

W. Hardle and L. Simar. *Applied Multivariate Statistical Analysis*. Springer Berlin Heidelberg, 2007. (page 70)

T. Hastie, R. Tibshirani, and A. Buja. Flexible discriminant analysis by optimal scoring. *Journal of the American Statistical Association*, 89(428) :1255–1270, 1994. (page 70)

T. Hastie, A. Buja, and R. Tibshirani. Penalized discriminant analysis. *The Annals of Statistics*, 23(1) :73–102, 1995. (page 66)

T. Hastie, R. Tibshirani, and J. Friedman. *The elements of statistical learning*. Springer, 2009. (page 71)

J. Hausseur and K. Strimmer. Entropy inference and the james-stein estimator, with applications to nonlinear gene associations networks. *Journal of Machine Learning Research*, 10 :1469–1484, 2009. (page 22)

N. Hazarika, J. Chen, C. Tsoi, and A. Sergejew. Classification of eeg signals using the wavelet transform. *Signal Processing*, 59 :61–72, 1997. (page 132)

I. Hedenfalk, D. Duggan, Y. D. Chen, M. Radmacher, M. Bittner, R. Simon, P. Meltzer, B. Gusterson, M. Esteller, O. P. Kallioniemi, B. Wilfond, A. Borg, and J. Trent. Gene expression profiles in hereditary breast cancer. *New England Journal of Medicine*, 344 :539–548, 2001. (page 81)

J. Hilbe. *Logistic regression models*. Chapman et Hall, 2009. (page 72)

A. Hoerl and R. Kennard. Ridge regression : Biased estimation for nonorthogonal problems. *Technometrics*, 12 :55–67, 1970. (pages 12, 15 et 73)

R. Hornung, C. Bernau, C. Truntzer, T. Stadler, and A.-L. Boulesteix. Full versus incomplete cross-validation : measuring the impact of imperfect separation between training and test sets in prediction error estimation. *Technical Report 159, Department of Statistics, LMU. In revision.*, 2014. (page 87)

E.-A. Houseman, K.-T. Kelsey, W. J.-K., and C.-J. Marsit. Cell-composition effects in the analysis of dna methylation array data : a mathematical perspective. *BMC Bioinformatics*, 16 - 95, 2015. (pages 14, 17, 19, 24 et 84)

D. Hunter and K. Lange. A tutorial on mm algorithms. *The American Statistician*, 58 :30–37, 2004. (page 72)

Y. Ingster. Some problems of hypothesis testing leading to infinitely divisible distribution. *Mathematical Methods of Statistics*, 6 :47–69, 1997. (pages 14 et 127)

Y. Ingster. Minimax detection of a signal for ℓ_n^p balls. *Mathematical Methods of Statistics*, 7 :401–428, 1999. (page 108)

J. Jin. Impossibility of successful classification when useful features are rare and weak. *Proceedings of the National Academy of Sciences of United States of America*, 106(22) :8859–8864, 2009. (page 59)

K.-G. Jöreskog. Some contributions to maximum likelihood factor analysis. *Psychometrika*, 32 :443–482, 1967. (page 26)

H. F. Kaiser. The application of electronic computer to factor analysis. *Educational and Psychological Measurement*, 20 :141–151, 1960. (page 26)

J. Khan et al. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nature medicine*, 7 :673–679, 2001. (page 17)

Y. Kim, S. Kwon, and H. Choi. Consistent model selection criteria on high dimension. *Journal of Machine Learning Research*, 13 :1037–1057, 2012. (page 13)

B. Klaus and K. Strimmer. Signal identification for rare and weak features : higher criticism or false discovery rates ? *Biostatistics*, 14 :1 :129–143, 2013. (pages 14, 108 et 113)

R. Kustra, R. Shioda, and M. Zhu. A factor analysis model for functional genomics. *BMC Bioinformatics*, 7, 2006. (pages 14, 17, 19 et 24)

A. Lage-Castellanos, E. Matínez-Montes, J.-A. Hernández-Cabrera, and L. Galán. False Discovery Rate and permutation test : An evaluation in ERP data analysis. *Statistics in Medicine*, 29 :63–74, 2010. (pages 40 et 62)

K. Lange. *Optimization*. Springer, New York, 2004. (page 72)

S. Langevin, D. Koestler, B. Christensen, R. Butler, J. Wiencke, H. Nelson, et al. Peripheral blood dna methylation profiles are indicative of head and neck squamous cell carcinoma. *Epigenetics*, 7(3) :291–299, 2012. (page 84)

D. N. Lawley. Tests of significance for the latent roots of covariance and correlation matrix. *Biometrika*, 43(1/2) :128–136, 1956. (page 26)

S. Lee and S. Batzoglou. Application of independent component analysis to microarrays. *Genome Biology*, 4 :11 :R76, 2003. (page 17)

Y.-S. Lee, H.-M. Lee, and J.-M. Fawcett. Intentional forgetting reduces color-naming interference : evidence from item-method directed forgetting. *Journal of Experimental Psychology. Learning, Memory and Cognition*, 39(1) :220–236, 2013. (page 36)

J. T. Leek and J. Storey. Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genetics*, 3(9) :e161, 2007. (pages 14, 17, 19, 21, 25, 26, 35, 52 et 65)

J. T. Leek and J. Storey. A general framework for multiple testing dependence. *Proceedings of the National Academy of Sciences*, 105 :18718–18723, 2008. (pages 21, 24, 25, 35, 54 et 65)

J.-T. Leek, W.-E. Johnson, H.-S. Parker, A.-E. Jaffe, and J.-D. Storey. *SVA : Surrogate Variable Analysis*, 2014. R package version 3.12.0. (pages 25, 27 et 54)

E. Levina. *Statistical issues in texture analysis. PhD thesis*. University of California, Berkeley, 2002. (page 65)

K. V. Mardia, J. T. Kent, and J. M. Bibby. *Multivariate Analysis*. 1979. (page 26)

N. Meinshausen and P. Bühlmann. High-dimensional graphs and variable selection with the LASSO. *The Annals of Statistics*, 34 :3 :1436–1462, 2006. (page 85)

N. Meinshausen and P. Bühlmann. Stability selection. *Journal of the Royal Statistical Society, series B*, 72 :4 :417–473, 2010. (pages 16 et 74)

R. Opgen-Rhein and K. Strimmer. Accurate ranking of differentially expressed genes by a distribution-free shrinkage approach. *Statistical Applications in Genetics and Molecular Biology*, 6(9), 2007. (page 17)

E. Perthame, C. Friguet, and D. Causeur. *FADA : Variable selection for supervised classification in high dimension*, 2014. URL `http://CRAN.R-project.org/package=FADA`. R package version 1.2. (pages 31, 88 et 131)

E. Perthame, C. Friguet, and D. Causeur. Stability of feature selection in classification issues for high-dimensional correlated data. *Statistics and Computing*, pages 1–14, 2015. (pages 31, 65 et 114)

E. I. Petricoin et al. Use of proteomic patterns in serum to identify ovarian cancer. *The Lancet*, 359(2) :572–577, 2002. (page 132)

T. W. Picton. The p300 wave of the human event-related potential. *Journal of Clinical Neurophysiology*, 9(4) :456–479, 1992. (page 96)

R.-A. Poldrack, J.-A. Mumford, and T.-E. Nichols. *Handbook of functional fMRI data analysis*. Cambridge University Press : New York, 2011. (pages 13 et 34)

I. Pournara and L. Wernisch. Factor analysis for gene regulatory networks and transcription factor activity profiles. *BMC Bioinformatics*, 8 :61, 2007. (pages 19 et 24)

W.-H. Press, S.-A. Teukolsky, W.-T. Vetterling, and B.-P. Flannery. *Numerical Recipes : The Art of Scientific Computing (3rd edn)*. Cambridge University Press : New York, 2007. (pages 25 et 51)

D.-B. Rubin and D.-T. Thayer. EM algorithms for ML factor analysis. *Psychometrika*, 47 :69–76, 1982. (pages 26 et 51)

M.-D. Rugg and T. Curran. Event-related potentials and recognition memory. *Trends in Cognitive Sciences*, 11(6) :251–257, 2007. (pages 36 et 59)

J. Schäfer and K. Strimmer. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical Applications in Genetics and Molecular Biology*, 4(32), 2005. (pages 17, 22, 113 et 115)

G. Schwarz. Estimating the dimension of a model. *Annals of statistics*, 6 :461–464, 1978. (page 12)

D. Shalon, S. J. Smith, and P. O. Brown. A dna microarray system for analyzing complex dna samples using two-color fluorescent probe hybridization. *Genome Research*, 6(7) :638–645, 1996. (pages 13 et 17)

J. Shao. An asymptotic theory for linear model selection. *Statistica Sinica*, 7 : 221–264, 1997. (page 13)

C. Sheu, E. Perthame, D. Causeur, and Y. Lee. Accounting for time dependence in large-scale multiple testing of event-related potential data. *In revision*, 2015. (pages 30, 34, 88 et 115)

D. Singh et al. Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell*, 1 :203–209, 2002. (page 17)

N.-J. Smith and M. Kutas. Regression-based estimation of erp waveforms : I. the rerp framework. *Psychophysiology*, 52(2) :157–168, 2015a. (page 101)

N.-J. Smith and M. Kutas. Regression-based estimation of erp waveforms : Ii. nonlinear effects, overlap correction, and practical considerations. *Psychophysiology*, 52(2) :169–181, 2015b. (page 101)

J.-D. Storey and R. Tibshirani. Statistical significance for genome-wide experiments. *Proceedings of the National Academy of Sciences of the United States of America*, 100 :9440–9445, 2003. (page 47)

J. D. Storey, J. Dai, and J. T. Leek. The optimal discovery procedure for large-scale significance testing, with application to comparative microarray experiments. *Biostatistics*, 8 :414–432, 2007. (page 114)

A. Subasi, M. Akin, K. Kiymik, and O. Erogul. Automatic recognition of vigilance state by using a wavelet-based artificial neural network. *Neural Computing and Applications*, 14 :45–55, 2005. (page 132)

W. Sun and T.-T. Cai. Large-scale multiple testing under dependence. *Journal of the Royal Statistical Society, Series B*, 71(2) :1–32, 2009. (pages 24 et 35)

Y. Sun, N. Zhang, and A. Owen. Multiple hypothesis testing adjusted for latent variables, with an application to the AGEMAP gene expression data. *The Annals of Applied Statistics*, 6(4) :1664–1688, 2012. (pages 14, 17, 19, 21, 25, 26, 35, 52, 65 et 114)

Y. Sun, N. Zhang, and A. Owen. *leapp : latent effect adjustment after primary projection*, 2014. URL `http://CRAN.R-project.org/package=leapp`. R package version 1.1. (pages 25 et 54)

A. Teschendorff, J. Zhuang, and M. Widschwendter. Independent surrogate variable analysis to deconvolve confounding factors in large-scale microarray profiling studies. *Bioinformatics*, 27 :11 :1496–1505, 2011. (pages 17 et 24)

G. Thomson. *The Factorial Analysis of Human Ability.* 1951. (page 26)

R. Tibshirani. Regression shrinkage and selection via LASSO. *Journal of the Royal Statistical Society, series B*, 58 :267–288, 1996. (pages 12, 15, 19, 73 et 75)

R. Tibshirani, T. Hastie, B. Narasimhan, and G. Chu. Diagnosis of multiple cancer type by shrunken centroids of gene expression. *Proc. Natl. Acad. Sci. USA*, 99 : 6567–6572, 2002. (pages 16, 27, 66 et 71)

R. Tibshirani, T. Hastie, B. Narsimhan, and G. Chu. Class prediction by nearest shrunken centroids, with applications to DNA microarrays. *Statistical Science*, 18 :104–117, 2003. (pages 15 et 66)

J. Tukey. T13 N : the higher criticism. *Course Notes*, Princeton University, 1976. (pages 14, 95 et 105)

S. Van de Geer. L1-regularization in high-dimensional statistical models. *Proceedings of the International Congress of Mathematicians*, 2010. (page 16)

M.-J. van der Laan and S. Dudoit. *Multiple Testing Procedures with Applications to Genomics*. Springer : New York, 2007. (pages 14 et 40)

W. Van Wieringen and C. Peeters. Ridge estimation of inverse covariance matrices from high-dimensional data. `arXiv:1403.0904 [stat.ME]`, 2014. (page 132)

N. Verzelen. Minimax risks for sparse regressions : Ultra-high dimensional phenomenons. *Electronic Journal of Statistics*, 6 :38–90, 2012. (page 30)

E. Vul, C. Harris, P. Winkielman, and H. Pashler. Puzzlingly high correlations in fMRI studies of emotion, personality, and social cognition. *Perspectives on Psychological Science*, 4(3) :274–290, 2009. (page 62)

M. West et al. Predicting the clinical status of human breast cancer by using gene expression profiles. *Proceedings of the National Academy of Sciences*, 98 : 11462–11467, 2001. (page 17)

P.-H. Westfall and S.-S. Young. *Resampling-Based Multiple Testing : Examples and Methods for p-value Adjustment*. Wiley : New York, 1993. (page 62)

M. Wiesenfarth, T. Krivobokova, S. Klasen, and S. Sperlich. Direct simultaneous inference in additive models and its application to model undernutrition. *Journal of the American Statistical Association*, 107(500) :1286–1296, 2012. (page 132)

L. Williams, E. Simms, C. Clark, , and R. Paul. The test-retest reliability of a standardized neurocognitive and neurophysiological test battery : neuromarker. *International Journal of Neuroscience*, 115 :1605–1630, 2005. (page 97)

D. Witten. *penalizedLDA : Penalized classification using Fisher's linear discriminant*, 2011. URL `http://CRAN.R-project.org/package=penalizedLDA`. R package version 1.0. (page 72)

D. Witten and R. Tibshirani. Penalized classification using fisher's linear discriminant. *Journal of the Royal Statistical Society, Series B*, 73(5) :753–772, 2011. (pages 16 et 72)

S.-N. Wood. On p-values for smooth components of an extended generalized additive model. *Biometrika*, 2012. doi : 10.1093/biomet/ass048. (page 132)

M.-W. Woolrich, C.-F. Beckmann, T.-E. Nichols, and S.-M. Smith. *Statistical analysis of fMRI data. In M. Filippi (Ed.), fMRI techniques and protocols*. Humana Press : New York, 2009. (page 59)

P. Xu, G. Brock, and R. S. Parrish. Modified linear discriminant analysis approaches for classification of high-dimensional microarray data. *Computational Statistics and Data Analysis*, 53 :1674–1687, 2009. (page 65)

Y. Yang. Can the strengths of AIC and BIC be shared ? A conflict between model identification and regression estimation. *Biometrika*, 92 :4 :937–950, 2005. (page 13)

N. Yeung, R. Bogacz, C.-B. Holroyd, and J.-D. Cohen. Detection of synchronized oscillations in the electroencephalogram : An evaluation of methods. *Psychophysiology*, 41 :822–832, 2004. (pages 17 et 47)

K. Zoski and S. Jurs. Using multiple regression to determine the number of factors to retain in factor analysis. *Multiple Linear Regression Viewpoints*, 20(1) :5–9, 1993. (page 26)

H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, series B*, 67(2) :301–320, 2005. (pages 16 et 73)

H. Zouridis et al. Methylation subtypes and large-scale epigenetic alterations in gastric cancer. *Science Translational Medicine*, 4(156) :156–140, 2012. (page 84)

V. Zuber and K. Strimmer. Gene ranking and biomarker discovery under correlation. *Bioinformatics*, 25 :2700–2707, 2009. (pages 14, 15, 17, 21, 22, 24, 65, 85, 96, 113, 114, 119 et 120)