# Prise en compte de la dépendance pour des problèmes de test global et de prédiction

Florian Hébert

# THESE DE DOCTORAT DE

AGROCAMPUS OUEST
DELIVREE CONJOINTEMENT AVEC
L'UNIVERSITE DE RENNES 1
COMUE UNIVERSITE BRETAGNE LOIRE

ECOLE DOCTORALE N° 601
*Mathématiques et Sciences et Technologies de l'Information et de la Communication*
Spécialité : *Mathématiques et leurs Interactions*

Par

**Florian HÉBERT**

# Prise en compte de la dépendance pour des problèmes de test global et de prédiction

**Thèse présentée et soutenue à Agrocampus Ouest, le 28 novembre 2019**
**Unité de recherche : IRMAR, UMR CNRS 6625**
**Thèse N° : 2019-19 / G-18**

**Rapporteurs avant soutenance** :

Nathalie VIALANEIX Directrice de recherche, INRA Toulouse
Ruth HELLER Associate Professor, Tel Aviv University

**Composition du Jury :**

| | | |
|---|---|---|
| Président : | Eric MATZNER-LØBER | Professeur des universités, ENSAE Formation continue |
| Examinateurs : | Pierre NEUVIAL | Chargé de recherche, Institut de Mathématiques de Toulouse |
| | Eric MATZNER-LØBER | Professeur des universités, ENSAE Formation continue |
| Dir. de thèse : | David CAUSEUR | Professeur des universités, Agrocampus Ouest |
| Co-dir. de thèse : | Mathieu EMILY | Maître de conférences HDR, Agrocampus Ouest |

**Invitée**
Madison GIACOFCI Maître de conférences, Université Rennes 2

# Contents

**List of Figures** **x**

**List of Tables** **xii**

**Remerciements** **1**

**Résumé** **3**

1 Contexte général . . . . . 4
2 Prise en compte de la dépendance pour les problèmes de prédiction . 5
2.1 Prédiction en petite dimension . . . . . 5
2.2 Prédiction en grande dimension ou en situation de forte dépendance : prendre en compte ou ignorer la dépendance . . . . 7
3 Prise en compte de la dépendance en test global . . . . . 13
3.1 Test global dans un modèle de régression standard . . . . . 13
3.2 Test global en grande dimension ou en présence de forte dépendance . . . . . 14
3.3 Test global et prise en compte de la dépendance en études d'association pangénomiques . . . . . 17
4 Quelques exemples de structures de dépendance spécifiques . . . . . 20
5 Organisation de la thèse . . . . . 22

**1 Introduction** **27**

1 General context . . . . . 28
2 Dependence handling for prediction problems . . . . . 29
2.1 Prediction in low dimension . . . . . 29
2.2 Prediction in high dimension or under strong dependence: handling or ignoring dependence . . . . . 31

3 Dependence handling for global testing . . . . . 36
3.1 Global testing in a standard regression model . . . . . 36
3.2 Global testing in high dimension or under strong dependence . 37
3.3 Global testing and dependence handling in Genome-Wide Association Studies . . . . . 40
4 A focus on some specific dependence patterns . . . . . 43
5 Organization of the thesis . . . . . 45

**2 Adaptive Handling of Dependence for Global Testing 49**
1 Introduction . . . . . 51
2 Signal detection by aggregation of pointwise test statistics . . . . . 53
2.1 A general framework for the detection of an association signal 54
2.2 Aggregation procedures of pointwise statistics . . . . . 56
2.3 Impact of the pattern of association signal on detection performance . . . . . 57
3 Weighted decorrelation of pointwise test statistics . . . . . 60
3.1 Oracle weighting procedure . . . . . 60
3.2 Implementation of the MGF-R weighting procedure . . . . . 64
4 Comparison of detection methods for association signals . . . . . 65
4.1 Simulation study . . . . . 65
4.2 Application to real datasets . . . . . 69
5 Discussion . . . . . 71
6 Appendix: proof of Theorem 3.1 . . . . . 74

**3 Detection of Interaction Effects Between a Gene and an Environmental Factor 77**
1 Introduction . . . . . 79
2 A general framework for the detection of interaction effects between a gene and an environmental factor . . . . . 81
2.1 Definition of a general framework adapted to the detection of interaction effects . . . . . 81
2.2 Estimation of the distribution of a global test statistic under the null hypothesis . . . . . 83
2.3 Existing methods specifically designed for the detection of gene - environmental factor interaction effects . . . . . 87

3 Assessment of the MGF-R test for the detection of gene - environmental factor interaction effects . . . . . 89
3.1 Control of the type I error rate . . . . . 89
3.2 Power study . . . . . 89
4 Discussion . . . . . 94

**4 Detection of Interaction Effects Between Two Genes 97**
1 Introduction . . . . . 99
2 Gene - gene interaction effects modeling . . . . . 101
2.1 A generalized linear model for the detection of gene - gene interaction effects . . . . . 101
2.2 Estimation of the distribution of a global test statistic under the null hypothesis . . . . . 103
2.3 Estimation of the marginal effects parameters . . . . . 104
3 Properties of the correlation matrix of the test statistics vector . . . . 105
3.1 Modelization of the dependence structure . . . . . 105
3.2 Eigendecomposition . . . . . 106
3.3 Sparsity and relative weakness of the correlation coefficients . 107
4 Comparison of global testing methods for gene - gene interaction effects detection . . . . . 111
4.1 Simulation procedure . . . . . 114
4.2 Results . . . . . 115
5 Extension of the proposed model for other interaction effects . . . . . 117
5.1 Dummy coding - based SNP - SNP interaction model . . . . . 119
5.2 Construction of a dummy coding - based gene - gene interaction model . . . . . 120
5.3 Correlation structure of the test statistics vector . . . . . 121
5.4 Comparison of the continuous and dummy codings . . . . . 123
6 Discussion . . . . . 127

**5 Adaptive Handling of Dependence for Regression Modeling 129**
1 Introduction . . . . . 131
2 Optimal versus naive linear prediction . . . . . 134
2.1 Optimal linear prediction . . . . . 135
2.2 Naive linear prediction . . . . . 136

2.3 Prediction performance of naive linear prediction . . . . . 138
2.4 Ordinary Least Squares and naive linear prediction in high dimension . . . . . 139
2.5 Illustrative comparative study in high dimension . . . . . 140
3 A new class of prediction scores . . . . . 144
3.1 Introduction of a new class of prediction scores . . . . . 144
3.2 $\mathcal{L}$ contains Ridge predictions scores . . . . . 145
3.3 $\mathcal{L}$ contains Partial Least Squares (PLS) prediction scores . . . 146
3.4 Alternative prediction scores . . . . . 147
4 Optimal prediction within $\mathcal{L}$ . . . . . 149
4.1 Closed-form expression of the best predictor . . . . . 150
4.2 Estimation of the optimal predictor within $\mathcal{L}$ . . . . . 151
5 Comparative study . . . . . 154
5.1 Simulation study . . . . . 154
5.2 Performance comparisons on datasets . . . . . 157
6 Discussion . . . . . 160
7 Appendix: proof of Theorem 2.1 . . . . . 161
8 Appendix: proof of Theorem 3.1 . . . . . 163
9 Appendix: Lemma 9.1 . . . . . 164
10 Appendix: proof of Theorem 4.1 . . . . . 166

**6 Conclusion** **171**

**List of Works** **177**
1 Articles . . . . . 177
2 Conferences . . . . . 177
2.1 Peer-reviewed conferences . . . . . 177
2.2 Invited conference . . . . . 178
3 R packages . . . . . 178
3.1 Package SNPSetSimulations . . . . . 178
3.2 Package MGFRTest . . . . . 179
3.3 Package GeneEnvInteractions . . . . . 179
3.4 Package GeneGeneInteractions . . . . . 179
3.5 Package AdaptivePrediction . . . . . 180

**Appendix A Simulation of SNP Data: R Package SNPSetSimulations 181**
1 Introduction . . . . . 181
2 Generating correlated genotype data . . . . . 182
3 Generating phenotypes . . . . . 183
4 Generating samples of genotype and phenotype data . . . . . 185

**Appendix B Implementation of the MGF-R Test: R Package MGFRTest 187**
1 Introduction . . . . . 187
2 Computing the vector of association test statistics between a gene and a phenotype . . . . . 188
3 Computing the $p$-value of the MGF-R test . . . . . 189
4 Power study . . . . . 189

**Appendix C Gene - Environment Interaction Effects Detection: R Package GeneEnvInteractions 193**
1 Introduction . . . . . 193
2 Generating samples under a gene - environment interaction model . . 193
3 Computing the test statistics vector corresponding to a gene - environment interaction model . . . . . 195

**Appendix D Gene - Gene Interaction Effects Detection: R Package GeneGeneInteractions 197**
1 Introduction . . . . . 197
2 Generating samples under a gene - gene interaction model . . . . . 197
3 Computing the test statistics vector corresponding to a gene - gene interaction model . . . . . 199

**Appendix E Implementation of the Adaptive Regression Predictor: R Package AdaptivePrediction 201**

**Bibliography 203**

# List of Figures

1 Taux de mauvais classement des règles LDA et DDA sur les jeux de données `Leukemia` et `Colon` . . . . . 10
2 Coordonnées des vecteurs moyens $\boldsymbol{\mu_1}$ et $\boldsymbol{\mu_2}$ . . . . . 18
3 Exemples de structure de dépendance de jeux de données publics . . 23

1.1 Estimated misclassification rates for the LDA and DDA rules on the Leukemia and Colon datasets . . . . . 33
1.2 Coordinates of the mean vectors $\boldsymbol{\mu_1}$ and $\boldsymbol{\mu_2}$ . . . . . 41
1.3 Examples of dependence structures from public datasets . . . . . 46

2.1 Power of each test in two different scenarios . . . . . 61
2.2 Power of the oracle test statistic $T_2(\boldsymbol{\mathcal{Z}}^*, \boldsymbol{h}^*)$, the $L^2$-norm test and Hotelling's test . . . . . 64
2.3 Dependence structures of the DTD1, KCNN3 and PDZRN4 genes . . 67
2.4 Power curves for several scenarios on genes PDZRN4, KCNN3 and DTD1 . . . . . 73

3.1 Power curves for the detection of gene - environment interaction effects under several scenarios on genes PDZRN4, KCNN3 and DTD1 . 95

4.1 Correlation matrices of 5 SNPs from genes PDZRN4 and KCNN3 and corresponding correlation matrix for the test statistics vector . . . . . 108
4.2 Proportion $\psi$ of coefficients of $\boldsymbol{\Sigma}$ lower than $\varepsilon$ compared to proportions $a_1$ and $a_2$ of coefficients of $\boldsymbol{\Sigma_1}$ and $\boldsymbol{\Sigma_2}$ lower than $\varepsilon$ considering three different pairs of genes . . . . . 112
4.3 Heatmap of the lower bound $\hat{\psi}$ as a function of $a$ and $b$ . . . . . 113
4.4 Power curves for the detection of gene - gene interaction effects under scenarios 1 to 4 (see Table 4.1) . . . . . 118

4.5 Correlation matrices of 5 SNPs from genes PDZRN4 and KCNN3 and corresponding correlation matrix for the test statistics vector (dummy coding) . . . 122
4.6 Power curves under scenario 1 (see Table 4.1) using the dummy coding - based tests . . . 123
4.7 Power curves under a simulation scenario involving recessive - recessive effects . . . 125
4.8 Power curves for the omnibus combining methods . . . 126

5.1 Lower bound for $R^2_N$ as a function of $R_{opt}$ . . . 137
5.2 Distribution of squared correlations between the predicted and observed values of the response for the prediction of the age of a patient by the genomic profiles in the illustrative dataset, over 50 random splittings of the dataset in a 10-fold cross-validation set-up. Four prediction methods are compared: $L_N(\hat{\boldsymbol{Z}})$, $L_{OLS}(\hat{\boldsymbol{Z}})$, Ridge and PLS. 143
5.3 Squared correlation between the response variable and $L_{OLS}(\hat{\boldsymbol{Z}})$ (OLS), $L_N(\hat{\boldsymbol{Z}})$ (Naive) and the optimal predictor (Optimal) within $\mathcal{L}$ in two illustrative scenarios introduced in Section 2 based on Lu et al. (2004) data. . . . 152
5.4 Weights for the prediction of the age of a patient using the 403 gene expression levels with the data introduced in Lu et al. (2004) by the optimal predictor in $\mathcal{L}$ and the OLS and naive predictors. The weights after the 50th explanatory variable are not shown since they remain constant until the 403rd value. . . . 153
5.5 Correlation matrix of the explanatory variables in the simulation study (estimated on the `wine` dataset from the R package `cggd` (Zhang and Melnik, 2012)) . . . 155
5.6 Simulation study results: mean RMSEP for scenarios 1 to 4 . . . 157
5.7 Boxplots of 10-fold cross-validation RMSEP for the orange juice dataset 158
5.8 Boxplots of 10-fold cross-validation RMSEP for the soil dataset . . . 159

# List of Tables

1 Puissance estimée de deux méthodes de test global sous deux configurations différentes . . . . . 19

1.1 Estimated power of two global testing methods under two different configurations . . . . . 42

2.1 Definitions of popular aggregation methods . . . . . 58
2.2 Empirical type I error rates of the MGF-R test (1) . . . . . 68
2.3 Empirical type I error rates of the MGF-R test (2) . . . . . 68
2.4 Bonferroni corrected $p$-values for gene RSPO2 (dogs dataset) . . . . . 70
2.5 Corrected $p$-values corresponding to genes detected by at least one test (Crohn's disease dataset) . . . . . 72

3.1 Type I error rates of the minP procedure under several scenarios using permutations or parametric bootstrap (confidence intervals in brackets) 87
3.2 Estimated type I error rates of the MGF-R test under several scenarios on gene PDZRN4 . . . . . 90
3.3 Estimated type I error rates of the MGF-R test under several scenarios on gene KCNN3 . . . . . 91
3.4 Estimated type I error rates of the MGF-R test under several scenarios on gene DTD1 . . . . . 92

4.1 Simulation scenarios (causal pairs and corresponding effects) for genes PDZRN4 and KCNN3 (scenarios 1 and 2) and for genes PDZRN4 and DTD1 (scenarios 3 and 4) . . . . . 115
4.2 Estimated type I error rates of the MGF-R test under several scenarios for gene - gene interaction testing . . . . . 116

5.1 Mean squared correlation between the predicted and observed values of the response in the test dataset over 1000 simulations for each of 4 scenarios (two patterns of association parameters, two values for $n$) and four prediction methods (Naive, OLS, Ridge and PLS) . . . . . 142

# Remerciements

Mes premiers remerciements vont naturellement à mes directeurs de thèse, David Causeur et Mathieu Emily. Je vous remercie pour tout ce que vous m'avez appris et pour la confiance que vous m'avez accordée dès que je suis arrivé pour mon stage de fin d'études.

Je remercie ensuite Ruth Heller et Nathalie Vialaneix, les rapportrices de cette thèse, pour l'intérêt qu'elles ont accordé à ce travail, ainsi que pour leurs conseils, remarques et suggestions qui ont permis d'améliorer ce manuscrit. Merci également à Pierre Neuvial et Madison Giacofci pour avoir accepté de compléter le jury. Enfin, merci à Eric Matzner-Løber pour avoir présidé le jury.

Merci à l'ensemble des membres de l'unité pédagogique de mathématiques appliquées d'Agrocampus Ouest pour m'avoir accueilli pendant ces presque quatre ans. Cela a été un réel plaisir de travailler parmi vous.

Je profite de ces quelques lignes pour remercier quelques personnes que je n'aurais pas pu rencontrer sans cette thèse. Je pense en particulier à Margot Brard et Marie Morvan pour nos échanges sur nos galères doctorales respectives, et Chloé Friguet. Je remercie aussi les stagiaires qui sont passés dans le même bureau que moi au cours de ma thèse, en particulier Ousmane à qui je souhaite bon courage. Je remercie également mes enseignants de licence et de master, en particulier Magalie Fromont, Eric, Nicolas Jégou, Alain Mom, Pierre-André Cornillon, Laurent Rouvière et bien sûr Mathieu.

Enfin, je remercie ma famille et ma belle-famille pour leur soutien. Merci enfin à Océane pour ses encouragements, son soutien et sa patience.

# Résumé

## Sommaire


**1 Contexte général . . . . . 4**

**2 Prise en compte de la dépendance pour les problèmes de prédiction . . . . . 5**

2.1 Prédiction en petite dimension . . . . . 5

2.2 Prédiction en grande dimension ou en situation de forte dépendance : prendre en compte ou ignorer la dépendance 7

**3 Prise en compte de la dépendance en test global . . . . . 13**

3.1 Test global dans un modèle de régression standard . . . . 13

3.2 Test global en grande dimension ou en présence de forte dépendance . . . . . 14

3.3 Test global et prise en compte de la dépendance en études d'association pangénomiques . . . . . 17

**4 Quelques exemples de structures de dépendance spécifiques . . . . . 20**

**5 Organisation de la thèse . . . . . 22**

# 1 Contexte général

La prédiction d'une variable réponse $Y$ à partir d'un profil multivarié de variables explicatives $\boldsymbol{X} \in \mathbb{R}^p, p \geq 1$, et la construction d'un test de significativité de la relation entre $\boldsymbol{X}$ et $Y$ sont deux tâches fondamentales de la méthodologie statistique qui ont été nettement impactées par l'émergence des données de grande dimension ou fortement corrélées. Bien que ces deux méthodologies ne visent pas le même but, elles partagent des similarités, étant donné qu'elles reposent toutes les deux sur une estimation du modèle de régression liant $Y$ à $\boldsymbol{X}$. En pratique, une procédure complète d'analyse d'un jeu de données associe généralement l'évaluation de la performance de prédiction et des tests de comparaisons de modèles, fournissant deux points de vue complémentaires sur la relation entre $Y$ et $\boldsymbol{X}$. En épidémiologie génétique par exemple, le profil de variables explicatives est composé de génotypes pour des milliers de marqueurs, appelés *Single Nucleotide Polymorphisms* (SNPs), mesurés le long du génome, et la variable réponse est généralement un statut (sain ou non) associé à une maladie donnée. Par conséquent, tester l'association entre un profil génotypique et la maladie peut donner une idée à propos du déterminisme génétique de la maladie étudiée, tandis que la prédiction du statut à partir des génotypes vise à établir un pronostic concernant la maladie, par exemple en médecine personnalisée.

Les conclusions générales de cette thèse sont valides ou peuvent être étendues à un large champ de modèles de régression, mais elles seront essentiellement développées sous l'hypothèse d'un modèle paramétrique : $\mathbb{E}[Y|\boldsymbol{X} = \boldsymbol{x}] = f(\boldsymbol{x}; \boldsymbol{\beta}, \boldsymbol{b})$, où $f$ est une fonction de régression, connue à un vecteur $\boldsymbol{\beta}$ de paramètres d'association entre $\boldsymbol{X}$ et $Y$ près, et éventuellement à un vecteur $\boldsymbol{b}$ de paramètres additionnels près.

Dans les sections suivantes, la prise en compte de la dépendance est discutée pour des problèmes de prédiction et de test global. Il est d'abord montré que les méthodes optimales dans un cadre théorique pour l'analyse discriminante et la régression prennent explicitement en compte la dépendance en décorrélant les variables explicatives. Pour des problèmes de petite dimension, les contreparties empiriques de ces prédicteurs théoriques prennent naturellement en compte la dépendance par la décorrélation. Néanmoins, dans des situations de grande dimension, ces prédic-

teurs ne peuvent pas être directement utilisés. Des approches plus sophistiquées ont été introduites pour contourner ce problème ; nous montrons dans la suite que plusieurs de ces prédicteurs peuvent être considérés comme étant basés sur une prise en compte adaptative de la dépendance. La prise en compte de la dépendance est ensuite discutée pour des problèmes de test global. De façon similaire aux problèmes de prédiction, nous commençons par une revue des tests optimaux dans un contexte de faible dimension et montrons que ces méthodes prennent explicitement en compte la dépendance en décorrélant les variables explicatives. Toutefois, il est intéressant de noter que lorsqu'une forte dépendance est observée entre les variables explicatives, de façon semblable aux situations de grande dimension, ces tests optimaux ne peuvent pas être utilisés. Néanmoins, les points de vue et stratégies proposés par les différents auteurs étant très variés, la question de la prise en compte ou non de la dépendance demeure ouverte, en particulier en situation de forte dépendance. Dans la suite, une situation de forte dépendance est caractérisée par le fait que la matrice de corrélation des variables explicatives peut être approximée de manière satisfaisante par une approximation de rang réduit en n'utilisant qu'un nombre réduit de ses vecteurs propres.

# 2 Prise en compte de la dépendance pour les problèmes de prédiction

## 2.1 Prédiction en petite dimension

À partir de $n$ observations indépendantes $\{(\boldsymbol{X_i}, Y_i)\}_{1\leq i\leq n}$ des variables explicatives et de la variable réponse, la prédiction consiste à estimer une règle fournissant une valeur prédite $\hat{Y}^*$ pour un individu ayant un profil $\boldsymbol{X} = \boldsymbol{x^*}$ et une valeur associée $Y^*$ de la variable réponse. La règle construite est telle qu'une fonction de perte théorique $\mathbb{E}[\ell(Y^*, \hat{Y}^*)]$, mesurant l'écart entre $Y^*$ et $\hat{Y}^*$, est aussi faible que possible. La définition de la fonction de perte $\ell(Y^*, \hat{Y}^*)$ dépend de la nature de la réponse. Dans la suite, nous considérons les deux cas possibles d'une variable réponse continue ou catégorielle.

Considérons tout d'abord le cas particulier du contexte de l'Analyse Discriminante Linéaire (LDA), où $Y \in \{0, 1\}$ est une variable de classe (deux classes possibles) et, conditionnellement à $Y$, $\boldsymbol{X}$ est distribuée selon une loi normale multivariée avec la même matrice de variance-covariance intra-groupe $\boldsymbol{\Sigma_x}$. Dans ce cas, $f(\boldsymbol{x}; \boldsymbol{\beta}, \boldsymbol{b}) = \text{logit}^{-1}\{\log(\pi_1/\pi_0) + (\boldsymbol{x} - \boldsymbol{\mu_x})'\boldsymbol{\Sigma_x^{-1}}\boldsymbol{\delta})\}$, où $\text{logit}^{-1} : u \mapsto 1/\{1 + \exp(-u)\}$, $\pi_k = \mathbb{P}(Y = k)$ est la probabilité d'appartenance au groupe $k$, $\boldsymbol{\mu_x}$ est l'espérance non conditionnelle de $\boldsymbol{X}$ et $\boldsymbol{\delta}$ est la différence entre les espérances intra-groupe de $\boldsymbol{X}$. Dans ce contexte, la probabilité de mauvais classement $\mathbb{E}[\ell(Y^*, \hat{Y}^*)] = \mathbb{P}(\hat{Y}^* \neq Y^*)$ est généralement considérée comme la fonction de perte la plus adaptée. Si $\boldsymbol{\Sigma_x}$, $\boldsymbol{\delta}$ et $\boldsymbol{\mu_x}$ sont supposés connus, le classifieur de Bayes, consistant à seuiller le score linéaire $L(\boldsymbol{X^*}) = \log(\pi_1/\pi_0) + (\boldsymbol{X^*} - \boldsymbol{\mu_x})'\boldsymbol{\Sigma_x^{-1}}\boldsymbol{\delta}$, minimise la probabilité de mauvais classement. Par conséquent, dans ce contexte théorique où les paramètres de la distribution jointe de $\boldsymbol{X}$ et $Y$ sont supposés connus, l'expression explicite de la règle de classification optimale ne laisse aucun doute concernant la prise en compte de la dépendance entre les variables explicatives, mesurée par $\boldsymbol{\Sigma_x}$.

Il est intéressant de remarquer que le score de classification linéaire optimal $L(\boldsymbol{X^*})$ peut être reformulé comme $L(\boldsymbol{X^*}) = \log(\pi_1/\pi_0) + \{\boldsymbol{W_x}(\boldsymbol{X^*} - \boldsymbol{\mu_x})\}'\boldsymbol{W_x}\boldsymbol{\delta} = \log(\pi_1/\pi_0) + \boldsymbol{\tilde{X}^{*\prime}}\boldsymbol{\tilde{\delta}}$, où $\boldsymbol{W_x}$ est une matrice $p \times p$ telle que $\boldsymbol{W_x'W_x} = \boldsymbol{\Sigma_x^{-1}}$ (ou, de façon équivalente, $\boldsymbol{W_x\Sigma_x W_x'} = \mathbf{I}_p$), $\boldsymbol{\tilde{\delta}} = \boldsymbol{W_x\delta}$ et $\boldsymbol{\tilde{X}^*} = \boldsymbol{W_x}(\boldsymbol{X^*} - \boldsymbol{\mu_x})$. La pré-multiplication de $\boldsymbol{X^*} - \boldsymbol{\mu_x}$ par $\boldsymbol{W_x}$ est appelée blanchiment (*whitening*) ou décorrélation, car la matrice de variance-covariance intra-groupe du vecteur ainsi transformé $\boldsymbol{\tilde{X}^*}$ est $\boldsymbol{W_x\Sigma_x W_x'} = \mathbf{I}_p$ (Kessy et al., 2018). Le score de classification optimal $L(\boldsymbol{X^*})$ étant une combinaison linéaire des coordonnées du vecteur blanchi des variables explicatives, nous pourrions en conclure que le meilleur moyen de prendre en compte la dépendance est de décorréler les variables explicatives.

Considérons à présent le contexte de la régression linéaire, où $Y$ est une variable continue. Pour des raisons de simplicité, supposons que les variables explicatives et la réponse sont de moyenne nulle et de variance unité. Supposons de plus que :

$$\begin{pmatrix} \boldsymbol{X} \\ Y \end{pmatrix} \sim \mathcal{N}_{p+1}\left(\begin{pmatrix} \mathbf{0} \\ 0 \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma_x} & \boldsymbol{\sigma_{xy}} \\ \boldsymbol{\sigma_{xy}'} & 1 \end{pmatrix}\right)$$

où $\boldsymbol{\Sigma_x}$ est la matrice de corrélation de $\boldsymbol{X}$ et $\boldsymbol{\sigma_{xy}}$ est le $p$-vecteur dont la $i$-ème coordonnée est la corrélation entre $Y$ et la $i$-ème variable explicative $X_i$. Dans ce cadre, l'erreur quadratique moyenne de prédiction $\mathbb{E}[\ell(Y^*, \hat{Y}^*)] = \mathbb{E}[(Y^* - \hat{Y}^*)^2]$

est souvent choisie comme fonction de perte. En supposant que les paramètres sont connus, le meilleur prédicteur linéaire sans biais est $L(\boldsymbol{X}^*) = \boldsymbol{X}^{*\prime}\boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1}\boldsymbol{\sigma}_{\boldsymbol{xy}} = (\boldsymbol{W}_{\boldsymbol{x}}\boldsymbol{X}^*)'(\boldsymbol{W}_{\boldsymbol{x}}\boldsymbol{\sigma}_{\boldsymbol{xy}})$ où $\boldsymbol{W}_{\boldsymbol{x}}'\boldsymbol{W}_{\boldsymbol{x}} = \boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1}$. Par conséquent, le meilleur prédicteur linéaire implique une étape de décorrélation des variables explicatives, comme dans le contexte de la LDA.

Dans le contexte de la LDA comme dans celui de la régression, les prédicteurs optimaux nécessitent l'utilisation de paramètres inconnus dans les problèmes réels. L'idée la plus directe et la plus naturelle est de les remplacer par les estimateurs des moments correspondants, calculés sur l'échantillon observé. Sous certaines conditions, notamment que la matrice de variance-covariance estimée des variables explicatives n'est pas singulière, les prédicteurs ainsi obtenus (la règle de Fisher en LDA et le prédicteur des moindres carrés ordinaires - *Ordinary Least Squares*, *OLS* - en régression) sont connus pour avoir des performances généralement bonnes. Finalement, il s'avère que pour la LDA comme pour la régression, la prédiction est améliorée en décorrélant les variables explicatives.

## 2.2 Prédiction en grande dimension ou en situation de forte dépendance : prendre en compte ou ignorer la dépendance

Malheureusement, dans un contexte de grande dimension, où $n < p$, et pour une structure de dépendance arbitraire, il n'existe pas de procédure uniformément optimale. Une grande diversité de procédures existe, avec des points de vue divers concernant la prise en compte de la dépendance entre les variables explicatives, et sans solution clairement indiscutable. Curieusement, le principal problème rencontré en grande dimension, ou en large dimension (contexte où $n > p$ mais $n$ n'est pas beaucoup plus grand que $p$), est le même que celui rencontré dans des situations de forte dépendance : la matrice de corrélation (ou de covariance) des variables explicatives ne peut être inversée, ou son inverse a une grande variance, impactant ainsi la stabilité des procédures (Hoerl and Kennard, 1970). De nombreuses méthodes ont été proposées pour remédier à ce problème. Plusieurs sont passées en revue dans la suite, pouvant être utilisées (au sens où le prédicteur peut être calculé) en grande

dimension ou en forte dépendance.

La pratique statistique de la régression hérite des procédures construites dans des cadres réguliers où $n > p$ la conviction que, si les variables explicatives sont corrélées, alors cette corrélation ne devrait pas être ignorée pour la construction de règles de prédiction ou de méthodes de test. Bien que cette conviction soit parfois confirmée dans des cadres de grande dimension, elle est aussi remise en question par de nombreux auteurs. En particulier, des prédicteurs dits "naïfs" ignorant la dépendance offrent parfois des performances clairement meilleures que les approches plus sophistiquées. Ce point est exposé plus en détail en résumant des arguments en faveur de la prise en compte ou non de la dépendance.

### 2.2.1 LDA en grande dimension

Considérons de nouveau le contexte de la LDA à deux groupes introduite dans la section précédente. La règle de prédiction de Fisher (Fisher, 1936) est déduite du score linéaire de classification de Bayes en remplaçant les paramètres inconnus par les estimateurs $\boldsymbol{S_x}$, $\hat{\boldsymbol{\delta}}$ et $\hat{\boldsymbol{\mu}}_{\boldsymbol{x}}$ de $\boldsymbol{\Sigma_x}$, $\boldsymbol{\delta}$ et $\boldsymbol{\mu_x}$, respectivement :

$$L_{\mathrm{LDA}}(\boldsymbol{X^*}) = \log(\hat{\pi}_1/\hat{\pi}_0) + (\boldsymbol{X^*} - \hat{\boldsymbol{\mu}}_{\boldsymbol{x}})' \boldsymbol{S}_{\boldsymbol{x}}^{-1} \hat{\boldsymbol{\delta}},$$

où $\hat{\pi}_k = \#\{i = 1, \ldots, n,\ Y_i = k\}/n$. En tant que contrepartie empirique de la règle de Bayes, la règle de Fisher est asymptotiquement optimale pour la minimisation du taux de mauvais classement.

Cependant, dans une étude comparative de méthodes de prédiction appliquées à une sélection de jeux de données d'expression de gènes en grande dimension, Dudoit et al. (2002) ont introduit une version modifiée de la règle de Fisher en remplaçant $\boldsymbol{S_x}$ par $\boldsymbol{D_s} = \mathrm{diag}(\boldsymbol{S_x})$, ignorant ainsi la corrélation entre les variables explicatives. Les performances de la règle ainsi obtenue, nommée *Diagonal Discriminant Analysis* (DDA), peuvent être dans certains cas largement meilleures que celles de la règle de Fisher, bien que l'hypothèse d'indépendance sur laquelle repose la règle DDA soit évidemment fausse. Ce point est confirmé par Bickel and Levina (2004), qui établissent dans un cadre théorique des conditions sous lesquelles les performances non-asymptotiques de la DDA sont meilleures que celles de la LDA, notamment

dans des situations de grande dimension (voir aussi Efron (2009); Tibshirani et al. (2003); Bodnar and Okhrin (2011)).

Néanmoins, même dans des cas de grande dimension, les performances de la LDA ne sont pas forcément inférieures à celles de la DDA. Ce point est illustré par deux jeux de données d'expression de gènes en grande dimension, donc appartenant au même champ d'application et générés par des technologies similaires. Les jeux de données sont disponibles dans le package R `plsgenomics` (Boulesteix et al., 2018). Le premier, `Leukemia`, est également inclus dans l'étude comparative de Dudoit et al. (2002), et contient 3 051 variables explicatives mesurées pour 38 individus. Le second, `Colon`, contient 2 000 variables pour 62 individus. Dans les deux cas, la variable réponse est une variable de statut, sain ou non, d'un patient à propos d'une maladie donnée. Étant donné que le nombre de variables explicatives est largement supérieur au nombre d'individus, l'inverse de $\boldsymbol{S_x}$ dans l'expression du score de la LDA est remplacé par l'inverse généralisé de Moore-Penrose (voir Bickel and Levina (2004)).

Pour comparer les performances de la LDA et de la DDA, le taux de mauvais classement est estimé par une procédure de validation croisée 10 blocs. La procédure de validation croisée est répétée sur 50 partitions aléatoires en 10 blocs du jeu de données. Les boîtes de dispersion des 50 taux de mauvais classement obtenus sont représentées sur la Figure 1. Conformément aux résultats de Dudoit et al. (2002), la DDA est bien meilleure que la LDA sur le jeu de données `Leukemia`. Le taux de mauvais classement moyen de la DDA est d'environ 2,5%, contre approximativement 9% pour la LDA. En revanche, sur le jeu de données `Colon`, les performances de la LDA sont largement supérieures : le taux de mauvais classement moyen de la LDA est d'environ 22%, contre 38% pour la DDA, dont les performances sont donc médiocres. Par conséquent, bien que les deux jeux de données semblent similaires au regard de leur nature et leur champ d'application, choisir la DDA plutôt que la LDA (ou l'inverse) donnerait des performances sous-optimales pour un jeu de données.

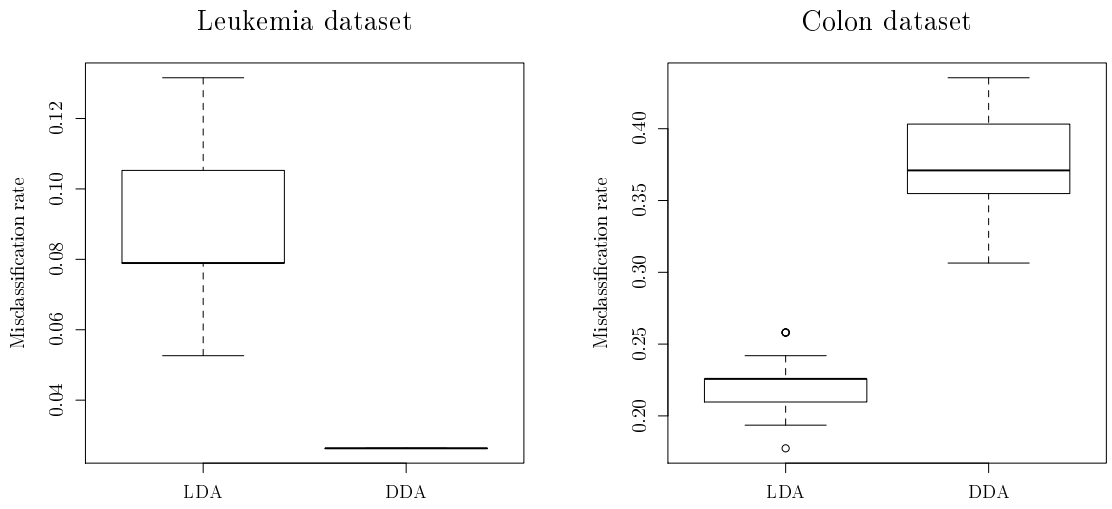


Figure 1 : Boîtes de dispersion des taux de mauvais classement des règles LDA et DDA sur les jeux de données `Leukemia` et `Colon` ; les taux de mauvais classement sont calculés en utilisant 50 procédures de validation croisée 10 blocs

### 2.2.2 Régression en grande dimension

Dans le contexte de la régression, de nombreuses variantes des moindres carrés ordinaires (OLS), ou, plus généralement, de la procédure d'estimation du maximum de vraisemblance, ont été proposées pour contourner les problèmes liés à la grande dimension. Parmi les méthodes les plus populaires, certaines sont basées sur une estimation *shrinkée* de la matrice de variance-covariance des variables explicatives, soit par la pénalisation du critère des moindres carrées ordinaires en régression Ridge (Hoerl and Kennard, 1970) ou en considérant une modélisation de rang réduit en régression sur composantes principales (PCR) (Jolliffe, 1982) ou en régression des moindres carrés partiels (PLSR) (Wold et al., 1983, 1984).

En supposant sans perte de généralité que les variables explicatives et la variable réponse ont été préalablement centrées et réduites, le prédicteur des moindres carrés ordinaires peut s'écrire comme $L_{\text{OLS}}(\boldsymbol{X}^*) = \boldsymbol{X}^{*\prime}\boldsymbol{S}_{\boldsymbol{x}}^{-}\boldsymbol{s}_{\boldsymbol{xy}}$ où $\boldsymbol{S}_{\boldsymbol{x}}$ et $\boldsymbol{s}_{\boldsymbol{xy}}$ sont les estimateurs des moments de $\boldsymbol{\Sigma}_{\boldsymbol{x}}$ et $\boldsymbol{\sigma}_{\boldsymbol{xy}}$, respectivement, et $\boldsymbol{S}_{\boldsymbol{x}}^{-}$ est l'inverse généralisée de $\boldsymbol{S}_{\boldsymbol{x}}$. Une des premières méthodes introduites pour aborder le problème de la singularité de $\boldsymbol{S}_{\boldsymbol{x}}$ est la régression Ridge. Le prédicteur Ridge peut être écrit

comme :

$$L_{\text{Ridge}}(\boldsymbol{X}^*) = \boldsymbol{X}^{*\prime}(\boldsymbol{S}_{\boldsymbol{x}} + \kappa \mathbf{I}_p)^{-}\boldsymbol{s}_{\boldsymbol{xy}}$$

où $\kappa$ est le coefficient de pénalité Ridge, généralement choisi par l'optimisation d'un critère par une procédure de validation croisée. Il peut être souligné que le prédicteur Ridge peut être vu comme un compromis entre une prise en compte de la dépendance (comme pour le prédicteur des moindres carrés ordinaires) et une ignorance de celle-ci. Selon la valeur du paramètre de régularisation $\kappa$, le prédicteur est plus ou moins proche du prédicteur OLS. En effet, considérons la décomposition en valeurs propres de $\boldsymbol{S}_{\boldsymbol{x}}$ : $\boldsymbol{S}_{\boldsymbol{x}} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}'$, $\boldsymbol{\Lambda} = \text{diag}(\lambda_i)$. Alors

$$L_{\text{Ridge}}(\boldsymbol{X}^*) = \boldsymbol{X}^{*\prime}\boldsymbol{U}\,\text{diag}\left(\frac{1}{\lambda_i + \kappa}\right)\boldsymbol{U}'\boldsymbol{s}_{\boldsymbol{xy}}.$$

Si $\kappa = 0$, de façon évidente $L_{\text{Ridge}}(\boldsymbol{X}^*) = L_{\text{OLS}}(\boldsymbol{X}^*)$. D'autre part, si $\kappa \gg \lambda_1$ où $\lambda_1$ est la plus grande valeur propre, alors $\lambda_i + \kappa \simeq \kappa$ et $L_{\text{Ridge}}(\boldsymbol{X}^*) \propto \boldsymbol{X}^{*\prime}\boldsymbol{s}_{\boldsymbol{xy}}$. Dans ce cas, le prédicteur Ridge fait abstraction de la dépendance entre les variables explicatives.

Opgen-Rhein and Strimmer (2007b) (voir également Zuber and Strimmer (2011)) ont aussi proposé d'aborder la singularité de $\boldsymbol{S}_{\boldsymbol{x}}$ en utilisant des estimateurs biaisés des variances et corrélations apparaissant dans le prédicteur OLS (Schäfer and Strimmer, 2005). Notons $\boldsymbol{D}$ la matrice diagonale contenant les variances estimées des variables explicatives, $\boldsymbol{R}_{\boldsymbol{x}} = \boldsymbol{D}^{-1/2}\boldsymbol{S}_{\boldsymbol{x}}\boldsymbol{D}^{-1/2}$ la matrice de corrélation estimée des variables explicatives et $\boldsymbol{r}_{\boldsymbol{xy}} = \boldsymbol{D}^{-1/2}\boldsymbol{s}_{\boldsymbol{xy}}/s_y$ le vecteur des corrélations estimées entre la variable réponse et chaque variable explicative. Ensuite, en remplaçant les estimateurs usuels par des estimateurs biaisés, Opgen-Rhein and Strimmer (2007b) ont introduit le prédicteur suivant :

$$L_{\text{SLM}}(\boldsymbol{X}^*) = \boldsymbol{X}^{*\prime}(s_y^{(\tau)}(\boldsymbol{D}^{(\boldsymbol{\tau})})^{-1/2}(\boldsymbol{R}_{\boldsymbol{x}}^{(\boldsymbol{\gamma})})^{-1}\boldsymbol{r}_{\boldsymbol{xy}}^{(\boldsymbol{\gamma})})$$

où $\gamma$ et $\tau$ sont des paramètres de *shrinkage*. En particulier, $\boldsymbol{R}_{\boldsymbol{x}}^{(\boldsymbol{\gamma})} = \gamma\mathbf{I}_p + (1-\gamma)\boldsymbol{R}_{\boldsymbol{x}}$ où $\gamma \in [0,1]$. Il peut ainsi être remarqué que pour de grandes valeurs de $\gamma$, $\boldsymbol{R}_{\boldsymbol{x}}^{(\boldsymbol{\gamma})}$ tend vers la matrice identité et la dépendance entre les variables explicatives tend à être ignorée. Néanmoins, $\boldsymbol{r}_{\boldsymbol{xy}}^{(\boldsymbol{\gamma})} = (1-\gamma)\boldsymbol{r}_{\boldsymbol{xy}}$ ; par conséquent, la dépendance entre les variables explicatives et la réponse tend également à être ignorée, et le vecteur de coefficients estimés tend vers $\mathbf{0}$. Finalement, la réponse est simplement prédite par sa valeur moyenne. À l'opposé, si $\gamma = 0$, le prédicteur écrit ci-dessus est similaire au

prédicteur OLS, la seule différence résidant dans l'estimation des variances. Selon la valeur du paramètre $\gamma$, $\boldsymbol{R}_{\boldsymbol{x}}^{(\gamma)}$ est plus ou moins proche de $\boldsymbol{R}_{\boldsymbol{x}}$. En particulier, $\boldsymbol{R}_{\boldsymbol{x}}^{(\gamma)} = \boldsymbol{W}(\gamma \mathbf{I}_p + (1-\gamma)\boldsymbol{\Psi})\boldsymbol{W}'$ où $\boldsymbol{W}\boldsymbol{\Psi}\boldsymbol{W}'$ est la décomposition en valeurs propres de $\boldsymbol{R}_{\boldsymbol{x}}$, ce qui signifie que les valeurs propres de $\boldsymbol{R}_{\boldsymbol{x}}$ qui sont plus grandes que 1 sont réduites. Cette idée a également été appliquée à l'analyse discriminante par Ahdesmäki and Strimmer (2010); Zuber and Strimmer (2009), où l'estimateur de la matrice de variance-covariance dans l'expression du score de Fisher est remplacé par un estimateur biaisé.

Le prédicteur PCR (Jolliffe, 1982) est défini comme suit :

$$L_{\text{PCR}}(\boldsymbol{X}^*) = \boldsymbol{X}^{*\prime} \left( \sum_{i=1}^{k} \frac{1}{\lambda_i} \boldsymbol{u_i} \boldsymbol{u_i'} \right) \boldsymbol{s_{xy}}$$

où $\boldsymbol{u_i}$ est le $i$-ème vecteur propre de $\boldsymbol{S}_{\boldsymbol{x}}$, associé à la valeur propre $\lambda_i$. Le terme $\sum_{i=1}^{k} \frac{1}{\lambda_i} \boldsymbol{u_i} \boldsymbol{u_i'}$, qui peut être vu comme un noyau de variables prédictives latentes, est une approximation de rang réduit de $\boldsymbol{S}_{\boldsymbol{x}}^{-}$. Par conséquent, la PCR permet une prise en compte intermédiaire de la dépendance à travers la dimension $k$ du noyau. En effet, accroître la valeur de $k$ revient à considérer plus de vecteurs propres pour approximer l'inverse de $\boldsymbol{S}_{\boldsymbol{x}}$. De ce fait, quand la valeur de $k$ augmente, $\sum_{i=1}^{k} \frac{1}{\lambda_i} \boldsymbol{u_i} \boldsymbol{u_i'}$ se rapproche de l'inverse de $\boldsymbol{S}_{\boldsymbol{x}}$.

Enfin, le prédicteur PLSR (Wold et al., 1983, 1984), souvent préféré au précédent prédicteur PCR, peut être écrit comme (Helland, 1988; Lingjaerde and Christophersen, 2000; Blazère et al., 2014) :

$$L_{\text{PLSR}}(\boldsymbol{X}^*) = \boldsymbol{X}^{*\prime} \left( \sum_{i=1}^{k} \alpha_i \boldsymbol{S}_{\boldsymbol{x}}^{i-1} \right) \boldsymbol{s_{xy}},$$

où $k$ est le nombre de facteurs PLS latents. En particulier, si $k = 1$, alors $L_{\text{PLSR}}(\boldsymbol{X}^*) \propto \boldsymbol{X}^{*\prime} \boldsymbol{s_{xy}}$ et la dépendance entre les variables explicatives est ignorée. De plus, le terme $\sum_{i=1}^{k} \alpha_i \boldsymbol{S}_{\boldsymbol{x}}^{i-1}$ est une approximation polynomiale de $\boldsymbol{S}_{\boldsymbol{x}}^{-1}$ ; plus $k$ est grand, plus le terme est proche de $\boldsymbol{S}_{\boldsymbol{x}}^{-1}$. Finalement, la quantité de dépendance prise en compte augmente avec $k$.

Pour chacune de ces quatre approches, un hyperparamètre est introduit, que ce soit un terme de régularisation en régression Ridge ou dans l'estimation *shrinkée* de Opgen-Rhein and Strimmer (2007b), ou la dimension d'un noyau de variables latentes prédictives en PCR ou PLSR. Dans chaque cas, il peut être remarqué que cet

hyperparamètre définit une prise en compte intermédiaire de la dépendance entre une ignorance totale et une décorrélation complète. Pour les méthodes de régression biaisée, la dépendance tend à être ignorée pour de grandes valeurs de l'hyperparamètre, et à être considérée pour de faibles valeurs. Au contraire, la quantité de dépendance prise en compte augmente avec le nombre $k$ de variables latentes considérées dans les prédicteurs PCR et PLSR. Cela montre que différentes approches sont adaptées pour une prise en compte flexible de la dépendence dans un cadre de régression. De plus, certaines approches ont également été transposées au cadre de l'analyse discriminante, notamment l'analyse discriminante PLS (Boulesteix, 2004; Gottfries et al., 1995; Barker and Rayens, 2003) ou des méthodes d'analyse discriminante biaisées (Friedman, 1989; Guo et al., 2005), ce qui montre également qu'une prise en compte adaptative de la dépendance peut être une option adéquate dans des contextes nombreux et variés.

# 3 Prise en compte de la dépendance en test global

## 3.1 Test global dans un modèle de régression standard

Dans le même contexte de régression paramétrique que celui considéré précédemment, un test global (Arias-Castro et al., 2011) consiste à tester $H_0 : \boldsymbol{\beta} = \boldsymbol{\beta_0}$ où $\boldsymbol{\beta_0}$ est un vecteur de valeurs cibles. Par exemple, dans le cas particulier du modèle de régression linéaire où, conditionnellement à $\boldsymbol{X} = \boldsymbol{x}$, $Y$ est distribuée selon une loi normale de moyenne $f(\boldsymbol{x}; \boldsymbol{\beta}, b_0) = b_0 + \boldsymbol{x}'\boldsymbol{\beta}$, le test global le plus usuel est le test de significativité de la relation entre $Y$ et $\boldsymbol{X}$, c'est-à-dire le test de $H_0 : \boldsymbol{\beta} = \mathbf{0}$.

Toujours dans le cadre du modèle de régression linéaire, où il est de plus supposé que $n > p$, l'approche traditionnelle de l'analyse de la variance (Fisher, 1925) est le moyen le plus souvent utilisé pour réaliser le test précédent. En effet, le test uniformément le plus puissant pour $H_0$ au sens du lemme de Neyman-Pearson est le test $F$, qui peut s'écrire comme une statistique de Wald (Wald, 1943) : $F = \hat{\boldsymbol{\beta}}'\hat{\boldsymbol{V}}_{\hat{\boldsymbol{\beta}}}^{-1}\hat{\boldsymbol{\beta}}/p$, où $\hat{\boldsymbol{\beta}} = \boldsymbol{S}_{\boldsymbol{x}}^{-1}\boldsymbol{s}_{\boldsymbol{xy}}$ est l'estimateur des moindres carrés ordinaires de $\boldsymbol{\beta}$, $\hat{\boldsymbol{V}}_{\hat{\boldsymbol{\beta}}} = \frac{\hat{\sigma}^2}{n}\boldsymbol{S}_{\boldsymbol{x}}^{-1}$ est la matrice de variance-covariance estimée de $\hat{\boldsymbol{\beta}}$, $\boldsymbol{S}_{\boldsymbol{x}}$ et $\boldsymbol{s}_{\boldsymbol{xy}}$ sont respectivement la matrice de variance-covariance estimée des variables explicatives

et le vecteur des covariances estimées entre les variables explicatives et la variable réponse, et $\hat{\sigma}^2$ et l'estimateur corrigé (sans biais) de la variance résiduelle. Dans ce cadre usuel, où il est supposé que $\boldsymbol{S}_{\boldsymbol{x}}^{-1}$ existe, la dépendance entre les componsantes de $\hat{\boldsymbol{\beta}}$ est héritée de la dépendance entre les variables explicatives, puisque $\hat{\boldsymbol{V}}_{\hat{\boldsymbol{\beta}}} \propto \boldsymbol{S}_{\boldsymbol{x}}^{-1}$.

Ici, que ce soit par la théorie du maximum de vraisemblance ou par des arguments géométriques, une solution optimale est donnée aux problèmes d'estimation et de test : l'estimateur des moindres carrés ordinaires et le test $F$. Cela impose sans aucun doute possible ces approches comme des approches de référence. De façon intéressante, la statistique du test $F$ peut aussi être vue comme la moyenne des carrés des coordonnées d'une version décorrélée du vecteur de paramètres de régression estimés. En effet, soit $\tilde{\boldsymbol{\beta}} = \boldsymbol{W}_{\hat{\boldsymbol{\beta}}}\hat{\boldsymbol{\beta}}$, où $\boldsymbol{W}_{\hat{\boldsymbol{\beta}}}$ est une matrice telle que $\boldsymbol{W}'_{\hat{\boldsymbol{\beta}}}\boldsymbol{W}_{\hat{\boldsymbol{\beta}}} = \hat{\boldsymbol{V}}_{\hat{\boldsymbol{\beta}}}^{-1}$. Alors la matrice de variance-covariance asymptotique de $\tilde{\boldsymbol{\beta}}$ est la matrice identité de dimension $p$ $\mathbf{I}_p$ et $F$ peut être réécrite comme la norme $\mathrm{L}^2$ de $\tilde{\boldsymbol{\beta}}$ divisée par $p$ : $F = ||\tilde{\boldsymbol{\beta}}||_2^2/p$. Comme précédemment, cela mène à la conclusion que la meilleure façon de prendre en compte la dépendance en test global est de décorréler $\hat{\boldsymbol{\beta}}$. Comme $\hat{\boldsymbol{\beta}} = \boldsymbol{S}_{\boldsymbol{x}}^{-1}\boldsymbol{s}_{\boldsymbol{xy}}$, $F \propto \boldsymbol{s}'_{\boldsymbol{xy}}\boldsymbol{S}_{\boldsymbol{x}}^{-1}\boldsymbol{s}_{\boldsymbol{xy}} = ||\boldsymbol{W}_{\boldsymbol{x}}\boldsymbol{s}_{\boldsymbol{xy}}||_2^2$ où $\boldsymbol{W}'_{\boldsymbol{x}}\boldsymbol{W}_{\boldsymbol{x}} = \boldsymbol{S}_{\boldsymbol{x}}^{-1}$. De plus, $\mathrm{Var}(\boldsymbol{s}_{\boldsymbol{xy}}) \propto \boldsymbol{S}_{\boldsymbol{x}}$, ce qui signifie que la statistique optimale $F$ peut aussi être obtenue en décorrélant les estimateurs de pente univariés, concaténés dans le vecteur $\boldsymbol{s}_{\boldsymbol{xy}}$. La structure de dépendance de $\hat{\boldsymbol{\beta}}$ est complètement différente de celle du vecteur d'estimateurs univariés $\boldsymbol{s}_{\boldsymbol{xy}}$. Cependant, les deux vecteurs, après décorrélation, donnent la même statistique optimale.

## 3.2 Test global en grande dimension ou en présence de forte dépendance

Dans un modèle en grande dimension ou en présence de forte dépendance entre les variables explicatives, l'estimateur OLS $\hat{\boldsymbol{\beta}}$ ne peut être calculé, ou est instable du fait de sa grande variance. Sous de telles conditions, pour tester la significativité du vecteur $\boldsymbol{\beta}$, de nombreux auteurs (voir Arias-Castro et al. (2011)) reformulent l'hypothèse nulle comme $H_0 : \ \boldsymbol{\mu} = \mathbf{0}$, où $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_p)'$ est le $p$-vecteur contenant les paramètres de pente dans les modèles de régression univariés liant $Y$ à chaque variable explicative $X_j : \mathbb{E}[Y|X_j = x_j] = m_{0j} + \mu_j x_j$, $m_{0j}$ étant l'ordonnée à l'origine.

En effet, l'estimation de $\boldsymbol{\mu}$ n'est pas affectée par la dimension de l'espace paramétrique, ni par la forte dépendance entre les variables explicatives. Cependant, comme pour $\hat{\boldsymbol{\beta}}$, les coordonnées de l'estimateur $\hat{\boldsymbol{\mu}}$ de $\boldsymbol{\mu}$ héritent leur dépendance de celle des variables explicatives. Cela pose la question de la meilleure façon de prendre en compte la dépendance pour construire un test global sur la base de $\hat{\boldsymbol{\mu}}$.

Par ailleurs, l'estimateur de $\boldsymbol{\mu}$ peut s'exprimer comme $\hat{\boldsymbol{\mu}} = \bar{\boldsymbol{Z}}$, où $\bar{\boldsymbol{Z}}$ est la moyenne des variables $\boldsymbol{Z_i} = Y_i.\boldsymbol{D_s}^{-1}(\boldsymbol{X_i} - \bar{\boldsymbol{X}})$ et $\boldsymbol{D_s} = \text{diag}(\boldsymbol{S_x})$. De plus, pour tout indice $i = 1, \ldots, n$, $\boldsymbol{Z_i}$ est asymptotiquement distribuée selon la loi normale multivariée de moyenne $\boldsymbol{\mu}$ et de matrice de variance-covariance $\boldsymbol{V_Z}$. En outre, tester la significativité de $\boldsymbol{\mu}$ revient à tester la significativité de l'espérance d'un vecteur gaussien $\boldsymbol{Z}$, sur la base de $n$ observations $\boldsymbol{Z_i}$, $i = 1, \ldots, n$. Une solution connue pour ce problème est le test de Hotelling (Hotelling, 1931), basé sur la statistique $T^2 = n\bar{\boldsymbol{Z}}'\boldsymbol{S_Z}^{-1}\bar{\boldsymbol{Z}}$, où $\boldsymbol{S_Z}$ est la matrice de variance-covariance estimée de $\boldsymbol{Z}$. Dans un cadre régulier où $n > p$, ce test est uniformément plus puissant parmi les statistiques invariantes par transformation linéaire de $\boldsymbol{Z}$ (voir Anderson (2003), chapitre 5, section 5.6). Néanmoins, cette solution suppose que $\boldsymbol{S_Z}^{-1}$ existe ; en grande dimension ou en situation de forte dépendance, $\boldsymbol{S_Z}^{-1}$ n'existe pas ou est trop instable pour produire des résultats fiables.

L'approche générale de test global basée sur $\boldsymbol{\mu}$ plutôt que sur $\boldsymbol{\beta}$ est considérée sous l'angle de la recherche de la statistique de test uniformément plus puissante en agrégeant les coordonnées de $\hat{\boldsymbol{\mu}}$. En études d'association pangénomiques, des tests simple marqueur sont parfois agrégés à l'intérieur de blocs de marqueurs adjacents afin d'identifier des régions du génome pouvant être impliquées dans l'apparition d'une maladie. Dans ce contexte, de nombreux articles incluent des études comparatives de méthodes d'agrégation, la plupart d'entre elles ignorant la dépendance entre les coordonnées de $\hat{\boldsymbol{\mu}}$, comme la norme $\text{L}^2$ ou la norme $\text{L}^\infty$ de $\hat{\boldsymbol{\mu}}$ (voir par exemple Wu et al. (2014)). Les conditions pour qu'une procédure d'agrégation soit optimale en grande dimension sont discutées par Donoho and Jin (2004); Hall and Jin (2008); Arias-Castro et al. (2011). Donoho and Jin (2004) considèrent le problème de test global par l'agrégation de statistiques de test $Z_j$, $j = 1, \ldots, p$, sous l'hypothèse que $\boldsymbol{\mathcal{Z}} = (Z_1, \ldots, Z_p)'$ est distribué selon un modèle de mélange gaussien : $\forall j$, $Z_j \sim (1-\varepsilon)\mathcal{N}(0,1) + \varepsilon\mathcal{N}(\delta, 1)$, où le paramètre de mélange $0 \leq \varepsilon \leq 1$ est la proportion d'hypothèses nulles fausses, $\delta \geq 0$ est l'amplitude du signal et $\text{Var}(\boldsymbol{\mathcal{Z}}) = \boldsymbol{\Sigma}$, où

$\boldsymbol{\Sigma}$ est une matrice de variance-covariance définie-positive. Donoho and Jin (2004) introduisent également des bornes de détectabilité sur $\varepsilon$ et $\delta$, en-dessous desquelles le signal est considéré comme rare et faible.

Donoho and Jin (2004) ont proposé la statistique du *Higher Criticism* (HC), définie comme une distance de type Kolmogorov-Smirnov entre la distribution empirique des $p$-values $p_j$ associées aux statistiques de test $Z_j$ et la distribution uniforme :

$$\mathrm{HC} = \max_{1 \leq j \leq p/2} \sqrt{p} \frac{j/p - p_{(j)}}{\sqrt{p_{(j)}(1 - p_{(j)})}}$$

où $p_{(j)}$ est la $j$-ème statistique d'ordre de $(p_j)_{\{1 \leq j \leq p\}}$. En supposant que $\boldsymbol{\Sigma} = \mathbf{I}_p$, Donoho and Jin (2004) montrent que le HC est optimal, au sens où il atteint les meilleures bornes de détectabilité du modèle de signal rare et faible (Ingster, 1997; Donoho and Jin, 2004, 2008). Quand l'hypothèse d'indépendance n'est pas vérifiée, le HC n'est plus optimal et ses propriétés sont sévèrement affectées (Hall and Jin, 2008). Afin de restituer son optimalité, Hall and Jin (2010) proposent de prendre en compte la corrélation entre les statistiques de test par une étape préliminaire de décorrélation de $\boldsymbol{\mathcal{Z}}$. L'optimalité de cette version du HC basée sur une décorrélation, nommée *innovated Higher Criticism*, a été établie par Hall and Jin (2010). En particulier, en observant que la décorrélation amplifie les coordonnées non nulles de l'espérance de $\boldsymbol{\mathcal{Z}}$, Hall and Jin (2010) affirment que la corrélation facilite la détection d'un signal. Néanmoins, Arias-Castro et al. (2011) ont étendu l'optimalité asymptotique (par rapport à $p$) du HC standard sous des conditions de faible dépendance. La différence principale entre Hall and Jin (2008, 2010) et Arias-Castro et al. (2011) est la suivante : dans Hall and Jin (2008, 2010), $\boldsymbol{\mu}$ est supposé vérifier les hypothèses du modèle de signal rare et faible, tandis que le vecteur de paramètres de régression $\boldsymbol{\beta}$ est supposé vérifier ces hypothèses dans Arias-Castro et al. (2011). Bien que cela ne fasse aucune différence lorsque les variables explicatives sont indépendantes, $\boldsymbol{\beta}$ et $\boldsymbol{\mu}$ peuvent avoir des formes très différentes en présence de dépendance. En particulier, pour un vecteur $\boldsymbol{\beta}$ vérifiant les hypothèses du modèle rare et faible, le vecteur $\boldsymbol{\mu}$ ne vérifie pas nécessairement ces hypothèses.

## 3.3 Test global et prise en compte de la dépendance en études d'association pangénomiques

Les méthodes de test global sont couramment utilisées en études d'association pangénomiques (Conneely and Boehnke, 2007; Wu et al., 2014; Luo et al., 2010; Barnett et al., 2017; Basu and Pan, 2011), dans lesquelles elles sont généralement dénommées méthodes de combinaison de tests ou de $p$-values. Comme mentionné précédemment, ces approches visent à découvrir des associations entre des régions du génome et un phénotype (par exemple une maladie) d'intérêt en agrégeant des statistiques de test univariées, chaque statistique univariée correspondant au test d'association entre un marqueur et le phénotype. À l'intérieur d'une région, la dépendance entre les marqueurs et, par conséquent, entre les statistiques de test correspondantes, peut être forte et avoir des formes diverses. Pour ce problème de test, en suivant Arias-Castro et al. (2011), Wu et al. (2014) ont proposé d'utiliser la statistique du HC initial, définie par Donoho and Jin (2004) et dont la définition ignore la dépendance entre les statistiques de test. Wu et al. (2014) ont également montré son optimalité sous différentes hypothèses sur la matrice de corrélation, toujours dans un cadre de signal rare et faible comme introduit par Donoho and Jin (2004). De façon surprenante, Wu et al. (2014) ont même montré qu'une méthode introduite par Luo et al. (2010) et basée sur une étape de décorrélation (similaire à celle introduite par Hall and Jin (2010)) n'est pas optimale. De même, Barnett et al. (2017) avancent qu'une telle étape de décorrélation pourrait détériorer le signal et amplifier le bruit, réduisant ainsi la puissance d'une méthode de test global. Néanmoins, il est montré par Liu and Xie (2018) que si les effets conditionnels d'un modèle linéaire sont plus forts que ses effets marginaux, alors une méthode de test global basée sur les effets marginaux telle que celles proposées par les précédents auteurs peut avoir une faible puissance. En outre, les auteurs ont proposé de pré-multiplier le vecteur d'effets marginaux par l'inverse de sa matrice de variance-covariance, proposant ainsi une transformation similaire à la décorrélation. En effet, la décorrélation consiste à pré-multiplier par l'inverse d'une racine carrée de la matrice de variance-covariance, tandis que dans Liu and Xie (2018), il s'agit simplement de l'inverse. De plus, les auteurs montrent théoriquement et à travers une étude par simulations qu'un test basé sur le vecteur transformé peut être plus puissant que les tests basés sur les effets marginaux. Cela plaide évidemment en faveur d'une prise en compte explicite de la dépendance, d'une

manière similaire à Hall and Jin (2010).

Cette discussion illustre la difficulté à obtenir un consensus concernant la prise en compte de la dépendance. Il apparaît que, selon la situation, décorréler les statistiques de test peut être préjudiciable ou bénéfique.

Nous proposons d'illustrer ce point en utilisant des simulations basées sur des données. Dans la suite, deux vecteurs moyens possibles $\boldsymbol{\mu_1}$ et $\boldsymbol{\mu_2}$ (représentés sur la Figure 2) sont considérés pour un vecteur $\boldsymbol{\mathcal{Z}}$ de statistiques de test. Ensuite, 10 000 vecteurs gaussiens sont simulés, de vecteur moyen $\boldsymbol{\mu_1}$ et de matrice de corrélation identique à celle estimée sur le gène PDZRN4 du génome humain, extrait de données de génomique publiques (Wellcome Trust Case Control Consortium, 2007) (voir Figure 3(a)). De même, 10 000 vecteurs gaussiens sont générés, de vecteur moyen $\boldsymbol{\mu_2}$ et ayant la même matrice de corrélation.

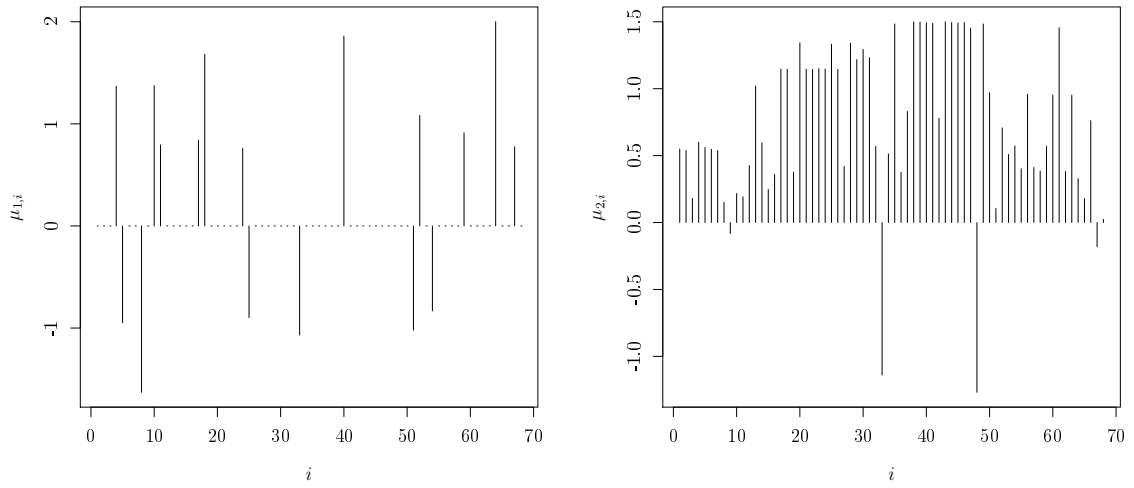


Figure 2 : Coordonnées des vecteurs moyens $\boldsymbol{\mu_1}$ et $\boldsymbol{\mu_2}$

Nous proposons de comparer la puissance de deux méthodes d'agrégation simples sous ces deux configurations. La première méthode est basée sur la norme $\mathrm{L}^2$ du vecteur $\boldsymbol{\mathcal{Z}}$, et la seconde sur la norme $\mathrm{L}^2$ de $\boldsymbol{\mathcal{Z}}$ après une étape préliminaire de décorrélation. Cette étape de décorrélation est réalisée en pré-multipliant $\boldsymbol{\mathcal{Z}}$ par l'inverse d'une racine carrée de la matrice de corrélation de $\boldsymbol{\mathcal{Z}}$, et les $p$-values sont obtenues en utilisant des vecteurs gaussiens simulés de même matrice de corrélation et de vecteur moyen $\mathbf{0}$. La puissance de détection estimée au niveau 0,05 est donnée

pour chaque approche et dans chaque configuration dans le Tableau 1. Dans la première configuration, le test basé sur la décorrélation est bien plus puissant, le test ignorant la dépendance ayant une puissance nulle, tandis que dans la seconde configuration, le test ignorant la dépendance est beaucoup plus puissant. Même si ces résultats ont été obtenus en utilisant des exemples très simples, ils démontrent que le choix de décorréler les statistiques de test ou non ne dépend pas seulement de la structure de dépendance des statistiques de test, mais de la configuration de cette structure de dépendance par rapport au signal d'association entre les variables explicatives et la variable réponse.

Tableau 1 : Puissance estimée de deux méthodes de test global sous deux configurations différentes

| Vecteur moyen | Avec décorrélation | Sans décorrélation |
|---|---|---|
| $\boldsymbol{\mu_1}$ | 0.72 | 0.05 |
| $\boldsymbol{\mu_2}$ | 0.09 | 0.33 |

Pour résumer, pour des problèmes de prédiction comme pour des problèmes de test global, l'impact de la dépendance entre les variables explicatives n'est pas un problème particulier dans des cadres réguliers où $n > p$. En effet, la dépendance est prise en compte explicitement dans l'expression des procédures optimales, pour des structures de dépendance arbitraires. De plus, elles peuvent être vues comme l'application de procédures optimales sous indépendance après une étape de décorrélation du profil de variables explicatives. Toutefois, dans des cadres de grande dimension, la plupart de ces solutions optimales ne peut être utilisée, ce qui pose la question du meilleur moyen de prendre en compte la dépendance dans de telles situations.

De façon surprenante, de nombreux auteurs observent que les procédures ignorant la dépendance entre les variables explicatives donnent de bonnes performances, parfois même supérieures à celles de méthodes construites spécifiquement pour prendre en compte la dépendance (Wu et al., 2014; Barnett et al., 2017; Dudoit et al., 2002; Bickel and Levina, 2004). Cette thèse vise à démontrer que la meilleure façon de prendre en compte la dépendance ne devrait pas se limiter à un choix binaire entre décorrélation et ignorance de la dépendance. En effet, des approches plus flexibles sont requises pour prendre en compte de façon optimale la configura-

tion entre le signal d'association $\boldsymbol{\beta}$ et la structure de dépendance entre les variables explicatives.

Dans le reste de ce manuscrit, ces points seront discutés plus en détail et les conséquences d'une ignorance de la dépendance ou d'une décorrélation seront exposées. Il sera montré que la puissance des méthodes de test global dépend fortement à la fois de la forme du signal et de la matrice de corrélation des statistiques de test. En particulier, pour une matrice de corrélation donnée, deux vecteurs $\boldsymbol{\beta}$ différents donnent deux vecteurs $\boldsymbol{\mu}$ très différents ; bien que ce point puisse paraître évident, il induit la question du choix entre la considération et l'ignorance de la dépendance. Il apparaît que, comme démontré par l'exemple précédent, choisir de décorréler les statistiques de test peut améliorer fortement la puissance, ou peut au contraire la détériorer. Des méthodes adaptatives sont développées et proposées comme alternatives possibles afin de permettre une prise en compte flexible de la dépendance. Cette prise en compte adaptative dépend à la fois de la matrice de corrélation des variables explicatives et du signal d'association de celles-ci avec la variable réponse.

# 4 Quelques exemples de structures de dépendance spécifiques

Bien que nous ne ferons pas d'hypothèses spécifiques sur la forme de la dépendance entre les variables explicatives, les approches proposées seront illustrées principalement dans des champs d'application pour lesquels la prise en compte de la dépendance est d'un intérêt primordial. C'est particulièrement le cas en génomique, où le profil de variables explicatives est composé de milliers de variables génotypiques (SNPs) le long du génome. La matrice de corrélation des SNPs formant le gène PDZRN4 du génome humain (Wellcome Trust Case Control Consortium, 2007) est représentée sur la Figure 3(a). Comme illustré sur cette figure, une part de la dépendance de ces variables est héritée de leur position le long du génome, notamment du fait du mécanisme de déséquilibre de liaison (*linkage disequilibrium*, LD). Bien que chaque variable soit catégorielle avec trois modalités correspondant au nombre (0, 1 ou 2) de copies de l'allèle mineur pour un SNP, ces variables sont généralement considérées comme quantitatives et le coefficient de corrélation de Pearson est considéré

par la plupart des auteurs comme une bonne mesure de la dépendance entre deux variables. La Figure 3(a) montre clairement une structure de dépendance forte et hétérogène, contenant des sous-blocs superposés de SNPs adjacents fortement corrélés. Ce partitionnement en sous-blocs est observé pour la plupart des gènes mais les tailles et nombres de sous-blocs varient d'une région du génome à l'autre. Ces blocs sont souvent nommés blocs de déséquilibre de liaison et il est généralement supposé qu'ils fournissent une partition adaptée du génome (Wall and Pritchard, 2003; Dehman et al., 2015; Nakamoto et al., 2006; Twells et al., 2003).

Dans la recherche d'associations entre des gènes et un phénotype $Y$, les statistiques de test simple marqueur correspondant au test d'association entre chaque SNP et $Y$ sont agrégées pour former une unique statistique de test, correspondant au test d'association entre le gène et $Y$. Comme la même approche est répétée pour des centaines de gènes, celle-ci doit donner des performances stables, quelles que soient la forme de la structure de dépendance et la forme du signal d'association.

Des approches de test global similaires sont utilisées en analyse de la variance fonctionnelle, qui étend l'analyse de la variance usuelle à des situations où la variable réponse est une courbe (voir Zhang (2013)). Pour tester la significativité de la différence moyenne entre des groupes de courbes, le modèle à un facteur est le plus communément utilisé (Cuevas et al., 2004; Zhang and Liang, 2014). En supposant que ces courbes sont des observations de processus gaussiens stochastiques temporels continus mesurés sur une échelle de temps discrète et dans un intervalle fermé $\mathcal{T}$, Ramsay and Silverman (2005) proposent de construire un test de significativité pour une courbe $\beta : \ t \mapsto \beta(t)$ à partir de statistiques individuelles $F_t$ pour les hypothèses nulles $H_{0t} : \ \beta(t) = 0, \ t \in \mathcal{T}$. Zhang and Liang (2014) décrivent les méthodes d'agrégation principales pour tester l'hypothèse nulle $H_0 = \bigcap_{t\in\mathcal{T}} H_{0t}$ à partir des statistiques individuelles $F_t$. La supériorité du test basé sur la norme $\mathrm{L}^\infty$ de $(F_t)_{t\in\mathcal{T}}$ par rapport au test basé sur la norme $\mathrm{L}^2$ est démontrée dans Zhang et al. (2019). Cependant, de telles approches d'agrégation de statistiques de test ignorent la dépendance temporelle entre les statistiques individuelles.

Les Figures 3(b) et (c) représentent des matrices de corrélation estimées sur des données fonctionnelles ; il s'agit de jeux de données publics de spectres proche infrarouges. Ces données sont caractérisées par une forte dépendance temporelle bien

plus complexe que les structures d'autocorrélation ou de Toeplitz couramment supposées dans la littérature des données fonctionnelles. Les structures de dépendance observées ici contiennent également des blocs de forte corrélation positive. Dans la suite, nous cherchons à montrer que prendre en compte la dépendance temporelle dans une procédure d'agrégation peut améliorer sensiblement les performances.

Comme affirmé précédemment, dans cette thèse, l'expression "forte dépendance" fait référence à une situation où la matrice de corrélation peut être approximée de manière satisfaisante en n'utilisant qu'un petit nombre de ses vecteurs propres, les vecteurs propres restants pouvant être considérés comme du bruit. Cela est étroitement lié au choix d'un nombre $K$ d'axes en analyse en composantes principales. Bien qu'il n'y ait pas de consensus définitif sur un critère pour ce choix, conserver les vecteurs propres associés à une valeur propre supérieure à 1 (Kaiser, 1960) est un critère simple et suffisant pour donnée une idée sur la notion de forte dépendance. Pour le gène PDZRN4, 11 valeurs propres sur 68 sont plus grandes que 1. Pour les jeux de données de spectres proche-infrarouges de jus d'orange et de vin, respectivement 3 valeurs propres sur 700 et 6 sur 256 sont plus grandes que 1. Dans chaque situation, le nombre estimé de vecteurs propres "utiles", est beaucoup plus petit que la dimension de la matrice de corrélation.

# 5 Organisation de la thèse

Le manuscrit est organisé comme suit.

Le Chapitre 2 est dédié à l'étude de l'impact de la dépendance sur les approches de test global, en considérant en particulier les approches de test d'association gène - phénotype utilisées en génomique. D'abord, une sélection de méthodes de test global sont comparées et l'impact de la dépendance sur celles-ci est étudié. Il apparaît que le classement des performances des différentes méthodes dépend de la configuration de la structure de dépendance et du signal d'association. Sur la base de cette observation, une méthode adaptative est proposée, nommée test MGF-R (*Moment Generating Function - Ratio based test*). Cette méthode vise à obtenir des performances plus stables d'une configuration à une autre. Ce chapitre a mené au développement d'un package R implémentant cette méthode, nommé `MGFRTest`, et

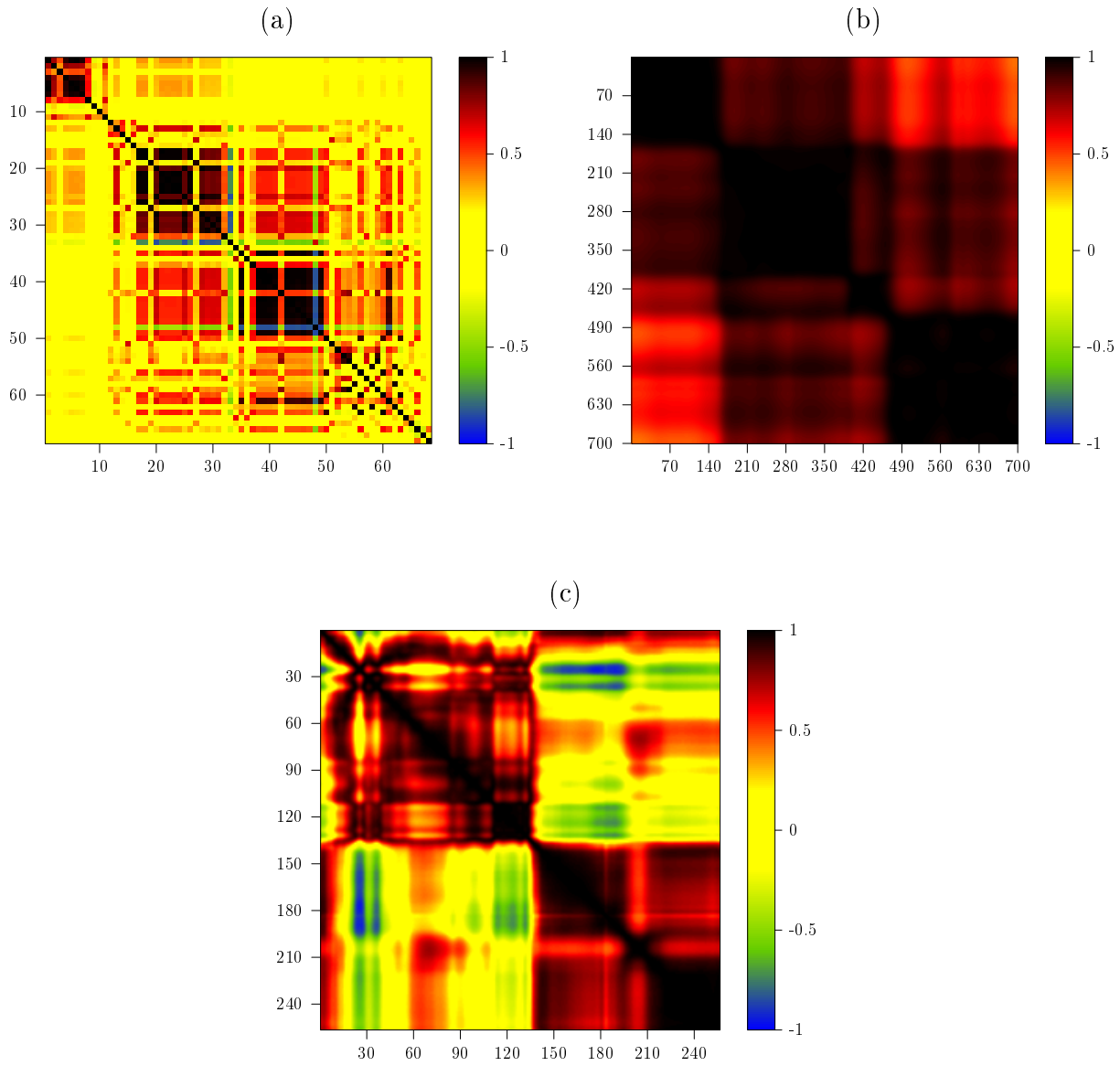


Figure 3 : Exemples de structures de dépendence (matrices de corrélation) estimées sur des jeux de données publics : données de SNP du gène PDZRN4 du génome humain (a) (Wellcome Trust Case Control Consortium, 2007), spectres proche infrarouges de jus d'orange (b) et spectres proche infrarouges de vin (c) (Zhang and Melnik, 2012)

à la soumission d'un article, actuellement en révision.

Dans le Chapitre 3, quelques développements mathématiques sont introduits pour appliquer les méthodes de test global au problème de détection d'effets d'interaction entre un gène et un facteur environnemental. En particulier, un modèle adapté est introduit pour la détection de ces effets d'interaction. Les résultats obtenus dans le Chapitre 2 sont étendus à ce contexte et la méthode adaptative proposée est appliquée à ce problème.

Dans le Chapitre 4, le problème de test global est étendu au problème de détection d'effets d'interaction entre deux gènes. Un modèle adapté à la détection de ces effets est introduit. Cependant, ce problème amène différentes questions. En particulier, la dimension du vecteur de statistiques de test est beaucoup plus grande que précédemment. De plus, il sera mis en évidence que la structure de dépendance du vecteur de statistiques de test est très particulière, étant donné qu'elle est héritée des deux gènes inclus dans le modèle. Les résultats obtenus montrent que même si la matrice de corrélation a une structure très particulière, elle est globalement plus parcimonieuse que les matrices de corrélation des deux gènes considérés, et ses coefficients sont plus faibles. Néanmoins, il est difficile d'évaluer l'influence de cette structure sur les performances des méthodes de test global. En effet, celles-ci sont également affectées par la dimension importante du vecteur de statistiques de test.

Dans le Chapitre 5, l'effet de la dépendance sur les méthodes de classification et de régression est étudié. De façon similaire aux résultats des chapitres précédents, il est mis en évidence que le classement des performances des différentes méthodes dépend non seulement de la structure de dépendance des variables explicatives, mais aussi du signal d'association entre les variables explicatives et la variable réponse. Dans des cadres de grande dimension, des approches dites naïves ignorant la dépendance entre les variables explicatives peuvent donner des performances meilleures que les méthodes prenant en compte la dépendance. Une nouvelle classe de méthodes est introduite, incluant à la fois la règle de prédiction naïve et la règle basée sur une décorrélation complète des variables explicatives. Une stratégie pour déterminer une procédure optimale dans cette classe est proposée. Les résultats obtenus sur des simulations et des jeux de données non simulés montrent que la méthode ainsi proposée donne généralement de très bonnes performances.

Enfin, dans le Chapitre 6, une conclusion générale de cette thèse est donnée et les principaux résultats sont rappelés et discutés. Les résultats obtenus au cours de cette thèse ont été présentés dans des présentations orales et des articles. Une liste complète des travaux (packages R implémentant les méthodes proposées, présentations orales et articles publiés ou soumis) réalisés au cours de la thèse est donnée après le Chapitre 6, avec des annexes détaillant les fonctionnalités des packages R.

# Chapter 1

# Introduction

## Contents


**1 General context . . . . . 28**

**2 Dependence handling for prediction problems . . . . . 29**

2.1 Prediction in low dimension . . . . . 29

2.2 Prediction in high dimension or under strong dependence: handling or ignoring dependence . . . . . 31

**3 Dependence handling for global testing . . . . . 36**

3.1 Global testing in a standard regression model . . . . . 36

3.2 Global testing in high dimension or under strong dependence 37

3.3 Global testing and dependence handling in Genome-Wide Association Studies . . . . . 40

**4 A focus on some specific dependence patterns . . . . . 43**

**5 Organization of the thesis . . . . . 45**

# 1 General context

Predicting a response variable $Y$ from a multivariate profile of explanatory variables $\boldsymbol{X} \in \mathbb{R}^p$, with $p \geq 1$, and testing the significance of the relationship between $\boldsymbol{X}$ and $Y$ are two cornerstones of the statistical methodology that have been markedly renewed by the emergence of high-dimensional or strongly correlated data. Although they do not aim at the same goal, prediction and testing share similarities since they both rely on the estimation of the regression model relating $Y$ and $\boldsymbol{X}$. In practice, a complete data analysis procedure usually associates the evaluation of the prediction performance and tests for model comparison, considering they offer two complementary viewpoints on the relationship between $Y$ and $\boldsymbol{X}$. In genetic epidemiology for example, the profile of explanatory variables is made of genotypes for thousands of markers, the so-called Single Nucleotide Polymorphisms (SNPs), along the genome and the response is usually a status, healthy or not, regarding a disease. Thus, testing for the association between SNPs and the disease gives insight on the genetic determinism of this disease, whereas predicting the disease from genotypes aims at the more ambitious goal of a genome-based prognosis of the disease, for instance in personalized medicine.

The general conclusions of the present thesis are valid for a wide scope of regression models but they will essentially be developed under the assumption of a parametric model: $\mathbb{E}[Y|\boldsymbol{X} = \boldsymbol{x}] = f(\boldsymbol{x}; \boldsymbol{\beta}, \boldsymbol{b})$, where $f$ is a regression function, known up to a vector $\boldsymbol{\beta}$ of association parameters between $\boldsymbol{X}$ and $Y$ and possibly a vector $\boldsymbol{b}$ of additional parameters.

In the following sections, dependence handling for prediction problems and for global testing problems is discussed. It is first shown that optimal theoretical methods for linear discrimination and regression explicitly account for dependence by decorrelating the explanatory variables. In small dimensional problems, the empirical counterparts of these theoretical predictors naturally account for dependence by decorrelation. However, in high dimensional situations, these predictors cannot be used directly. More sophisticated approaches were introduced to counteract this problem. It is shown that several of these predictors can be considered as being based on an adaptive handling of dependence. We will then discuss dependence handling for global testing problems. Similarly as for prediction problems, we first review

optimal tests in a small dimensional context and show that these methods explicitly handle dependence by decorrelating the explanatory variables. Nevertheless, it is interesting to note that when strong dependence is observed among explanatory variables, similarly as in high dimensional situations, these optimal tests cannot be used. However, due to the very diverse points of view and strategies proposed by different authors, whether or not should dependence be taken explicitly into account remains unclear, especially under strong dependence. In the following, strong dependence refers to a situation where the correlation matrix of the explanatory variables can be accurately approximated with a low-rank approximation considering a reduced number of its eigenvectors.

# 2 Dependence handling for prediction problems

## 2.1 Prediction in low dimension

Based on $n$ independent joint observations $\{(\boldsymbol{X_i}, Y_i)\}_{1\leq i\leq n}$ of the explanatory variables and the response, prediction will refer to the estimation of a rule that produces, for an item with profile $\boldsymbol{X} = \boldsymbol{x^*}$ and response value $Y^*$, a predicted value $\hat{Y}^*$ such that an expected loss function $\mathbb{E}[\ell(Y^*, \hat{Y}^*)]$ measuring the average deviation between $\hat{Y}^*$ and $Y^*$ is as small as possible. The way the loss function $\ell(Y^*, \hat{Y}^*)$ is defined depends on the nature of the response. The present manuscript will address the two cases of a real-valued and a categorical response.

Let us first consider the special case of the Linear Discriminant Analysis (LDA) framework, where $Y \in \{0, 1\}$ is a two-group categorical variable and, conditionally on $Y$, $\boldsymbol{X}$ is normally distributed with the same positive within-group variance-covariance matrix $\boldsymbol{\Sigma_x}$. Then, $f(\boldsymbol{x}; \boldsymbol{\beta}, \boldsymbol{b}) = \text{logit}^{-1}\{\log(\pi_1/\pi_0) + (\boldsymbol{x} - \boldsymbol{\mu_x})'\boldsymbol{\Sigma_x^{-1}}\boldsymbol{\delta})\}$, where $\text{logit}^{-1} : u \mapsto 1/\{1+\exp(-u)\}$, $\pi_k = \mathbb{P}(Y = k)$ is the prior group probability, $\boldsymbol{\mu_x}$ is the unconditional expectation of $\boldsymbol{X}$ and $\boldsymbol{\delta}$ is the difference between the group mean expectations of $\boldsymbol{X}$. In the present classification context, the probability of misclassification $\mathbb{E}[\ell(Y^*, \hat{Y}^*)] = \mathbb{P}(\hat{Y}^* \neq Y^*)$ is often considered as the most suited expected loss function. If $\boldsymbol{\Sigma_x}$, $\boldsymbol{\delta}$ and $\boldsymbol{\mu_x}$ are supposed to be known, the probability of misclassification is minimized by the Bayes classifier, consisting in thresholding

the linear score $L(\boldsymbol{X}^*) = \log(\pi_1/\pi_0) + (\boldsymbol{X}^* - \boldsymbol{\mu_x})'\boldsymbol{\Sigma_x}^{-1}\boldsymbol{\delta}$. Therefore, in the present purely probabilistic framework where parameters of the joint distribution of $\boldsymbol{X}$ and $Y$ are supposed to be known, the former closed-form expression of the optimal classification rule does not leave any option for the handling of dependence across explanatory variables measured by $\boldsymbol{\Sigma_x}$.

It is interesting to notice that the optimal linear classification score $L(\boldsymbol{X}^*)$ can be reformulated as $L(\boldsymbol{X}^*) = \log(\pi_1/\pi_0) + \{\boldsymbol{W_x}(\boldsymbol{X}^* - \boldsymbol{\mu_x})\}'\boldsymbol{W_x}\boldsymbol{\delta} = \log(\pi_1/\pi_0) + \tilde{\boldsymbol{X}}^{*\prime}\tilde{\boldsymbol{\delta}}$, where $\boldsymbol{W_x}$ is any $p \times p$ matrix such that $\boldsymbol{W_x'W_x} = \boldsymbol{\Sigma_x}^{-1}$ (or equivalently $\boldsymbol{W_x\Sigma_x W_x'} = \mathbf{I_p}$), $\tilde{\boldsymbol{\delta}} = \boldsymbol{W_x\delta}$ and $\tilde{\boldsymbol{X}}^* = \boldsymbol{W_x}(\boldsymbol{X}^* - \boldsymbol{\mu_x})$. Pre-multiplying $\boldsymbol{X}^* - \boldsymbol{\mu_x}$ by $\boldsymbol{W_x}$ is called a *whitening* transformation since the within-group variance of the whitened vector $\tilde{\boldsymbol{X}}^*$ is $\boldsymbol{W_x\Sigma_x W_x'} = \mathbf{I_p}$ (Kessy et al., 2018). The optimal classification score $L(\boldsymbol{X}^*)$ being a linear combination of the coordinates of the whitened vector of explanatory variables, a general conclusion could be that the best way to handle dependence is by decorrelating the explanatory variables.

Let us now consider the linear regression framework, where $Y$ is a real-valued variable. For simplicity, let us assume the explanatory variables and the response are scaled to zero mean and unit variance. Let us further assume that:

$$\begin{pmatrix} \boldsymbol{X} \\ Y \end{pmatrix} \sim \mathcal{N}_{p+1}\left(\begin{pmatrix} \mathbf{0} \\ 0 \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma_x} & \boldsymbol{\sigma_{xy}} \\ \boldsymbol{\sigma_{xy}'} & 1 \end{pmatrix}\right)$$

where $\boldsymbol{\Sigma_x}$ is the correlation matrix of $\boldsymbol{X}$ and $\boldsymbol{\sigma_{xy}}$ is the $p$-vector whose $i$-th coordinate is the correlation between $Y$ and the $i$-th explanatory variable $X_i$. The mean squared error of prediction $\mathbb{E}[\ell(Y^*, \hat{Y}^*)] = \mathbb{E}[(Y^* - \hat{Y}^*)^2]$ is often chosen as the expected loss function in this context. Assuming the parameters are known, the best linear unbiased predictor is $L(\boldsymbol{X}^*) = \boldsymbol{X}^{*\prime}\boldsymbol{\Sigma_x}^{-1}\boldsymbol{\sigma_{xy}} = (\boldsymbol{W_x X}^*)'(\boldsymbol{W_x\sigma_{xy}})$ where $\boldsymbol{W_x'W_x} = \boldsymbol{\Sigma_x}^{-1}$. Consequently, the best linear predictor involves a decorrelation of the explanatory variables, as in the LDA framework.

For both the LDA and the regression contexts, the optimal predictors require the use of unknown parameters in real problems. The most straightforward and natural idea is to replace them with their sample moment estimates. Under some conditions, namely that the estimated covariance matrix of the explanatory variables is not ill-conditioned, the corresponding predictors (Fisher's rule in LDA and the ordinary least squares (OLS) predictor in regression) are known to perform generally

well. Finally, it turns out that for LDA as for regression, prediction is enhanced by decorrelating the explanatory variables.

## 2.2 Prediction in high dimension or under strong dependence: handling or ignoring dependence

Unfortunately, in a high-dimensional context, where $n < p$, and for an arbitrary dependence pattern, uniformly optimal procedures do not exist. The variety of procedures is huge, with diverse viewpoints about the handling of dependence across the explanatory variables, and no clearly indisputable solutions. Interestingly, the main issue encountered in an high dimensional context, or in a large dimensional context (where $n > p$ but $n$ is not much larger than $p$), is the same as that encountered in strong dependence situations: the correlation (or covariance) matrix cannot be inverted, or its inverse has a very important variance, thus impacting the stability of the procedures (Hoerl and Kennard, 1970). Numerous methods were proposed to counteract this problem. Several of them are reviewed hereafter, which can be used (in the sense that the predictor can be computed) under high dimension or strong dependence.

Statistical practice of regression modeling inherits from the proven procedures designed for regular settings where $n > p$ the belief that, if the explanatory variables are dependent, then this dependence should not be ignored in the derivation of prediction rules or testing strategies. Although this belief is sometimes confirmed in high-dimensional designs, it is also questioned by many authors. In particular, "naive" predictors ignoring dependence sometimes have the ability to outperform more sophisticated methods. We detail this point by summarizing arguments in favor of ignoring dependence or not.

### 2.2.1 High dimensional LDA

Let us come back to the two-group LDA framework introduced in the previous section. Fisher's prediction rule (Fisher, 1936) is deduced from the linear Bayes classification score by plugging-in sample estimates $\boldsymbol{S}_{\boldsymbol{x}}$, $\hat{\boldsymbol{\delta}}$ and $\hat{\boldsymbol{\mu}}_{\boldsymbol{x}}$ of $\boldsymbol{\Sigma}_{\boldsymbol{x}}$, $\boldsymbol{\delta}$ and $\boldsymbol{\mu}_{\boldsymbol{x}}$

respectively:

$$L_{\text{LDA}}(\boldsymbol{X}^*) = \log(\hat{\pi}_1/\hat{\pi}_0) + (\boldsymbol{X}^* - \hat{\boldsymbol{\mu}}_{\boldsymbol{x}})' \boldsymbol{S}_{\boldsymbol{x}}^{-1} \hat{\boldsymbol{\delta}},$$

where $\hat{\pi}_k = \#\{i = 1, \ldots, n,\ Y_i = k\}/n$. As the empirical counterpart to the Bayes rule, Fisher's rule is asymptotically optimal for minimizing the misclassification rate.

However, in a comparative study of machine learning methods applied on a selection of public high-dimensional gene expression datasets, Dudoit et al. (2002) introduced a modified version of Fisher's rule by replacing $\boldsymbol{S}_{\boldsymbol{x}}$ by $\boldsymbol{D}_{\boldsymbol{s}} = \text{diag}(\boldsymbol{S}_{\boldsymbol{x}})$, thus ignoring the correlation between any pair of predicting variables. Even though the former independence assumption is obviously wrong, the resulting so-called Diagonal Discriminant Analysis (DDA) rule turns out to outperform Fisher's rule in some studies. This is confirmed by Bickel and Levina (2004) who establish the conditions under which the non-asymptotic classification performance of DDA is better than LDA in large or high dimensional situations (see also Efron (2009); Tibshirani et al. (2003); Bodnar and Okhrin (2011)).

Nevertheless, even in high dimensional problems, LDA is not always outperformed by DDA. We illustrate this point by using two high-dimensional gene expression datasets, thus belonging to the same field of application and generated by similar technology. The datasets are available in the `R` package `plsgenomics` (Boulesteix et al., 2018). The first one, `Leukemia`, is also included in the comparative study of Dudoit et al. (2002) and contains 3,051 gene expressions measured for 38 individuals. The second one, `Colon`, contains 2,000 gene expressions for 62 individuals. In both cases, the response variable is a status, healthy or not, of a patient regarding a disease. Since the number of explanatory variables is much larger than the sample size, the inverse matrix of $\boldsymbol{S}_{\boldsymbol{x}}$ in the LDA classification score is replaced by the Moore-Penrose generalized inverse (see Bickel and Levina (2004)).

To compare the performance of LDA and DDA, the misclassification rate of both methods is estimated using 10-fold cross-validation. The cross-validation procedure is repeated with 50 random splittings of the dataset in 10 segments. The boxplots of the 50 estimated misclassification rates are given in Figure 1.1. In agreement with the results of Dudoit et al. (2002), DDA outperforms LDA for the `Leukemia` dataset. The average misclassification rate obtained with the DDA rule is approximately 2.5%, against approximately 9% with the LDA rule. However, on the `Colon` dataset,

LDA outperforms DDA; the average misclassification rates are approximately 22% and 38%, respectively. Consequently, even though the two datasets seem similar regarding their nature and field of application, choosing the DDA rule over the LDA rule (or the other way around) would yield suboptimal performances for one dataset.

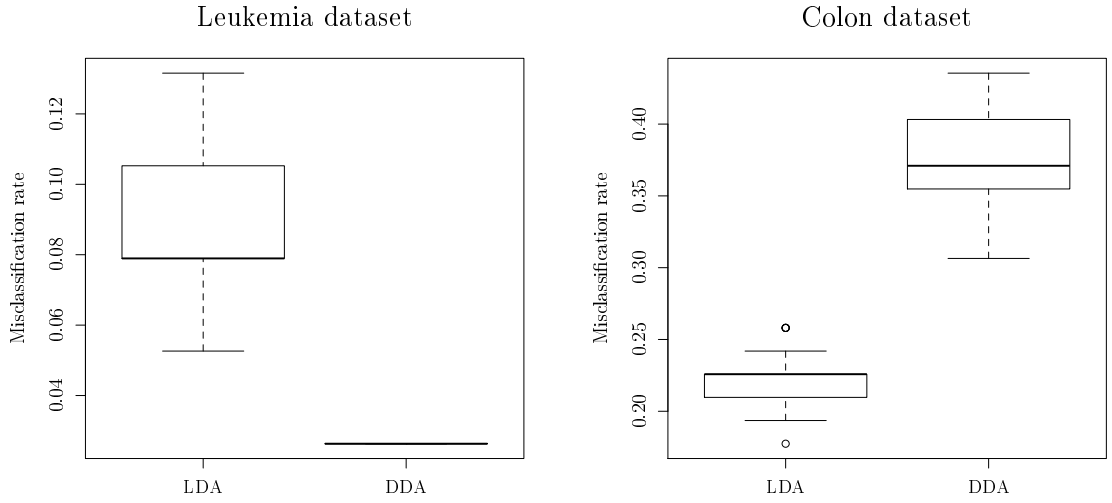


Figure 1.1: Boxplots of the estimated misclassification rates for the LDA and DDA rules on the Leukemia and Colon datasets, using 50 repeated 10-fold cross validation procedures

### 2.2.2 High dimensional regression

In the regression context, many variants of the Ordinary Least Squares (OLS), or, more generally, of the Maximum-Likelihood (ML) estimation procedure have been proposed to circumvent the problems raised by high dimension. Among the most popular of these methods, some opt for a shrunken estimation of the variance-covariance matrix of the explanatory variables, either by regularization of the least-squares minimization algorithm in Ridge regression (Hoerl and Kennard, 1970) or by a reduced-rank modeling in Principal Component Regression (PCR, Jolliffe (1982)) and Partial Least-Squares Regression (PLSR, Wold et al. (1983, 1984)).

Assuming without loss of generality that the explanatory variables and the response were scaled to zero mean and unit variance, the OLS predictor can be written

as $L_{\text{OLS}}(\boldsymbol{X}^*) = \boldsymbol{X}^{*\prime}\boldsymbol{S}_{\boldsymbol{x}}^{-}\boldsymbol{s}_{\boldsymbol{xy}}$ where $\boldsymbol{S}_{\boldsymbol{x}}$ and $\boldsymbol{s}_{\boldsymbol{xy}}$ are the sample estimates of $\boldsymbol{\Sigma}_{\boldsymbol{x}}$ and $\boldsymbol{\sigma}_{\boldsymbol{xy}}$, respectively, and $\boldsymbol{S}_{\boldsymbol{x}}^{-}$ is the generalized inverse of $\boldsymbol{S}_{\boldsymbol{x}}$. One of the first methods designed to address the singularity of $\boldsymbol{S}_{\boldsymbol{x}}$ is the Ridge regression (Hoerl and Kennard, 1970). The Ridge predictor can be written as:

$$L_{\text{Ridge}}(\boldsymbol{X}^*) = \boldsymbol{X}^{*\prime}(\boldsymbol{S}_{\boldsymbol{x}} + \kappa\mathbf{I}_{\boldsymbol{p}})^{-}\boldsymbol{s}_{\boldsymbol{xy}}$$

where $\kappa$ is the Ridge penalty, usually chosen by optimizing a criterion through a cross-validation procedure. It can be seen that the Ridge predictor can be considered as a trade-off between taking dependence into account, as for the OLS predictor, or on the contrary ignoring dependence. Depending on the value of the Ridge regularization parameter $\kappa$, the predictor is more or less close to the OLS predictor. Indeed, let us introduce the eigenvalue decomposition of $\boldsymbol{S}_{\boldsymbol{x}}$ as $\boldsymbol{S}_{\boldsymbol{x}} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}'$, $\boldsymbol{\Lambda} = \text{diag}(\lambda_i)$. Then

$$L_{\text{Ridge}}(\boldsymbol{X}^*) = \boldsymbol{X}^{*\prime}\boldsymbol{U}\,\text{diag}\left(\frac{1}{\lambda_i + \kappa}\right)\boldsymbol{U}'\boldsymbol{s}_{\boldsymbol{xy}}.$$

For $\kappa = 0$, obviously $L_{\text{Ridge}}(\boldsymbol{X}^*) = L_{\text{OLS}}(\boldsymbol{X}^*)$. On the other hand, if $\kappa \gg \lambda_1$ where $\lambda_1$ is the greatest eigenvalue, then $\lambda_i + \kappa \simeq \kappa$ and $L_{\text{Ridge}}(\boldsymbol{X}^*) \propto \boldsymbol{X}^{*\prime}\boldsymbol{s}_{\boldsymbol{xy}}$; in this case, the Ridge predictor ignores dependence between explanatory variables.

Interestingly, Opgen-Rhein and Strimmer (2007b) (see also Zuber and Strimmer (2011)) also proposed to address the ill-conditioning of $\boldsymbol{S}_{\boldsymbol{x}}$ by using shrinkage-based estimates of the variances and correlations involved in the OLS predictor (Schäfer and Strimmer, 2005). Let us denote $\boldsymbol{D}$ the diagonal matrix containing the estimated variances of the explanatory variables, $\boldsymbol{R}_{\boldsymbol{x}} = \boldsymbol{D}^{-1/2}\boldsymbol{S}_{\boldsymbol{x}}\boldsymbol{D}^{-1/2}$ the estimated correlation matrix of the explanatory variables and $\boldsymbol{r}_{\boldsymbol{xy}} = \boldsymbol{D}^{-1/2}\boldsymbol{s}_{\boldsymbol{xy}}/s_y$ the vector of estimated correlations between the response and each explanatory variable. Then, by replacing the usual esimates by shruken estimates, Opgen-Rhein and Strimmer (2007b) introduced the following predictor:

$$L_{\text{SLM}}(\boldsymbol{X}^*) = \boldsymbol{X}^{*\prime}(s_y^{(\tau)}(\boldsymbol{D}^{(\boldsymbol{\tau})})^{-1/2}(\boldsymbol{R}_{\boldsymbol{x}}^{(\boldsymbol{\gamma})})^{-1}\boldsymbol{r}_{\boldsymbol{xy}}^{(\boldsymbol{\gamma})})$$

where $\gamma$ and $\tau$ are shrinkage parameters. In particular, $\boldsymbol{R}_{\boldsymbol{x}}^{(\boldsymbol{\gamma})} = \gamma\mathbf{I}_{\mathbf{p}} + (1-\gamma)\boldsymbol{R}_{\boldsymbol{x}}$ where $\gamma \in [0,1]$. It can thus be remarked that for great values of $\gamma$, $\boldsymbol{R}_{\boldsymbol{x}}^{(\boldsymbol{\gamma})}$ tends to the identity matrix and dependence among explanatory variables tends to be ignored. Nevertheless, $\boldsymbol{r}_{\boldsymbol{xy}}^{(\boldsymbol{\gamma})} = (1-\gamma)\boldsymbol{r}_{\boldsymbol{xy}}$; consequently, dependence between the explanatory variables and the response also tends to be ignored. The vector of

estimated coefficients thus tends to $\mathbf{0}$. Finally, the response is simply predicted by its mean. On the contrary, if $\gamma = 0$, the above predictor is similar to the OLS predictor; the only difference resides in the estimation of variances. Depending on the value of the shrinkage parameter $\gamma$, $\boldsymbol{R}_{\boldsymbol{x}}^{(\gamma)}$ is more or less close to $\boldsymbol{R}_{\boldsymbol{x}}$. In particular, $\boldsymbol{R}_{\boldsymbol{x}}^{(\gamma)} = \boldsymbol{W}(\gamma \mathbf{I}_{\boldsymbol{p}} + (1-\gamma)\boldsymbol{\Psi})\boldsymbol{W}'$ where $\boldsymbol{W}\boldsymbol{\Psi}\boldsymbol{W}'$ is the eigenvalue decomposition of $\boldsymbol{R}_{\boldsymbol{x}}$, meaning that the eigenvalues of $\boldsymbol{R}_{\boldsymbol{x}}$ which are greater than 1 are shrunken. The same idea was also transposed to discriminant analysis in Ahdesmäki and Strimmer (2010); Zuber and Strimmer (2009), where the estimates involved in the expression of Fisher's score are replaced with shrunken-based estimates.

The PCR predictor (Jolliffe, 1982) is defined as follows:

$$L_{\text{PCR}}(\boldsymbol{X}^*) = \boldsymbol{X}^{*\prime} \left( \sum_{i=1}^{k} \frac{1}{\lambda_i} \boldsymbol{u_i} \boldsymbol{u_i'} \right) \boldsymbol{s_{xy}}$$

where $\boldsymbol{u_i}$ is the $i$-th eigenvector of $\boldsymbol{S_x}$, associated to the eigenvalue $\lambda_i$. The term $\sum_{i=1}^{k} \frac{1}{\lambda_i} \boldsymbol{u_i} \boldsymbol{u_i'}$, which can be seen as a kernel of latent predictive variables, is a rank-reduced approximation to $\boldsymbol{S_x^-}$. Thus, PCR allows for an intermediate handling of dependence through the dimension $k$ of the kernel. Indeed, increasing $k$ amounts to use more eigenvectors to approximate the inverse of $\boldsymbol{S_x}$. Consequently, as $k$ increases, $\sum_{i=1}^{k} \frac{1}{\lambda_i} \boldsymbol{u_i} \boldsymbol{u_i'}$ gets closer to the inverse of $\boldsymbol{S_x}$.

Finally, the PLS regression predictor (Wold et al., 1983, 1984), often preferred to the former PCR predictor, can be written as (Helland, 1988; Lingjaerde and Christophersen, 2000; Blazère et al., 2014):

$$L_{\text{PLS}}(\boldsymbol{X}^*) = \boldsymbol{X}^{*\prime} \left( \sum_{i=1}^{k} \alpha_i \boldsymbol{S_x^{i-1}} \right) \boldsymbol{s_{xy}},$$

where $k$ is the number of latent PLS factors. In particular, if $k = 1$, then $L_{\text{PLS}}(\boldsymbol{X}^*) \propto \boldsymbol{X}^{*\prime} \boldsymbol{s_{xy}}$ and dependence among explanatory variables is ignored. Moreover, the term $\sum_{i=1}^{k} \alpha_i \boldsymbol{S_x^{i-1}}$ is a polynomial approximation to $\boldsymbol{S_x^{-1}}$; the greater is $k$, the closer the term gets to $\boldsymbol{S_x^{-1}}$. Finally, the amount of dependence taken into account increases with $k$.

For each of these four approaches, an hyperparameter is introduced, either a regularization term in Ridge regression or in the shrinkage-based regression of Opgen-Rhein and Strimmer (2007b), or the dimension of the kernel of latent predicting

variables in PCR or PLSR. In each case, it can be seen that this hyperparameter defines an intermediate handling of dependence between ignorance and complete whitening. For shrinkage-based regression methods, dependence tends to be ignored for high values of the shrinkage parameter, and to be taken into account for low values of the shrinkage parameter. On the other hand, the amount of dependence taken into account is controlled by increasing or decreasing the number $k$ of latent variables included in the PCR and PLSR predictors. These points show that several approaches are suitable for an adaptive handling of dependence in a regression framework. Moreover, some of these approaches were transposed to the former discriminant analysis framework, namely as PLS discriminant analysis (Boulesteix, 2004; Gottfries et al., 1995; Barker and Rayens, 2003) or shrinkage-based discriminant analysis (Friedman, 1989; Guo et al., 2005), thus showing that adaptively handling dependence can be a suitable approach in a broad range of frameworks.

# 3 Dependence handling for global testing

In the same general framework of a parametric model, global testing (Arias-Castro et al., 2011) refers to the test of $H_0 : \boldsymbol{\beta} = \boldsymbol{\beta_0}$, where $\boldsymbol{\beta_0}$ is a vector of target values. For instance, in the special case of the linear regression model where, conditionally on $\boldsymbol{X} = \boldsymbol{x}$, $Y$ is normally distributed with mean $f(\boldsymbol{x}; \boldsymbol{\beta}, b_0) = b_0 + \boldsymbol{x}'\boldsymbol{\beta}$, one of the most usual examples of such a global testing issue is the test for the significance of a relationship between $Y$ and $\boldsymbol{X}$, expressed as the test of $H_0 : \boldsymbol{\beta} = \mathbf{0}$. More generally, global testing can refer to any problem of testing for a null hypothesis $H_0$ that can be written as a collection of pointwise null hypotheses $H_0^{(i)}$, $i = 1, \ldots, p$, *i.e.* $H_0 = \bigcap_{1 \leq i \leq p} H_0^{(i)}$.

## 3.1 Global testing in a standard regression model

Still in the linear regression model settings, where it is moreover supposed that $n > p$, the traditional Analysis of Variance (ANOVA, Fisher (1925)) testing approach is the most common way to address the above global testing issue. Indeed, the uniformly most powerful test statistic for $H_0$, in the usual sense of the Neyman-

Pearson lemma, is the $F$-test, that can be expressed as a Wald-type statistic (Wald, 1943): $F = \hat{\boldsymbol{\beta}}' \hat{\boldsymbol{V}}_{\hat{\boldsymbol{\beta}}}^{-1} \hat{\boldsymbol{\beta}}/p$, where $\hat{\boldsymbol{\beta}} = \boldsymbol{S}_{\boldsymbol{x}}^{-1} \boldsymbol{s}_{\boldsymbol{xy}}$ is the Ordinary Least Squares (OLS) estimate of $\boldsymbol{\beta}$, $\hat{\boldsymbol{V}}_{\hat{\boldsymbol{\beta}}} = \frac{\hat{\sigma}^2}{n} \boldsymbol{S}_{\boldsymbol{x}}^{-1}$ estimates the variance of $\hat{\boldsymbol{\beta}}$, $\boldsymbol{S}_{\boldsymbol{x}}$ and $\boldsymbol{s}_{\boldsymbol{xy}}$ are respectively the sample variance-covariance matrix of the explanatory variables and the sample vector of covariances between the response and the explanatory variables, and $\hat{\sigma}^2$ is the degree-of-freedom corrected mean of squared residuals. In this regular regression design, where it is supposed that $\boldsymbol{S}_{\boldsymbol{x}}^{-1}$ exists, dependence across the components of $\hat{\boldsymbol{\beta}}$ is readily inherited from the dependence across the explanatory variables since $\hat{\boldsymbol{V}}_{\hat{\boldsymbol{\beta}}} \propto \boldsymbol{S}_{\boldsymbol{x}}^{-1}$.

Here, either based on likelihood theory or on geometrical arguments, both estimation and testing find an optimal solution, the former leading to the OLS estimation procedure and the latter to the $F$–test. This undoubtedly imposes those approaches as gold standards. Interestingly, the $F$–test statistic can also be viewed as the mean of squared coordinates of a whitened version of the vector of estimated regression parameters. Indeed, let $\tilde{\boldsymbol{\beta}} = \boldsymbol{W}_{\hat{\boldsymbol{\beta}}} \hat{\boldsymbol{\beta}}$, where $\boldsymbol{W}_{\hat{\boldsymbol{\beta}}}$ is any matrix such that $\boldsymbol{W}_{\hat{\boldsymbol{\beta}}}' \boldsymbol{W}_{\hat{\boldsymbol{\beta}}} = \hat{\boldsymbol{V}}_{\hat{\boldsymbol{\beta}}}^{-1}$. Then, the asymptotic variance of $\tilde{\boldsymbol{\beta}}$ is the $p \times p$ identity matrix $\mathbf{I}_{\boldsymbol{p}}$ and $F$ can be expressed as the squared $\mathrm{L}^2$-norm of $\tilde{\boldsymbol{\beta}}$ divided by $p$: $F = ||\tilde{\boldsymbol{\beta}}||_2^2/p$. Here also, this readily leads to the conclusion that the optimal way of handling dependence in global testing is by decorrelating $\hat{\boldsymbol{\beta}}$. Interestingly, as $\hat{\boldsymbol{\beta}} = \boldsymbol{S}_{\boldsymbol{x}}^{-1} \boldsymbol{s}_{\boldsymbol{xy}}$, $F \propto \boldsymbol{s}_{\boldsymbol{xy}}' \boldsymbol{S}_{\boldsymbol{x}}^{-1} \boldsymbol{s}_{\boldsymbol{xy}} = ||\boldsymbol{W}_{\boldsymbol{x}} \boldsymbol{s}_{\boldsymbol{xy}}||_2^2$ where $\boldsymbol{W}_{\boldsymbol{x}}' \boldsymbol{W}_{\boldsymbol{x}} = \boldsymbol{S}_{\boldsymbol{x}}^{-1}$. Furthermore, $\mathrm{Var}(\boldsymbol{s}_{\boldsymbol{xy}}) \propto \boldsymbol{S}_{\boldsymbol{x}}$, meaning that the optimal $F$ statistic can also be obtained by whitening the univariate OLS slope estimators stacked in the vector $\boldsymbol{s}_{\boldsymbol{xy}}$. The dependence structure of the OLS estimator $\hat{\boldsymbol{\beta}}$ is completely different from that of the vector of univariate OLS estimators $\boldsymbol{s}_{\boldsymbol{xy}}$. However, both vectors, after whitening, lead to the same optimal test statistic.

## 3.2 Global testing in high dimension or under strong dependence

In a high dimensional model or under strong dependence among explanatory variables, the OLS estimator $\hat{\boldsymbol{\beta}}$ cannot be computed, or is unstable due to its high variance. Under such conditions, for the significance testing of the parameter vector

$\boldsymbol{\beta}$, many authors (see Arias-Castro et al. (2011) for a review) reformulate the null hypothesis as $H_0 : \boldsymbol{\mu} = \mathbf{0}$, where $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_p)'$ is the $p$–vector of slope parameters in the univariate regression models relating $Y$ to each of the explanatory variables: $\mathbb{E}[Y|X_j = x_j] = m_{0j} + \mu_j x_j$, $m_{0j}$ being an intercept parameter. Indeed, estimation of $\boldsymbol{\mu}$ is not affected by the dimension of the parametric space regarding the sample size, nor by the dependence among explanatory variables. However, as for $\hat{\boldsymbol{\beta}}$, the coordinates of the estimate $\hat{\boldsymbol{\mu}}$ of $\boldsymbol{\mu}$ inherit their dependence from the dependence across explanatory variables, which carries the question of the best way to handle dependence in the derivation of a global test statistic over $\hat{\boldsymbol{\mu}}$.

Interestingly, the estimate of $\boldsymbol{\mu}$ can be expressed as $\hat{\boldsymbol{\mu}} = \bar{\boldsymbol{Z}}$, where $\bar{\boldsymbol{Z}}$ is the average over $i = 1, \ldots, n$ of the variables $\boldsymbol{Z_i} = Y_i . \boldsymbol{D_s}^{-1}(\boldsymbol{X_i} - \bar{\boldsymbol{X}})$ and $\boldsymbol{D_s}$ is the $p \times p$ diagonal matrix with the same diagonal entries as $\boldsymbol{S_x}$. Moreover, for all $i = 1, \ldots, n$, $\boldsymbol{Z_i}$ is asymptotically normally distributed with mean $\boldsymbol{\mu}$ and variance $\boldsymbol{V_Z}$. Therefore, testing the significance of $\boldsymbol{\mu}$ amounts to testing the significance of the expectation of a normal vector $\boldsymbol{Z}$ based on $n$ observations $\boldsymbol{Z_i}$, $i = 1, \ldots, n$. A well-known solution to this problem is Hotelling's t-square test (Hotelling, 1931), based on the statistic $T^2 = n\bar{\boldsymbol{Z}}' \boldsymbol{S_Z}^{-1} \bar{\boldsymbol{Z}}$, where $\boldsymbol{S_Z}$ is the sample variance-covariance matrix of $\boldsymbol{Z}$. In regular settings where $n > p$, this test is uniformly most powerful within the class of test statistics invariant by a linear transformation on $\boldsymbol{Z}$ (see Anderson (2003), Chapter 5, section 5.6). However, implementing this solution supposes that $\boldsymbol{S_Z}^{-1}$ exists. In high dimension or under strong dependence, $\boldsymbol{S_Z}^{-1}$ does not exist or is too unstable to produce reliable results.

The former general approach of global testing based on $\boldsymbol{\mu}$ rather than $\boldsymbol{\beta}$ is therefore addressed as the search for the most powerful test statistic by aggregating the coordinates of $\hat{\boldsymbol{\mu}}$. In Genome-Wide Association Studies (GWAS), where single marker tests are sometimes aggregated within spatially contiguous blocks to identify regions of the genome that may be involved in a disease, many papers report comparative studies of aggregation methods, most of them ignoring dependence across the coordinates of $\hat{\boldsymbol{\mu}}$, including $\mathrm{L}^2$– or $\mathrm{L}^\infty$–norm based tests (see Wu et al. (2014) for a review). Conditions on which an aggregation procedure is optimal in large or high-dimensional contexts are discussed by Donoho and Jin (2004); Hall and Jin (2008); Arias-Castro et al. (2011). Donoho and Jin (2004) address the global testing issue by aggregating pointwise test statistics $Z_j$, $j = 1, \ldots, p$, under the assump-

tion that $\boldsymbol{\mathcal{Z}} = (Z_1, \ldots, Z_p)'$ is distributed according to the following sparse normal mixture model: $\forall j,\ Z_j \sim (1-\varepsilon)\mathcal{N}(0,1) + \varepsilon\mathcal{N}(\delta,1)$, where the mixing parameter $0 \leq \varepsilon \leq 1$ is the proportion of non-null features, $\delta \geq 0$ is the signal amplitude and moreover, $\text{Var}(\boldsymbol{\mathcal{Z}}) = \boldsymbol{\Sigma}$, where $\boldsymbol{\Sigma}$ is a positive-definite correlation matrix. Donoho and Jin (2004) also introduce upper bounds on $\varepsilon$ and $\delta$ under which the signal can be considered as both rare and weak.

Donoho and Jin (2004) proposed the so-called Higher Criticism (HC) statistic, defined as a Kolmogorov-Smirnov-type distance between the empirical distribution of the p-values $p_j$ associated to the pointwise tests $Z_j$, and the uniform distribution:

$$\text{HC} = \max_{1 \leq j \leq p/2} \sqrt{p} \frac{j/p - p_{(j)}}{\sqrt{p_{(j)}(1-p_{(j)})}}$$

where $p_{(j)}$ is the $j$-th ordered p-value. Assuming $\boldsymbol{\Sigma} = \mathbf{I}_{\boldsymbol{p}}$, Donoho and Jin (2004) show that HC is Chernoff-consistent and optimal, in the sense that it achieves the best detectability bounds in the rare-and-weak signal regime (Ingster, 1997; Donoho and Jin, 2004, 2008). When the independence assumption is not fulfilled, HC is no longer optimal and its properties can be seriously affected (Hall and Jin, 2008). To restore its optimality, Hall and Jin (2010) proposed to accurately take into account the correlation between the test statistics by a preliminary whitening of $\boldsymbol{\mathcal{Z}}$. Optimality of this decorrelation-based HC statistic, called innovated Higher Criticism, was established by Hall and Jin (2010). In particular, observing that whitening has the effect of amplifying the non-null expectation of $\boldsymbol{\mathcal{Z}}$, Hall and Jin (2010) claim that correlation strengthens the ability to detect the signal. However, Arias-Castro et al. (2011) extended the asymptotic (with respect to $p$) optimality of the standard HC under weakly correlated designs. The major difference between Hall and Jin (2008, 2010) and Arias-Castro et al. (2011) is the following: in Hall and Jin (2008, 2010), $\boldsymbol{\mu}$ is assumed to fulfill the assumptions of the rare and weak regime, whereas the vector of regression parameters $\boldsymbol{\beta}$ is supposed to fulfill these assumptions in Arias-Castro et al. (2011). Even though this makes no difference if the explanatory variables are independent, $\boldsymbol{\beta}$ and $\boldsymbol{\mu}$ can have very different patterns under dependence. In particular, having a sparse vector $\boldsymbol{\beta}$ does not necessary yield a sparse vector $\boldsymbol{\mu}$.

## 3.3 Global testing and dependence handling in Genome-Wide Association Studies

Global testing methods are commonly used in GWAS (Conneely and Boehnke, 2007; Wu et al., 2014; Luo et al., 2010; Barnett et al., 2017; Basu and Pan, 2011), where they are often referred to as tests or $p$-values combination methods. As mentioned above, these approaches aim at discovering associations between regions of the genome and a phenotype (*e.g.* a disease) of interest by aggregating pointwise test statistics for the association between each marker within the region and the response variable. Within-region dependence across single markers and consequently among the corresponding pointwise test statistics can be strong and describe various patterns. For this region-based testing issue, following Arias-Castro et al. (2011), Wu et al. (2014) proposed to use the initial HC statistic defined by Donoho and Jin (2004), which derivation ignores dependence, and showed its optimality under several assumptions on the correlation matrix, still in the rare-and-weak settings introduced by Donoho and Jin (2004). More strikingly, they even show that a method introduced by Luo et al. (2010) and based on a decorrelation step similar to that of Hall and Jin (2010) is not optimal. Also skeptical about the gain in whitening the pointwise test statistics, Barnett et al. (2017) argued that such a preliminary step of decorrelation could deteriorate the signal and amplify noise, thus reducing the power of a global testing procedure. Nevertheless, Liu and Xie (2018) showed that if the conditional effects of a linear model are larger than its marginal effects, then a detection procedure based on the marginal effects such as those proposed by the previous authors may be powerless. Therefore, they proposed to pre-multiply the vector of marginal effects by the inverse of its covariance matrix, thus considering a transformation similar to a decorrelation transformation. Indeed, decorrelation consists in pre-multiplying by the inverse of a square root of the covariance matrix, whereas Liu and Xie (2018) simply consider the inverse of the covariance matrix. They show both theoretically and through simulations that a test based on the transformed vector can outperform tests based on marginal effects. This obviously advocates for a proper and explicit dependence handling, similar to Hall and Jin (2010).

The above discussion illustrates the difficulty to draw a consensus on whether

dependence should be taken into account or not. It turns out that, depending on the situation, it can either be detrimental or beneficial.

Let us illustrate this point with data-driven simulations. In the following, two possible mean vectors $\boldsymbol{\mu_1}$ and $\boldsymbol{\mu_2}$ (represented on Figure 1.2) are considered for a vector $\boldsymbol{\mathcal{Z}}$ of test statistics. Then, 10,000 random generations of $\boldsymbol{\mathcal{Z}}$ are drawn according to the multivariate normal distribution with mean $\boldsymbol{\mu_1}$ and covariance matrix estimated using public GWAS data focusing on the PDZRN4 gene (see Figure 1.3(a)). Similarly, 10,000 simulations of $\boldsymbol{\mathcal{Z}}$ are generated with mean $\boldsymbol{\mu_2}$ and the same covariance matrix.

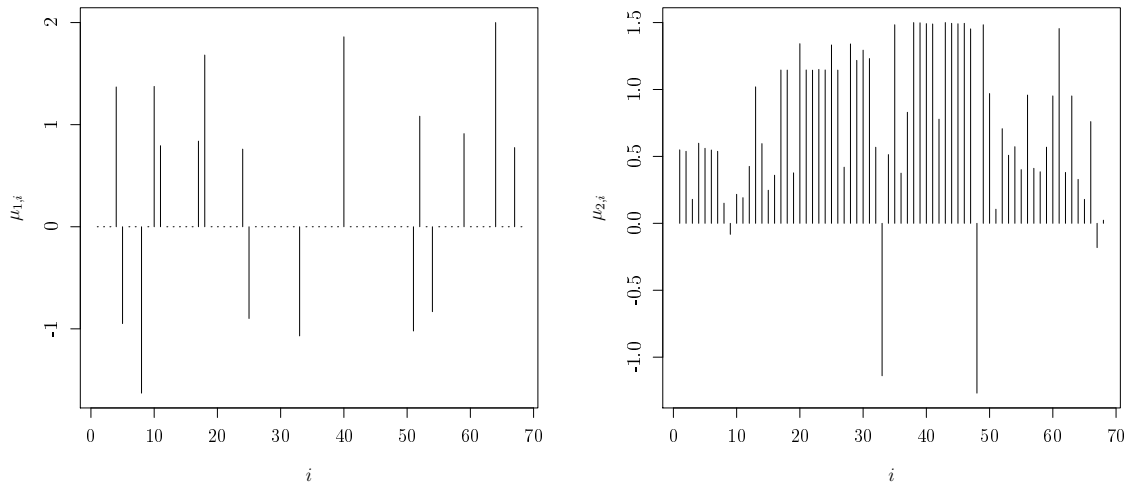


Figure 1.2: Coordinates of the mean vectors $\boldsymbol{\mu_1}$ and $\boldsymbol{\mu_2}$

We propose to compare the performance of two simple aggregation methods under those two configurations. The first method is based on the $L^2$-norm of the vector $\boldsymbol{\mathcal{Z}}$, and the second one on the $L^2$-norm of $\boldsymbol{\mathcal{Z}}$ after a preliminary decorrelation step. Decorrelation is obtained by pre-multiplying $\boldsymbol{\mathcal{Z}}$ by an inverse-square root of the true covariance matrix of the pointwise test statistics and exact $p$-values are deduced from the known null distribution of $L^2$-norm based statistics in the present normal framework. Detection rates, namely proportions of the 10,000 simulations for which the $p$-value is smaller than 0.05, are given in Table 1.1. In the first situation, the decorrelation-based test is much more powerful than the other test, which has no power at all, whereas on the contrary, in the second situation, the test ignoring

dependence is the most powerful. Even if these results were obtained using very simple examples, they demonstrate that choosing to decorrelate the test statistics or not is not just a matter of dependence pattern of the pointwise test statistics but of an interplay between this dependence pattern and the pattern of association between the pointwise features and the response variable.

Table 1.1: Estimated power of two global testing methods under two different configurations

| Mean vector | Decorrelation | No decorrelation |
|---|---|---|
| $\boldsymbol{\mu_1}$ | 0.72 | 0.05 |
| $\boldsymbol{\mu_2}$ | 0.09 | 0.33 |

To sum up, both for prediction or for global testing, the impact of dependence across explanatory variables is not identified as a specific issue in regular settings where $n > p$, since the way dependence is handled is explicit in optimal procedures under arbitrarily complex dependence structures. For most of them, it turns out that they can be viewed as applications of optimal procedures under independence after whitening of the profile of explanatory variables. However, in high-dimensional contexts, most of those optimal solutions cannot be implemented, which re-opens the question of the best handling of dependence.

Surprisingly, many authors report the good performance of procedures ignoring dependence across explanatory variables or even their superiority with respect to procedures especially designed to handle dependence (Wu et al., 2014; Barnett et al., 2017; Dudoit et al., 2002; Bickel and Levina, 2004). This thesis aims at demonstrating that the best handling of dependence should not be restricted to the dichotomy between ignorance and whitening. Indeed, more flexible approaches are needed to optimally account for the interplay between the association signal $\boldsymbol{\beta}$ and the dependence across explanatory variables.

In the remainder of this manuscript, these points will be discussed in further detail and some consequences of ignoring dependence or taking it explicitly into account will be exposed. It will be demonstrated that the power of global testing methods highly depends on both the pattern of the signal and the correlation matrix of the test statistics. In particular, it is shown that for a given correlation matrix, two

different vectors $\boldsymbol{\beta}$ yield two very different vectors $\boldsymbol{\mu}$; although this point might seem obvious, it yields a great issue on whether dependence should be explicitly taken into account or not. It appears that, as demonstrated by the above example, choosing to decorrelate the test statistics can greatly improve power, or on the contrary can deter power. Adaptive methods are developed and proposed as possible alternatives to allow for a flexible handling of dependence, depending on the correlation matrix of the explanatory variables and their association with the response.

# 4 A focus on some specific dependence patterns

Although we will not make specific assumptions on the pattern of dependence across explanatory variables, our approaches will mainly be illustrated in fields of applications for which the handling of dependence is of great interest. This is especially true in GWAS where the profile of explanatory variables is made of thousands of genotype variables for SNPs composing the genome. An image plot of the correlation matrix of continguous SNPs forming the PDZRN4 gene of the human genome (Wellcome Trust Case Control Consortium, 2007) is shown on Figure 1.3(a). As illustrated on this figure, those variables inherit part of their dependence from their spatial alignment along the genome, namely because of the linkage disequilibrium (LD) mechanism. Note that, although each genotype is a categorical variable with 3 classes corresponding to the number (0, 1 or 2) of copies of the minor allele for a SNP, it is often considered as quantitative and the Pearson correlation coefficient is used by most authors as a good indicator of the dependence between two genotypes. Figure 1.3(a) clearly shows a strong and heterogeneous dependence structure, with overlapping sub-blocks of contiguous SNPs more markedly correlated. This partitioning in sub-blocks is observed for most genes but the sizes and numbers of such sub-blocks is very different from a region of the genome to another region. These blocks are often referred to as LD blocks and it is often assumed that these blocks provide a suitable partition of the whole genome (Wall and Pritchard, 2003; Dehman et al., 2015; Nakamoto et al., 2006; Twells et al., 2003).

In SNP-set search for significant association with a response $Y$, most often a phenotype variable, the pointwise test statistics for the association between each

SNP and $Y$ are aggregated to form a single test statistic for the association between the whole block, here a gene, and $Y$. Since the same testing approach is reproduced for hundreds of such blocks, this approach has to show reproducible performance whatever the within-block pattern of dependence and whatever the pattern of the association signal.

Strikingly, similar global testing approaches are implemented for functional ANOVA, which extends standard ANOVA to situations where the response is a curve (see Zhang (2013) for a review). For significance testing of the mean difference of groups of curves, the one-way design is the most common (Cuevas et al., 2004; Zhang and Liang, 2014). Assuming that curves are observations of continuous time Gaussian stochastic processes at discrete time points in a bounded and closed interval $\mathcal{T}$ , Ramsay and Silverman (2005) addressed the significance testing of an effect curve $\beta : \ t \mapsto \beta(t)$ by deriving pointwise test statistics $F_t$ for the null hypotheses $H_{0t} : \ \beta(t) = 0, \ t \in \mathcal{T}$. Zhang and Liang (2014) reviewed the main strategies for testing the global null hypothesis $H_0 = \bigcap_{t\in\mathcal{T}} H_{0t}$ by aggregating the pointwise tests statistics $F_t$. The superiority of the $\mathrm{L}^\infty$–norm of $(F_t)_{t\in\mathcal{T}}$ over $\mathrm{L}^2$-norm based tests is demonstrated by Zhang et al. (2019). However, aggregating pointwise test statistics by their integral over time or their supremum ignores the time dependence.

Figures 1.3(b) and (c) display the image plot of correlation matrices of functional data, estimated using public datasets of Near-Infrared Spectroscopy (NIRS) data. They exhibit a pronounced time-dependence, which is far more complex than the auto-correlation or Toeplitz structures commonly assumed for functional data in the literature, with blocks of strong positive correlation. In the following, we aim to show that accounting for the time dependence in the aggregation procedure can lead to considerable improvements for significance testing.

As previously stated, in this thesis, "strong dependence" refers to a situation where a correlation matrix can be accurately approximated by using only a small number of its eigenvectors, the remaining eigenvectors being considered as noise. This is closely related to the choice of a number $K$ of axes in principal component analysis. Even though there is no definitive consensus on a criterion for this choice, selecting eigenvectors associated to an eigenvalue greater than 1 (Kaiser, 1960) is a simple and commonly used criterion, and will be sufficient to give an insight on

the notion of strong dependence. For the PDZRN4 gene, 11 eigenvalues out of 68 are greater than 1. For the orange juice and wine near-infrared spectra datasets, 3 out of 700 and 6 out of 256 eigenvalues are greater than 1, respectively. In each situation, the estimated number of "useful" eigenvectors is much smaller than the dimension of the correlation matrix.

# 5 Organization of the thesis

The manuscript is organized as follows.

Chapter 2 is dedicated to the study of the impact of dependence on global testing methods, with a special focus on SNP-set testing approaches in GWAS. First, a selection of global testing methods are compared and the extent of which dependence affects their performance is studied. The ranking of the approaches turns out to depend on the conjunction of the dependence structure and the association signal. Based on the former observation, an adaptive method is proposed, called MGF-R Test, for Moment Generating Function - Ratio based Test. This method aims at more reproducible performance when applied to a variety of dependence patterns and association signals. This chapter led to the development of an R package `MGFRTest` and the submission of an article, currently under review.

In Chapter 3, mathematical developments are introduced to apply the global testing methods to the problem of detection of interaction effects between a gene and an environmental factor. In particular, a suitable model is introduced for the detection of these interaction effects. Results obtained in Chapter 2 are further extended in the context of interaction effects detection and the former adaptive method is applied to this problem. The formerly proposed MGF-R test is naturally applied to the detection of gene - environment interaction effects.

In Chapter 4, the global testing problem is extended to the problem of detection of interaction effects between two genes. A suitable model is introduced to construct gene - gene interaction effects tests. However, this problem yields several important issues. In particular, the dimensionality is much greater than in the previous chapters. Moreover, it will be shown that the dependence structure of the test statistics

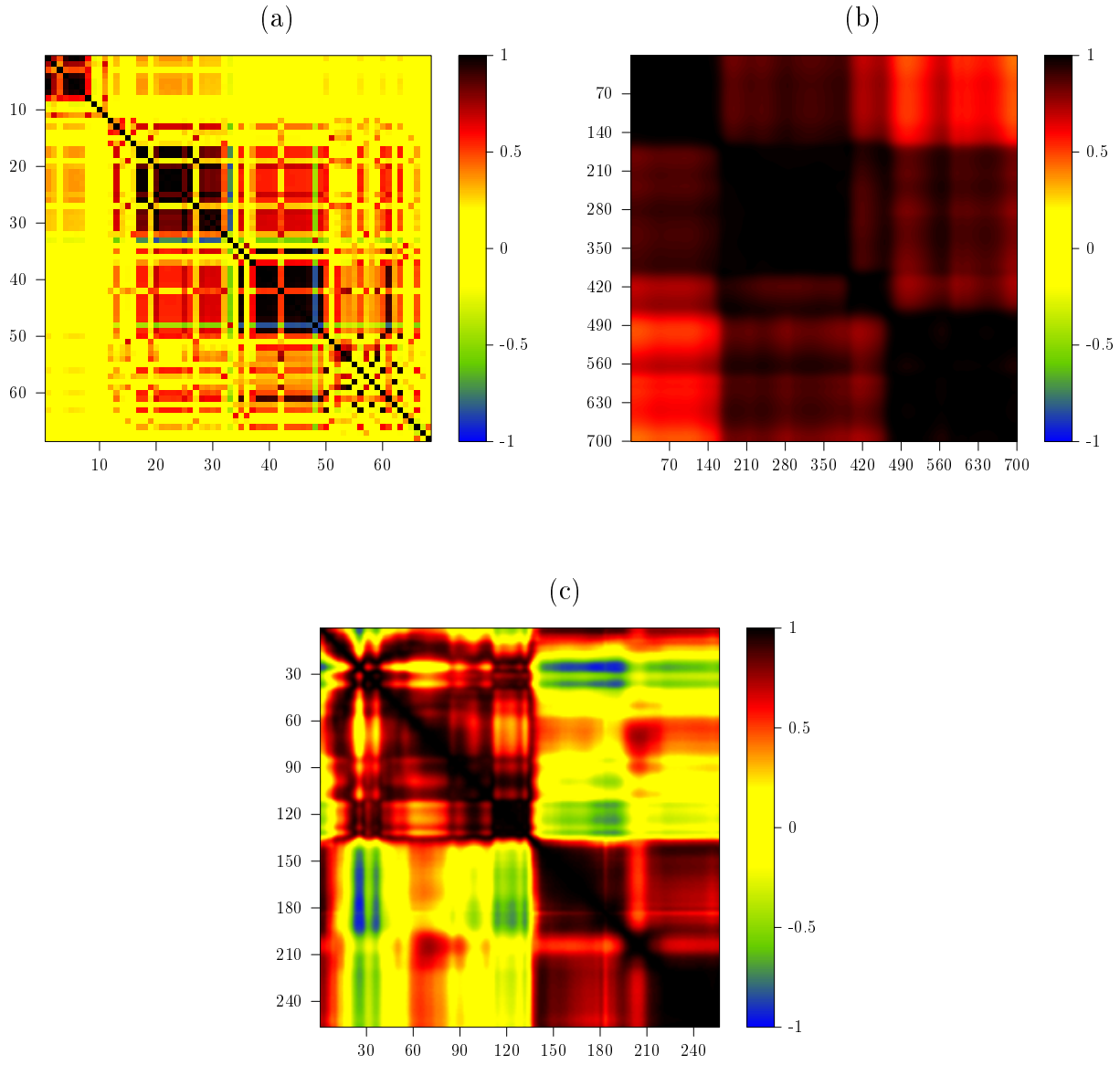


Figure 1.3: Examples of dependence structures (correlation matrices) from public datasets: SNP data from the PDZRN4 gene of the human genome (a) (Wellcome Trust Case Control Consortium, 2007), orange juice near-infrared spectra (b) and wine near-infrared spectra (c) (Zhang and Melnik, 2012)

vector has a very particular shape, as it is inherited from both genes included in the model. Some results are derived, indicating that despite the structure of the correlation matrix being very particular, it is globally sparser than the correlation matrices of the two considered genes, and its coefficients are smaller. Nevertheless, it is difficult to foresee how it will influence the performances of the global testing methods. Indeed, these performances are also affected by the great dimensionality of the test statistics vector.

In Chapter 5, the effect of dependence on classification and regression methods is studied. Similar to the results of the previous chapters, it is demonstrated that the ranking of a selection of prediction methods depends not only on the dependence structure of the explanatory variables, but also on the association signal between the predictors and the response variable. In high dimensional settings, "naive" approaches ignoring dependence among explanatory variables can outperform methods taking dependence into account. A new class of methods is introduced, embedding both naive and decorrelation-based prediction rules. A strategy to search for an optimal procedure within this class is proposed. The results obtained through a simulation study and on non-simulated datasets show that the proposed method generally yields very satisfying performances.

Finally, in Chapter 6, a conclusion is drawn and the main results of this thesis are recalled and discussed. The results obtained during this thesis were presented in conferences and articles. The full list of works (R packages implementing the proposed methods, oral presentations and submitted articles) is given after Chapter 6, with appendices detailing the features of the R packages.

# Chapter 2

# Adaptive Handling of Dependence for Global Testing

**Abstract.** Handling dependence or not remains an open issue in signal detection, where a large number of pointwise test statistics are aggregated to simultaneously test for a collection of null hypotheses. In various fields such as genetic epidemiology or functional data analysis, many testing methods for signal detection have been proposed, some ignoring dependence across pointwise test statistics whereas others introduce a model for decorrelation, with unclear conclusions on their relative performance. Indeed, the benefit that can be expected from decorrelation highly depends on the interplay of the patterns of dependence across pointwise test statistics and of the distribution of the non-null features in the true signal. Within a large class of test statistics covering a continuum of whitening approaches, we introduce an optimal procedure maximizing a Cumulant Generating Function-based distance between the null and nonnull distributions, in order to adapt the aggregation of the pointwise statistics to the pattern of non-null features. Using data-driven simulations, we demonstrate that the ability of the present test to detect a signal is more robust to the dependence structure than existing methods. We further apply the present test to two Genome-Wide Assosiation Studies and show its ability to efficiently detect validated associations in various situations.

# Contents

**1 Introduction . . . . . 51**

**2 Signal detection by aggregation of pointwise test statistics 53**

2.1 A general framework for the detection of an association signal . . . . . 54

2.2 Aggregation procedures of pointwise statistics . . . . . 56

2.3 Impact of the pattern of association signal on detection performance . . . . . 57

**3 Weighted decorrelation of pointwise test statistics . . . . 60**

3.1 Oracle weighting procedure . . . . . 60

3.2 Implementation of the MGF-R weighting procedure . . . . 64

**4 Comparison of detection methods for association signals 65**

4.1 Simulation study . . . . . 65

4.2 Application to real datasets . . . . . 69

**5 Discussion . . . . . 71**

**6 Appendix: proof of Theorem 3.1 . . . . . 74**

# 1 Introduction

In many research fields, signal detection is viewed as the simultaneous test of pointwise null hypotheses, *e.g.* over a time interval in functional Analysis of Variance (fANOVA), over a specific segment of the genome in Genome Wide Association Studies (GWAS) or over a two-dimensional region of an image in functional Magnetic Resonance Imaging (fMRI). In the former situations where the number of features is usually large, sometimes larger than the sample size, such testing issues are generally addressed by deriving a global test statistic for the conjunction of null hypotheses from the aggregation of the corresponding pointwise test statistics. The diversity of existing aggregation methods (see the reviews by Zhang and Liang (2014) for fANOVA and Derkach et al. (2014) for GWAS issues) reflects the difficulty to identify a method that would show a good detection performance in a wide scope of situations. As reported by Cai et al. (2014) for the two-group mean comparison issue in high-dimension, the possibly strong dependence across pointwise test statistics turns out to be a crucial point in the comparative studies of aggregation procedures. Besides, several studies also investigate the influence on statistical power of the pattern of the true signal with respect to its sparsity rate (Donoho and Jin, 2004; Arias-Castro et al., 2011; Zhong et al., 2013).

However, the most popular whole-interval or whole-region testing methods, both in fANOVA and in GWAS, are based on simple aggregations of pointwise test statistics, not especially designed to be optimal under dependence. For example, Ramsay et al. (2009) suggest using the maximum absolute pointwise test statistics, which turns out to be analogous to the famous maxT or minP procedure, proposed by Westfall and Young (1993) (see also Conneely and Boehnke (2007)) to test for significant relationship between genotypes of a given set of single nucleotide polymorphisms (SNPs) and a case/control group membership in the context of GWAS. A functional F-type test statistic based on the squared $L^2$–norm of the vector of pointwise test statistics is also introduced by Zhang (2013), whereas similar weighted or unweighted $L^2$-norm statistics are recommended by many authors (Liu et al., 2010; Wu et al., 2011; Derkach et al., 2014) for GWAS issues.

The choice of an appropriate method to aggregate pointwise test statistics falls into the general context of global testing as defined by Arias-Castro et al. (2011).

They especially focus on the impact of the sparsity rate of the true signal on the choice between the $L^2$-norm based test statistics of standard Analysis of Variance and the Higher Criticism (Donoho and Jin, 2004), under assumptions of weak dependence. The former Higher Criticism (HC) test statistic can be viewed as a Kolmogorov-Smirnov type distance between the standardized empirical distribution of the pointwise $p$-values and the theoretical uniform null distribution. If the pointwise test statistics are assumed to be independent and in the so-called Rare-and-Weak paradigm, defined by conditions on the amplitude and sparsity rate of the true signal, Donoho and Jin (2004, 2008) show that HC reaches the optimal detection bounds obtained by Ingster (1997).

In this general framework, it is commonly observed that, whatever the aggregation method, detection performance for a given true signal can be affected by dependence across pointwise test statistics. A growing number of studies therefore suggests that signal detection procedures can be improved by aggregating decorrelated pointwise test statistics, as for instance in Hall and Jin (2008) and Hall and Jin (2010) for HC and Ahdesmäki and Strimmer (2010) for the slightly different feature selection issue in two-group classification models. The innovated higher criticism (iHC) proposed by Hall and Jin (2010) first performs a whitening transformation of the original test statistics using the Cholesky decomposition of the correlation matrix and then applies the higher criticism after the transformation. Similarly, Ahdesmäki and Strimmer (2010) introduce Correlation-Adjusted t-scores based on a James-Stein shrinkage estimate of correlations.

However, as discussed in Bickel and Levina (2004) in the closely related two-group classification issue, the potential gain in detection performance that can be expected from decorrelation remains unclear (see Wu et al. (2014); Barnett et al. (2017) in the GWAS context). Indeed, arguing that whitening may generate noise and weaken the signal, Barnett et al. (2017) recently introduced the generalized higher criticism (GHC) where aggregation is performed on the raw pointwise statistics. Noticing that the benefit of using GHC highly depends on the pattern of dependence across pointwise statistics and on the signal sparsity rate, Barnett et al. (2017) propose to combine it with the maximum of the absolute pointwise test statistics and a weighted $L^2$-norm statistic in a simple omnibus test. While the omnibus test shows good detection performance in the simulation setup proposed in Barnett et al. (2017), it

raises limitations in terms of power and computational cost induced by the two-step Monte-Carlo calculation of the $p$-value.

To overcome these limitations, we propose a new approach, whose aim is to adapt the aggregation of pointwise tests to both the correlation structure and the pattern of the signal. For that purpose, we first introduce a class of test statistics defined as a weighted sum of the squared decorrelated statistics. Introducing weights in the sum of decorrelated statistics enables a flexible handling of correlation in the aggregation procedure and prevents from the dilution of the signal that can be induced from a complete whitening of the raw pointwise statistics. Optimal weights are derived by maximizing a Cumulant Generating Function-based distance between the null and non-null distributions of the test statistics. Our results, based on a large panel of data-driven simulations, demonstrate that the former method provides a good detection performance in a wide variety of situations.

In Section 2, we first introduce a general testing framework for signal detection based on a generalized linear model, which can be used in numerous situations. We next illustrate the joint role played by the spatial pattern of the association signal and the correlation among pointwise statistics in statistical power. In Section 3, we introduce a family of test statistics for signal detection defined as a weighted sum of squared decorrelated pointwise test statistics. We then propose a statistical procedure for an optimal choice of the vector of weights. A comparative study using simulations is presented in Section 4 followed by a study of two GWAS data.

# 2 Signal detection by aggregation of pointwise test statistics

The global testing issue is usually presented in the standard linear model framework under normality assumptions (Arias-Castro et al., 2011; Derkach et al., 2014). In order to cover a wider scope of applications, including especially GWAS issues where the outcome is a two-level categorical variable, we introduce a generalized linear model framework (Conneely and Boehnke, 2007; Wu et al., 2011; Barnett et al., 2017) in which we review the main existing aggregation methods for signal detection.

## 2.1 A general framework for the detection of an association signal

### 2.1.1 Signal detection

In the following, $Y$ denotes the outcome variable and $\boldsymbol{X} = (X_1, \ldots, X_p)$ a $p$–profile of explanatory variables. In most fANOVA settings, $\boldsymbol{X}$ contains the discretized observations of a curve and $Y$ can either be assumed to be normally distributed given $\boldsymbol{X} = \boldsymbol{x}$ in scalar-on-function regression issues or can be a grouping variable for supervised classification of functional data. In GWAS issues, $Y$ is often a two-class (case/control) variable and $\boldsymbol{X} = (X_1, \ldots, X_p)$ a profile of genetic markers observed for $p$ Single-Nucleotide Polymorphisms in a given region of the genome.

An optional set of $q$ covariates $U_1, \ldots, U_q$ is also considered, if available. Hereafter, the following generalized linear model is assumed:

$$h(\mathbb{E}[Y|\boldsymbol{U} = \boldsymbol{u}, \boldsymbol{X} = \boldsymbol{x}] = \boldsymbol{u}'\boldsymbol{\alpha} + \boldsymbol{x}'\boldsymbol{\beta} \tag{2.1}$$

where $h$ is a link function, $\boldsymbol{\alpha}$ is the $q$-vector of coefficients for the covariates and $\boldsymbol{\beta}$ is the $p$-vector of coefficients for the explanatory variables.

The link function $h$ is often chosen to be the identity function, as in continuous trait analysis, or the logistic function for case-control studies (Wu et al., 2014; Vukcevic et al., 2011; Zhao et al., 2017). In the above framework, signal detection is viewed as a global test for the significance of the association signal $\boldsymbol{\beta}$:

$$\begin{cases} H_0 : & \boldsymbol{\beta} = \mathbf{0} \\ H_1 : & \boldsymbol{\beta} \neq \mathbf{0} \end{cases} \tag{2.2}$$

### 2.1.2 A global framework for signal detection

Hereafter, $\boldsymbol{y} = (y_1, \ldots, y_n)'$ denotes the $n$–vector containing the observations of the outcome variable, $\mathbb{X}$ the $n \times p$ matrix whose $j$th column $\boldsymbol{X_j}$ contains the observed values $(x_{1j}, \ldots, x_{nj})$ of $X_j$, $j = 1, \ldots, p$. Similarly, the values of the covariates are stacked in a $n \times q$ matrix $\mathbb{U}$.

In large-dimensional situations, namely when $p$ is large with respect to the sample size $n$, standard Likelihood-Ratio Tests or Fisher tests for (2.2) cannot be implemented. An alternative approach is to form a global test statistic from the aggregation of pointwise test statistics $Z_j$, $j = 1, \ldots, p$ for the marginal association between the outcome and each explanatory variable. In the standard linear regression model, a usual choice for $Z_j$ is the t-test statistic for the significance of a regression parameter. Although many such statistics have been proposed in the literature, we focus here on the most used pointwise procedure in GWAS issues. Let $\hat{\boldsymbol{\alpha}}_{\mathbf{0}}$ be the estimator of $\boldsymbol{\alpha}$ under $H_0$. We denote $\hat{\boldsymbol{y}}_{\mathbf{0}}$ the $n$-vector whose $i$-th coordinate equals $h^{-1}(\boldsymbol{u}_i'\hat{\boldsymbol{\alpha}}_{\mathbf{0}})$, where $\boldsymbol{u}_i$ is the $i$th row of $\mathbb{U}$. For testing the significance of an association between $Y$ and $X_j$, the following marginal score test statistic is often used (Conneely and Boehnke, 2007; Barnett et al., 2017):

$$Z_j = \frac{\boldsymbol{X}_j'(\boldsymbol{y} - \hat{\boldsymbol{y}}_{\mathbf{0}})}{\sqrt{\widehat{\boldsymbol{\Gamma}}_{j,j}}}.$$

where $\widehat{\boldsymbol{\Gamma}}$ is the estimated covariance matrix of the vector $\mathbb{X}'(\boldsymbol{y} - \hat{\boldsymbol{y}}_{\mathbf{0}})$. Conneely and Boehnke (2007) proposed the following estimate:

$$\widehat{\boldsymbol{\Gamma}} = \hat{\sigma}_Y^2(\mathbb{X}'\mathbb{X} - \mathbb{X}'\mathbb{U}(\mathbb{U}'\mathbb{U})^{-1}\mathbb{U}'\mathbb{X})$$

where $\hat{\sigma}_Y^2 = \dfrac{1}{n}(\boldsymbol{y} - \hat{\boldsymbol{y}}_{\mathbf{0}})'(\boldsymbol{y} - \hat{\boldsymbol{y}}_{\mathbf{0}})$.

To sum up, as in the Rare-and-Weak paradigm introduced by Donoho and Jin (2004) for the Higher Criticism procedure, it will be assumed that the $p$-vector $\boldsymbol{\mathcal{Z}} = (Z_1, \ldots, Z_p)'$ of marginal test statistics for the association between $Y$ and $X_j$ is asymptotically normally distributed with mean $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_p)'$ and positive definite correlation matrix $\boldsymbol{\Sigma}$:

$$\boldsymbol{\mathcal{Z}} = (Z_1, \ldots, Z_p)' \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \tag{2.3}$$

The original testing problem (2.2) is then equivalent to testing the global nullity of $\boldsymbol{\mu}$:

$$\begin{cases} H_0: \ \boldsymbol{\mu} = \mathbf{0} \\ H_1: \ \boldsymbol{\mu} \neq \mathbf{0} \end{cases} \quad i.e. \quad \begin{cases} H_0: \ \forall j \in \{1, \ldots, p\}, \ \mu_j = 0 \\ H_1: \ \exists j \in \{1, \ldots, p\}, \ \mu_j \neq 0. \end{cases} \tag{2.4}$$

Note that the set of null $\boldsymbol{\beta}$ coefficients, $I_{\boldsymbol{\beta}} = \{j \in \{1, \ldots, p\}, \beta_j = 0\}$, is generally different from the set of null $\boldsymbol{\mu}$ coefficients $I_{\boldsymbol{\mu}} = \{j \in \{1, \ldots, p\}, \mu_j = 0\}$, thus

conferring a different sparsity rate for $\boldsymbol{\beta}$ and $\boldsymbol{\mu}$. This point is further illustrated through the detailed analysis of two simulated scenarios in Section 2.3.

## 2.2 Aggregation procedures of pointwise statistics

For the testing issue (2.4), a combined test statistic $T(\boldsymbol{\mathcal{Z}})$ is obtained by aggregating the pointwise statistics $Z_j$. In this section, we introduce the most commonly used aggregation methods. Table 2.1 provides a summary of the formulation of each global statistic.

One of the main aggregation methods is the maximum absolute pointwise test statistics (see Ramsay et al. (2009) for fANOVA and Westfall and Young (1993); Conneely and Boehnke (2007) for GWAS), for which $T(\boldsymbol{\mathcal{Z}}) = T_{\max}(\boldsymbol{\mathcal{Z}}) = \max_j |Z_j|$ (see Table 2.1). The former method, called minP or maxT by Westfall and Young (1993), is indeed often preferred to alternatives because its performances, both in terms of type I error control and ability to reveal a true signal, turn out to be generally good and also resistant to strong correlations between the elements of $\boldsymbol{\mathcal{Z}}$ (Wu et al., 2014).

The squared $L^2$-norm of $\boldsymbol{\mathcal{Z}}$, named as $T_2(\boldsymbol{\mathcal{Z}})$ and defined as the sum of the squared $Z_i$ was also proposed by Liu et al. (2010) in the GWAS context and by Shen and Faraway (2004) to define a functional F-test. In order to increase the contribution of explanatory variables with smaller variance, several weighted versions of the $L^2$-norm statistic were also introduced. Especially designed for the identification of rare variants in the GWAS context, the Sequence Kernel Association Test (SKAT, Wu et al. (2011)) is one of them, for which weights are defined via the Beta density function (see Table 2.1).

A Hotelling's t-square type statistic is also sometimes used, denoted by $T_{\text{Hotelling}}(\boldsymbol{\mathcal{Z}})$ (see Derkach et al. (2014) for applications to GWAS issues). It can just be viewed as a whitened version of the $L^2$-norm statistic. Indeed, provided $\widehat{\boldsymbol{\Sigma}}$ is a consistent estimate of $\boldsymbol{\Sigma}$ and $\widehat{\boldsymbol{\Sigma}}^{-1}$ exists, $T_{\text{Hotelling}}(\boldsymbol{\mathcal{Z}})$ is the $L^2$-norm of the decorrelated vector $\widehat{\boldsymbol{\Sigma}}^{-1/2}\boldsymbol{\mathcal{Z}}$, where $\widehat{\boldsymbol{\Sigma}}^{-1/2}$ is any $p \times p$ matrix satisfying $(\widehat{\boldsymbol{\Sigma}}^{-1/2})'\widehat{\boldsymbol{\Sigma}}^{-1/2} = \widehat{\boldsymbol{\Sigma}}^{-1}$.

The Higher Criticism (HC) signal detection method introduced by Donoho and

Jin (2004) can also be included into the above testing framework for signal detection. The HC test statistic aggregates the pointwise test statistics by forming a Kolmogorov-Smirnov type distance between the empirical probability distribution function of the individual $p$-values and the uniform null distribution. Some other versions are based on a distance between $S(t) = \sum_{j=1}^{p} \mathbf{1}_{\{|Z_j| \geq t\}}$ and its expectation under the null hypothesis, the distance being scaled by the theoretical standard error of $S(t)$. Under the null, the distribution of $S(t)$ is indeed binomial.

However, when the $Z_j$'s are not independent, $S(t)$ is no longer binomial and the scaling factor in the HC statistic is no longer valid. Consistently, Barnett et al. (2017) propose to scale the HC statistic by the empirical standard error of $S(t)$ under dependence in the GHC method. In order to improve the robustness of the former test to various patterns of dependence, Barnett et al. (2017) also suggest to add a second level of aggregation in the $T_{\text{Omnibus}}$ statistic: the $p$-value of the $T_{\text{Omnibus}}$ statistic is indeed the smallest of the $p$-values obtained with minP, SKAT and GHC.

The null distribution of the global test statistics mentioned above is generally unknown and is thus estimated by using permutations or resampling techniques. For simplicity, in the following, no covariates are used; therefore, simple permutations of the vector of phenotype values will be used to estimate the distributions. However, note that in presence of covariates having a significant effect on the phenotype, the type I error rate can be overcontrolled or undercontrolled by using simple permutations. Resampling techniques taking account of such effects, such as parametric bootstrap or biased permutation methods (Bůžková et al., 2011; Epstein et al., 2012) discussed in Chapters 3 and 4, are generally more appropriate to provide a proper control of the type I error rate.

## 2.3 Impact of the pattern of association signal on detection performance

Many authors (Derkach et al., 2014; Barnett et al., 2017) report that the relative power of aggregation tests highly depends on the pattern of correlation across the pointwise test statistics. The following illustrative study aims at highlighting the fact that, for a given correlation structure, the ranking of methods can also be

Table 2.1: Definitions of popular aggregation methods, where Beta is the Beta density function and $\text{MAF}_j$ is the Minor Allele Frequency of the $j$th SNP. $\Phi$ is the cumulative distribution of the standard normal distribution. $S(t)$ is defined as $S(t) = \sum_{j=1}^{p} \mathbf{1}_{\{|Z_j| \geq t\}}$. $p_{\text{GHC}}$, $p_{\text{minP}}$ and $p_{\text{SKAT}}$ are the $p$-values respectively obtained for the aggregation method GHC, minP and SKAT.

| Method | Statistic | References |
|---|---|---|
| minP | $T_{\max}(\boldsymbol{\mathcal{Z}}) = \max_{1 \leq j \leq p} \lvert Z_j \rvert$ | Westfall and Young (1993)<br>Conneely and Boehnke (2007) |
| $L^2$-norm | $T_2(\boldsymbol{\mathcal{Z}}) = \boldsymbol{\mathcal{Z}}'\boldsymbol{\mathcal{Z}} = \sum_{j=1}^{p} Z_j^2$ | Liu et al. (2010) |
| Hotelling | $T_{\text{Hotelling}}(\boldsymbol{\mathcal{Z}}) = \boldsymbol{\mathcal{Z}}'\widehat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{\mathcal{Z}}$ | Derkach et al. (2014)<br>Luo et al. (2010) |
| SKAT | $T_{\text{SKAT}}(\boldsymbol{\mathcal{Z}}) = (\boldsymbol{D}_{\widehat{\boldsymbol{\Gamma}}}^{1/2}\boldsymbol{\mathcal{Z}})'\boldsymbol{W}(\boldsymbol{D}_{\widehat{\boldsymbol{\Gamma}}}^{1/2}\boldsymbol{\mathcal{Z}})$<br>with $\boldsymbol{W} = \text{diag}(w_j)$, $\boldsymbol{D}_{\widehat{\boldsymbol{\Gamma}}} = \text{diag}(\widehat{\boldsymbol{\Gamma}})$,<br>$w_j = \text{Beta}(\text{MAF}_j, 1, 25)$ | Wu et al. (2011) |
| HC | $T_{\text{HC}}(\boldsymbol{\mathcal{Z}}) = \max_{1 \leq j \leq p/2} \sqrt{p} \dfrac{j/p - p_{(j)}}{\sqrt{p_{(j)}(1-p_{(j)})}}$ | Donoho and Jin (2004)<br>Wu et al. (2014) |
| GHC | $T_{\text{GHC}}(\boldsymbol{\mathcal{Z}}) = \max_{t \geq t_0} \dfrac{S(t) - 2p(1-\Phi(t))}{\sqrt{\widehat{\text{var}}(S(t))}}$<br>with $S(t) = \sum_{i=1}^{p} \mathbf{1}_{\{\lvert Z_i \rvert \geq t\}}$ | Barnett et al. (2017) |
| Omnibus | $T_{\text{Omnibus}} = \min(p_{\text{GHC}}, p_{\text{minP}}, p_{\text{SKAT}})$ | Barnett et al. (2017) |

highly modified according to the pattern of the true association signal $\boldsymbol{\mu}$ (see Equation (2.3)). Indeed, not only the sparsity rate but more generally the homogeneity of the distribution of non-null features define different patterns for $\boldsymbol{\mu}$. Note that in most papers addressing signal detection issues by aggregation tests, a rare signal usually refers to a sparse vector $\boldsymbol{\beta}$ in Equation (2.1), which, due to the correlation among the variables, does not imply the sparsity of $\boldsymbol{\mu}$. To illustrate this point, we investigate the relative power of aggregating methods in two slightly different simulated scenarios. In a first scenario (scenario 1), we assume that the vector $\boldsymbol{\beta}$ is sparse while in the second scenario (scenario 2), $\boldsymbol{\beta}$ is no longer sparse by considering additional non-zero coefficients.

In our simulation pipeline, we used a fixed realistic design based on a block of $p = 64$ genetic markers on chromosome 1 in a publicly available GWAS dataset (Wellcome Trust Case Control Consortium, 2007). For each of the 1,000 simulated datasets, 1,000 independent profiles of genotype variables whose both marginal and joint distributions correspond to the observed block are generated using the R package `GenOrd` (Barbiero and Ferrari, 2015a). Then, in the sparse case of scenario 1, only the SNPs located at the 5th and 10th positions are considered to have non-zero coefficients (*i.e.* $\beta_5 \neq 0$, $\beta_{10} \neq 0$ and $\beta_i = 0 \; \forall i \in \{1, \ldots, 64\}$ with $i \neq 5$ and $i \neq 10$). In scenario 2, a few other coefficients around those SNPs 5 and 10 are also set to nonzero. For each simulated dataset, the vector $\boldsymbol{\mathcal{Z}}$ of marginal score test statistics as defined above is calculated and the association signal $\boldsymbol{\mu}$ is estimated by averaging over these 1,000 simulated vectors. The two illustrative choices of $\boldsymbol{\beta}$ are plotted in gray in Figure 2.1, while the association signal $\boldsymbol{\mu}$ is displayed in black. It can be remarked that for each scenario, the two patterns of association $\boldsymbol{\beta}$ and $\boldsymbol{\mu}$ are very different, especially regarding their sparsity rates.

For each simulated dataset in these two scenarios, the power of the seven aggregation methods previously introduced ($L^2$-norm, minP, HC, GHC, Hotelling, SKAT and Omnibus) has been estimated with a Monte-Carlo estimation of the null distribution based on 1,000 random permutations of the observations of $Y$. To simulate $Y$, a vector $\boldsymbol{\beta}_{\mathbf{max}}$ corresponding to the maximal signal strength has been set so that the power of the most powerful test is close to 1. Then, $\boldsymbol{\beta}_{\mathbf{max}}$ was multiplied by a signal strength parameter $\xi$ taking values in $\{0, 0.1, 0.2, \ldots, 1\}$. As a result, the signal pattern remains unchanged but the signal strength varies regularly between 0 and 1, so that the power of each method is estimated as a function of $\xi$ in each scenario.

The estimated power curves displayed in Figure 2.1 first show that the ranking of aggregation methods highly depends on the pattern of the true association signal $\boldsymbol{\mu}$. Such a result confirms that no single aggregation test can be the most powerful in every design setup defined as a combination between the correlation structure and the pattern of the signal (Barnett et al., 2017). Let us further compare tests based on a quadratic form of $\boldsymbol{\mathcal{Z}}$, namely $L^2$-norm, Hotelling and SKAT. It can be remarked that, in scenario 1, the $L^2$-norm test is the most powerful method (with HC, GHC and minP) while Hotelling and SKAT are the two worst methods.

However, in scenario 2, rankings are reversed since Hotelling is undoubtedly the most powerful method while $L^2$-norm and SKAT perform as the worst methods. To sum up, our results show that choosing an appropriate quadratic form of $\boldsymbol{\mathcal{Z}}$ may allow to maximize the power of detection. However, recalling that the Hotelling's t-square test is a whitened version of the $L^2$-norm test, we also demonstrate that the potential benefit of decorrelating the test statistics highly depends on the pattern of the association signal $\boldsymbol{\mu}$. In the next section, we therefore introduce a new class of aggregation methods allowing for a flexible whitening of the test statistics.

# 3 Weighted decorrelation of pointwise test statistics

Suppose first that $\boldsymbol{\Sigma}$ is known, with eigendecomposition $\boldsymbol{P}\boldsymbol{\Lambda}\boldsymbol{P}'$ where $\boldsymbol{P}$ is a $p \times p$ matrix such that $\boldsymbol{P}'\boldsymbol{P} = \mathbf{I}_p$ and $\boldsymbol{\Lambda}$ is a $p \times p$ diagonal matrix whose diagonal entries are positive eigenvalues $\lambda_j,\ j = 1, \ldots, p$. Let $\boldsymbol{\mathcal{Z}}^* = \boldsymbol{P}'\boldsymbol{\mathcal{Z}}$ denote the decorrelated version of $\boldsymbol{\mathcal{Z}}$. Under the asymptotic normality assumption on $\boldsymbol{\mathcal{Z}}$ introduced in the previous section, $\boldsymbol{\mathcal{Z}}^*$ is also asymptotically normally distributed with $\mathbb{E}[\boldsymbol{\mathcal{Z}}^*] = \boldsymbol{P}'\boldsymbol{\mu}$ and $\mathrm{Var}(\boldsymbol{\mathcal{Z}}^*) = \boldsymbol{\Lambda}$.

Now, let us introduce the class of aggregated test statistics $T_2(\boldsymbol{\mathcal{Z}}^*, \boldsymbol{h})$ for $H_0 : \boldsymbol{\mu} = \mathbf{0}$ defined as weighted sums of the squared coordinates $Z_j^*$ of $\boldsymbol{\mathcal{Z}}^*$:

$$T_2(\boldsymbol{\mathcal{Z}}^*, \boldsymbol{h}) = \sum_{j=1}^{p} h_j [Z_j^*]^2. \tag{2.5}$$

It can be remarked that $T_2(\boldsymbol{\mathcal{Z}}^*, \boldsymbol{h})$ encompasses $T_2(\boldsymbol{\mathcal{Z}})$ and $T_{\text{Hotelling}}(\boldsymbol{\mathcal{Z}})$ defined respectively as the $L^2$-norm, with $h_j = 1,\ j = 1, \ldots, p$, and the Hotelling's t-square, with $h_j = 1/\lambda_j,\ j = 1, \ldots, p$, test statistics (see Table 2.1). We propose hereafter to search for an optimal vector of weights $\boldsymbol{h} = (h_1, \ldots, h_p)'$ and deduce relevant weighting procedures.

## 3.1 Oracle weighting procedure

Since $\boldsymbol{\mathcal{Z}}^*$ is normally distributed with mean $\boldsymbol{P}'\boldsymbol{\mu}$ and variance matrix $\boldsymbol{\Lambda}$, then $T_2(\boldsymbol{\mathcal{Z}}^*, \boldsymbol{h})$ is distributed as a weighted sum $\sum_{j=1}^{p} h_j \lambda_j \chi_1^2(\gamma_j)$, where the $\chi_1^2$ variables

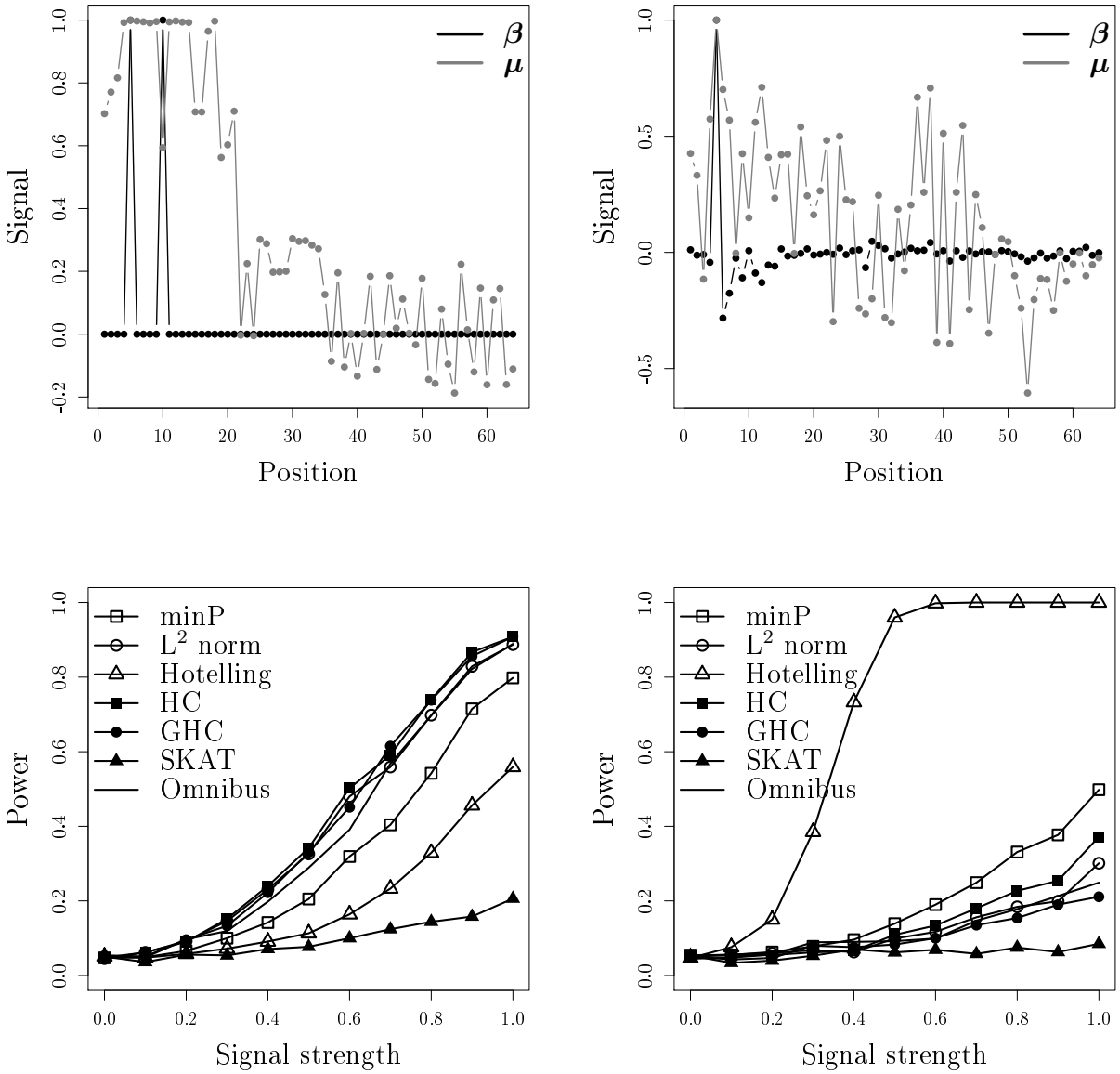


Figure 2.1: Power of each test in two different scenarios. Upper part: Regression and corresponding association signals in two GWAS scenarios. Lower part: Detection rates of the $L^2$-norm test, minP, HC and GHC, the Hotelling's t-square test, SKAT and the omnibus test (with type I error level $\alpha = 0.05$). Left part corresponds to scenario 1, right part to scenario 2.

in the former sum are mutually independent and $\gamma_j = (\boldsymbol{p}_j'\boldsymbol{\mu})^2/\lambda_j$ are non-centrality parameters and $\boldsymbol{p_j}$ is the $j$-th column of $\boldsymbol{P}$.

We propose to choose the vector $\boldsymbol{h}$ of weights that maximizes a distance between the null and non-null distributions of $T_2(\boldsymbol{\mathcal{Z}}^*, \boldsymbol{h})$. In the present context, such an optimization issue with the most classical distribution-based distances, such as Cramér-Von Mises-type distances or Kullback-Leibler divergence, requires untractable calculations of the cumulative or probability distribution functions of weighted sums of noncentral chi-square distributions.

However, the Moment Generating Function

$$\text{MGF}: \ t \mapsto \text{MGF}(t, \boldsymbol{h}, \boldsymbol{\lambda}, \boldsymbol{\gamma}) = \mathbb{E}[\exp(tT_2(\boldsymbol{\mathcal{Z}}^*, \boldsymbol{h}))],$$

or the Cumulant Generating Function

$$\text{CGF}: \ t \mapsto \text{CGF}(t, \boldsymbol{h}, \boldsymbol{\lambda}, \boldsymbol{\gamma}) = \log \text{MGF}(t, \boldsymbol{h}, \boldsymbol{\lambda}, \boldsymbol{\gamma}),$$

are convenient alternatives to characterize a distribution. Indeed, the CGF of mixtures of chi-square distributions has a simple closed form expression: for a given vector $\boldsymbol{h}$ of weights and for $t$ such that, for all $j = 1, \ldots, p$, $1 - 2th_j\lambda_j > 0$, the log-ratio of null and non-null Moment Generating Functions of $T_2(\boldsymbol{\mathcal{Z}}^*, \boldsymbol{h})$ is given by the following simple expression:

$$\text{CGF}(t, \boldsymbol{h}, \boldsymbol{\lambda}, \boldsymbol{\gamma}) - \text{CGF}_0(t, \boldsymbol{h}, \boldsymbol{\lambda}) \quad = \quad t \sum_{j=1}^{p} \frac{h_j \lambda_j \gamma_j}{1 - 2th_j\lambda_j}, \tag{2.6}$$

where $\text{CGF}_0(t, \boldsymbol{h}, \boldsymbol{\lambda}) = \text{CGF}(t, \boldsymbol{h}, \boldsymbol{\lambda}, \boldsymbol{\gamma} = \boldsymbol{0})$ is the null Cumulant Generating Function of $T_2(\boldsymbol{\mathcal{Z}}^*, \boldsymbol{h})$. It is deduced from expression (2.6) that the distance between the null and non-null distributions of $T_2(\boldsymbol{\mathcal{Z}}^*, \boldsymbol{h})$ depends jointly on the true association signal and the dependence across pointwise test statistics through the parameters $(\boldsymbol{\lambda}, \boldsymbol{\gamma})$.

The power of a test based on $T_2(\boldsymbol{\mathcal{Z}}^*, \boldsymbol{h})$ is invariant with respect to the multiplication of the vector of weights $\boldsymbol{h}$ by a scalar. Indeed, the distribution of the statistic is inevitably shifted and the critical value of the test has to be multiplied by the same scalar to maintain the control of the type I error rate, since $T_2(\boldsymbol{\mathcal{Z}}^*, \alpha\boldsymbol{h}) = \alpha T_2(\boldsymbol{\mathcal{Z}}^*, \boldsymbol{h})$ for any scalar $\alpha$. Therefore, we choose to consider that $\sum_{j=1}^{p} h_j = p$. First, using a

Lagrange multiplier technique to ensure that $\sum_{j=1}^{p} h_j = p$, it turns out that the optimal weights $\boldsymbol{h_t^*} = (h_{1t}^*, \ldots, h_{pt}^*)'$, obtained for a fixed value of $t$, admit closed-form expressions given by the following theorem.

**Theorem 3.1** *For a given value of $t$, the optimal weights maximizing the distance given by equation 2.6 have the following closed-form expression:*

$$h_{jt}^* = \frac{1}{2t\lambda_j} - \text{sign}(t)\left|p - \frac{1}{2t}\sum_{k=1}^{p}\frac{1}{\lambda_k}\right|\omega_j, \quad \omega_j = \frac{\sqrt{\frac{\gamma_j}{\lambda_j}}}{\sum_{k=1}^{p}\sqrt{\frac{\gamma_k}{\lambda_k}}}.$$

*It is straightforwardly checked that, for all $j = 1, \ldots, p$, $1 - 2th_{jt}^*\lambda_j > 0$, for all $t$.*

*Proof*: see Section 6.

Consequently, and still considering that the association signal $\boldsymbol{\mu}$, hence also $\boldsymbol{\gamma}$, is known, the optimal test statistic $T_2(\boldsymbol{\mathcal{Z}}^*, \boldsymbol{h_t^*})$ can be viewed as a linear combination of Hotelling's statistic $T_2(\boldsymbol{\mathcal{Z}}^*, 1/\boldsymbol{\lambda}) = \sum_{j=1}^{p}[Z_j^*]^2/\lambda_j$ and $T_2(\boldsymbol{\mathcal{Z}}^*, \boldsymbol{\omega}) = \sum_{j=1}^{p}\omega_j[Z_j^*]^2$:

$$T_2^*(\boldsymbol{\mathcal{Z}}^*, \boldsymbol{h_t^*}) = T_2(\boldsymbol{\mathcal{Z}}^*, 1/\boldsymbol{\lambda}) - u_t T_2(\boldsymbol{\mathcal{Z}}^*, \boldsymbol{\omega}),$$

where $u_t = 2p|t - \frac{1}{2p}\sum_{k=1}^{p}\frac{1}{\lambda_k}|$, with $u_t \geq 0$. Note that the special case $u_t = 0$, or equivalently $t = \frac{1}{2p}\sum_{k=1}^{p} 1/\lambda_k$, gives the Hotelling's t-square test. $u_t$ can therefore be viewed as a tuning parameter to maximize the power of the signal detection procedure within the family of Oracle test statistics $T_2(\boldsymbol{\mathcal{Z}}^*, \boldsymbol{h_t^*})$. Depending on the patterns of the true association signal and the within-block correlations among pointwise test-statistics, the best choice of $u_t$ may be close to zero to favor a full decorrelation of the pointwise test statistics or, on the contrary, may be very large to temper the decorrelation.

This is illustrated by Figure 2.2 that displays the detection rates of the $L^2-$norm test $T_2(\boldsymbol{\mathcal{Z}}^*, \mathbf{1})$, the Hotelling's t-square test $T_2(\boldsymbol{\mathcal{Z}}^*, 1/\boldsymbol{\lambda})$ and the optimal test statistic $T_2(\boldsymbol{\mathcal{Z}}^*, \boldsymbol{h}^*)$ within the family of Oracle test statistics $T_2(\boldsymbol{\mathcal{Z}}^*, \boldsymbol{h_t^*})$ in the two scenarios $(\boldsymbol{\lambda}, \boldsymbol{\gamma})$ for an association signal on an LD block in chromosome 1 introduced in Section 2 (see Figure 2.1). In scenario 1, whatever the signal strength, the best choice of $u_t$ is very large whereas it is close to 0 in scenario 2. In both scenarios, it

turns out that there exists a vector of weights $\boldsymbol{h}^*$ ensuring that $T_2(\boldsymbol{\mathcal{Z}}^*, \boldsymbol{h}^*)$ shows better detection rates than the $\mathrm{L}^2-$norm test of the pointwise test statistics and the Hotelling's t-square test.

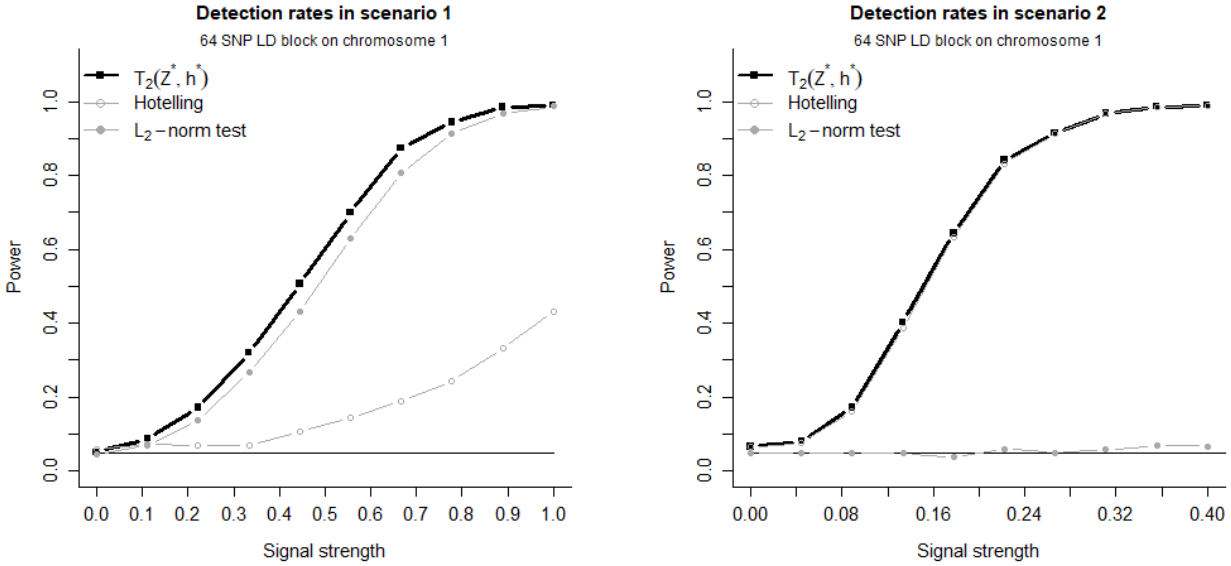


Figure 2.2: Power curves of the oracle test statistic $T_2(\boldsymbol{\mathcal{Z}}^*, \boldsymbol{h}^*)$, the $\mathrm{L}^2$-norm test $T_2(\boldsymbol{\mathcal{Z}}^*, \mathbf{1})$ and the Hotelling's t-square test $T_2(\boldsymbol{\mathcal{Z}}^*, 1/\lambda)$ in the two scenarios for an association signal $\boldsymbol{\mu}$ introduced in Section 2 (left: scenario 1, right: scenario 2).

## 3.2 Implementation of the MGF-R weighting procedure

A plug-in version $\hat{\boldsymbol{h}}_t^* = \boldsymbol{h}_t^*(\hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\lambda}})$ of the vector of oracle weights is now proposed, where $(\hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\lambda}})$ is a consistent estimate of $(\boldsymbol{\gamma}, \boldsymbol{\lambda})$. Hereafter, $\hat{\boldsymbol{\lambda}}$ is the $p$-vector of eigenvalues of the estimate $\widehat{\boldsymbol{\Sigma}}$ of $\boldsymbol{\Sigma}$. Similarly, the coordinates of $\hat{\boldsymbol{\gamma}}$ are defined by $\hat{\gamma}_i = [Z_i^*]^2/\hat{\lambda}_i$, where $\widehat{\boldsymbol{\mathcal{Z}}}^* = \hat{\boldsymbol{P}}'\boldsymbol{\mathcal{Z}}$ and $\hat{\boldsymbol{P}}$ is the $p \times p$ matrix of eigenvectors of $\widehat{\boldsymbol{\Sigma}}$.

The vector of weights $\hat{\boldsymbol{h}}_t^*$ is computed for a sequence of values for $t$. The distribution of $T_2(\widehat{\boldsymbol{\mathcal{Z}}}^*, \hat{\boldsymbol{h}}_t^*)$ can then be estimated by using permutations of $\boldsymbol{y}$. For each permutation, the vector of test statistics is obtained and the vector of optimal coefficients is computed for the same sequence of values of $t$. For each value of $t$, the $p$-value $p_t$ corresponding to $T_2(\widehat{\boldsymbol{\mathcal{Z}}}^*, \hat{\boldsymbol{h}}_t^*)$ is computed by using the analogous statistics obtained on permutations, and the smallest $p$-value $p^*$ is selected. To ensure that

the type I error rate is properly controlled, the same selection procedure is applied to the statistics obtained on permutations; the smallest $p$-value obtained on the $k$-th permutation if denoted $p^*_{(k)}$. The global $p$-value is finally defined as:

$$p = \frac{1}{K} \sum_{k=1}^{K} \mathbf{1}_{\{p^*_{(k)} \leq p^*\}},$$

where $K$ is the number of permutations. Hereafter, this procedure is named MGF-R (Moment Generating Function - Ratio) test, the proposed method being based on a distance defined as a difference between cumulant generating functions or, equivalently, as the logarithm of a ratio of moment generating functions.

# 4 Comparison of detection methods for association signals

## 4.1 Simulation study

In this section, the performance of our MGF-R procedure is investigated in terms of control of the type I error rate and statistical power for a large scope of simulation scenarios, either data-driven in various GWAS designs or under typical models for dependence. Each data-driven scenario corresponds to a block of SNPs and is characterized by both the marginal distributions of the SNPs and the dependence structure between the SNPs. For a given SNP-set, phenotypes are simulated based on various association models. For each setup, type I error rate and power are estimated on 1,000 simulated datasets, each one made of 1,000 controls and 1,000 cases. The set of individuals, cases and controls, is represented by a matrix of $p$-dimensional profiles of genotype variables generated using the R package `GenOrd` (Barbiero and Ferrari, 2015a).

### 4.1.1 Control of the type I error rate

To study the control of the type I error rate by the MGF-R test, two types of situations have been investigated, either based on equicorrelation and autocorrelation

models for dependence, parameterized by the correlation coefficient $\rho$, or corresponding to more realistic dependence structures observed in a real dataset. In each scenario, the phenotype, independent from the genotype profiles, is randomly assigned to each individual.

For the equicorrelation and autocorrelation structures, the number of SNPs $p$ is set to 20, 50 and 100 and $\rho$ to 0.2, 0.5 and 0.8. The marginal distributions are obtained assuming that SNPs are in Hardy-Weinberg equilibrium with minor allele frequency set to 0.4 for each SNP. The empirical type I error rate is then estimated on 1,000 simulations for different values of the nominal level $\alpha$. The type I error rate is therefore shown to be properly controlled, all empirical values being close to the theoretical level $\alpha$ (see Table 2.2). For data-driven simulations, a similar procedure is applied on three blocks of SNPs corresponding to genes PDZRN4 (68 SNPs), DTD1 (49 SNPs) and KCNN3 (37 SNPs) observed in the WTCCC dataset (Wellcome Trust Case Control Consortium, 2007). The within-gene dependence structure of these genes are displayed on Figure 2.3. These genes show different patterns of dependence: DTD1 has the strongest dependence structure, KCNN3 the weakest and PDZRN4 has an intermediate position. The empirical type I error rates are given in table 2.2 for different values of the nominal level $\alpha$. The empirical level is always close to its theoretical value; the type I error rate is accurately controlled under these three realistic scenarios. To sum up, it is demonstrated that the MGF-R test properly controls the type I error rate for several values $\alpha$ and various patterns of dependence.

### 4.1.2 Power study

The aim of this paragraph is to compare the power of MGF-R with the seven existing methods introduced in Section 2.2, namely minP, $L^2$-norm, Hotelling, SKAT, HC, GHC and the omnibus test. We focus here on the patterns of dependence observed within SNP-sets in the three genes PDZRN4, DTD1 and KCNN3 introduced in the previous paragraph. To simulate scenarios under the alternative hypothesis, we used a similar approach as that introduced in Section 2.3, based on model (2.1) without covariates. For that purpose, a small number of causal SNPs are chosen by setting the corresponding coordinates of $\boldsymbol{\beta}$ to nonzero values. The maximal signal strength

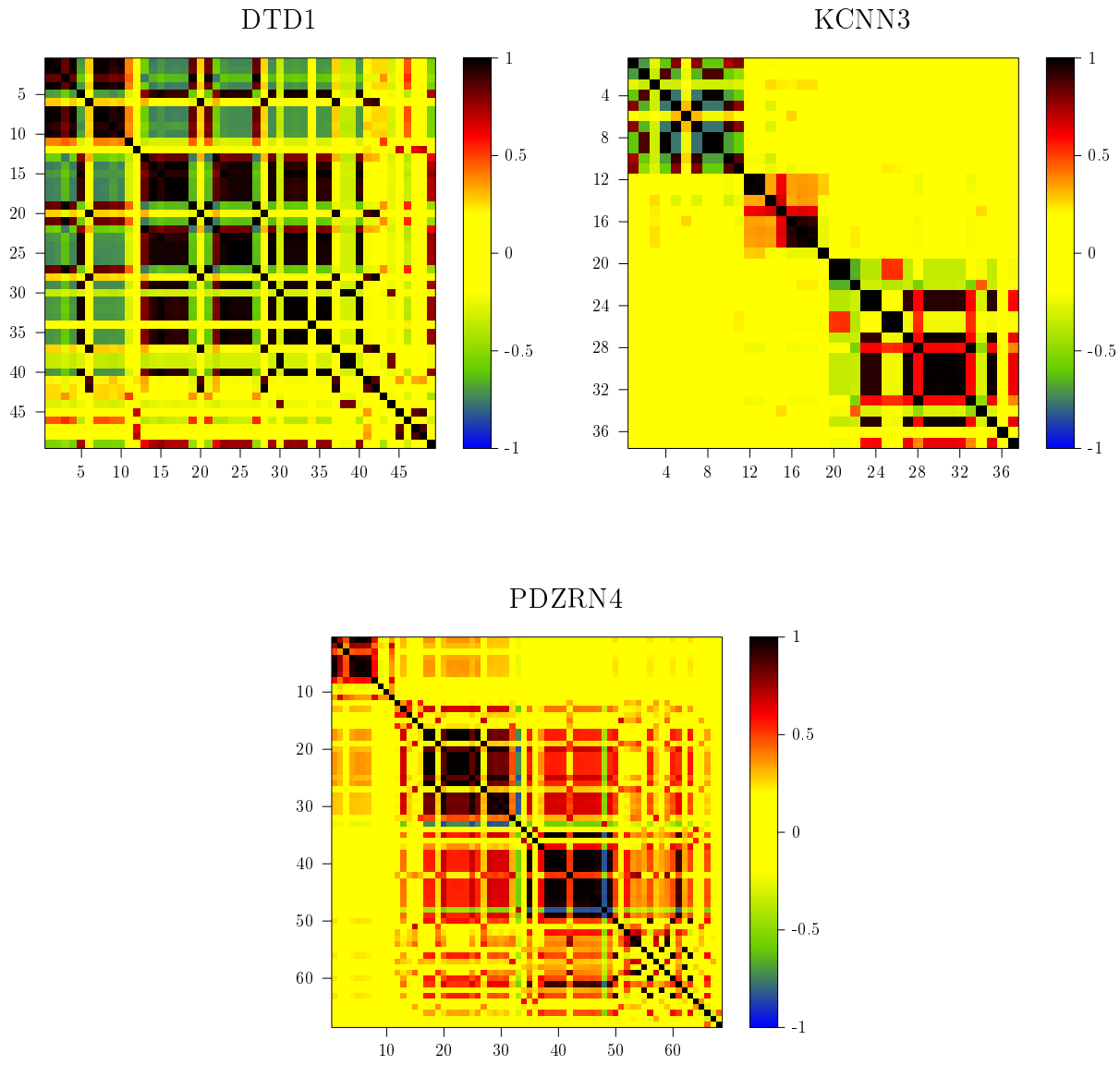


Figure 2.3: Dependence structures of the DTD1, KCNN3 and PDZRN4 genes

Table 2.2: Empirical type I error rates of the MGF-R test for different values of the nominal level $\alpha$, under different dependence structures (upper table: equicorrelation, lower table: autocorrelation).

| $\rho$ | 0.2 | | | 0.5 | | | 0.8 | | |
|---|---|---|---|---|---|---|---|---|---|
| $\alpha$ | 0.01 | 0.05 | 0.1 | 0.01 | 0.05 | 0.1 | 0.01 | 0.05 | 0.1 |
| $p = 20$ | 0.012 | 0.046 | 0.1 | 0.011 | 0.048 | 0.097 | 0.01 | 0.048 | 0.102 |
| $p = 50$ | 0.012 | 0.043 | 0.101 | 0.012 | 0.054 | 0.112 | 0.013 | 0.051 | 0.093 |
| $p = 100$ | 0.007 | 0.054 | 0.102 | 0.013 | 0.055 | 0.098 | 0.015 | 0.05 | 0.109 |

| $\rho$ | 0.2 | | | 0.5 | | | 0.8 | | |
|---|---|---|---|---|---|---|---|---|---|
| $\alpha$ | 0.01 | 0.05 | 0.1 | 0.01 | 0.05 | 0.1 | 0.01 | 0.05 | 0.1 |
| $p = 20$ | 0.008 | 0.054 | 0.097 | 0.015 | 0.055 | 0.103 | 0.01 | 0.047 | 0.095 |
| $p = 50$ | 0.01 | 0.046 | 0.098 | 0.014 | 0.05 | 0.099 | 0.007 | 0.049 | 0.098 |
| $p = 100$ | 0.013 | 0.053 | 0.095 | 0.013 | 0.057 | 0.1 | 0.015 | 0.047 | 0.112 |

Table 2.3: Empirical type I error rates of the MGF-R test for different values of the nominal level $\alpha$, for three different genes from WTCCC data (Wellcome Trust Case Control Consortium, 2007).

| $\alpha$ | 0.01 | 0.05 | 0.1 | 0.2 |
|---|---|---|---|---|
| PDZRN4 | 0.013 | 0.049 | 0.104 | 0.185 |
| DTD1 | 0.011 | 0.055 | 0.107 | 0.207 |
| KCNN3 | 0.015 | 0.050 | 0.098 | 0.207 |

is obtained for $\boldsymbol{\beta}_{\mathbf{max}}$ chosen to achieve a detection rate close to one and intermediate signal strengths are given by $\boldsymbol{\beta} = \xi\boldsymbol{\beta}_{\mathbf{max}}$ with $\xi \in \{0.1, 0.2, \ldots, 0.9, 1\}$.

Estimated power curves are displayed in Figure 2.4, with values of the non-zero coefficients above each subfigure. Results shows that MGF-R is the only method with an acceptable power in all scenarios. MGF-R is indeed close to the most powerful method in scenarios $\ell_1, \ell_2, \ell_3$ in Figure 2.4 and is even the best method in scenarios $r_1$, $r_2$ and $r_3$. In all scenarios, either the $L^2$-norm or Hotelling has high power thus confirming that the class of tests $T_2(\boldsymbol{\mathcal{Z}}^*, \boldsymbol{h})$ (see Equation (2.5))

covers a sufficiently large scope of test statistics. All the other methods, including the omnibus test, have low power in at least one scenario. It is worth noting that minP and GHC are closely related in all scenarios, which weakens the advantage of combining them in an omnibus test. Furthermore, the benefit of using GHC instead of HC is not clear in our results. To summarize, our results show a large inconsistency in the ranking of the global tests and MGF-R is undoubtedly the most robust method to variations in the dependence structure and in the pattern of the association signal.

## 4.2 Application to real datasets

The MGF-R test is now applied to two GWAS datasets in order to investigate its ability to detect biologically validated genes. For each dataset, the MGF-R test is compared to the alternative methods introduced in Section 2.2. It is demonstrated that the MGF-R test can successfully detect genes known to be associated to the phenotype, whereas other tests cannot.

### 4.2.1 Dogs dataset

The first dataset consists of 28 dogs for which 5615 genes were sequenced to investigate the genetic background of the furnishing phenotype, which is a characteristic pattern of a moustache and eyebrows (Broeckx et al., 2017). Among the 28 dogs, 16 are standard poodle, thus having a furnishing trait, while 12 are not furnished. Due to the small sample size, the correlation matrix of the test statistics vector may not be positive definite. When needed, it is replaced by the nearest positive definite correlation matrix, computed with the `nearPD` function of the R package `Matrix` (Bates and Maechler, 2018). To ensure this does not affect the control of the type I error rate, the procedure described in Section 4.1 has been performed on genes having a number of SNPs much larger than the sample size. Even for genes of more than 100 SNPs, the type I error rate is controlled.

According to Broeckx et al. (2017), the RSPO2 gene in chromosome 13 (119 SNPs) is strongly associated to the furnishing phenotype. This gene is correctly

detected by the MGF-R test and the $L^2$-norm test with $p$-values both lower than $10^{-6}$ after a Bonferroni correction. However minP, HC, GHC, SKAT, Omnibus and Hotelling's tests fail at detecting RSPO2 with corrected $p$-values much greater than 0.05. The corrected $p$-values are given in Table 2.4.

Table 2.4: Bonferroni corrected $p$-values for gene RSPO2 (dogs dataset)

| minP | $L^2$-norm | Hotelling | SKAT | HC | GHC | Omnibus | MGF-R |
|---|---|---|---|---|---|---|---|
| 0.48 | $< 10^{-6}$ | 1 | 1 | 1 | 0.44 | 1 | $< 10^{-6}$ |

### 4.2.2 Crohn's disease dataset

The second application to a real GWAS dataset focuses on the study of Crohn's disease, a type of chronic inflammatory bowel disease for which genetic factors are known to play a major role. In de Lange et al. (2017), a list of genes known to be involved in Crohn's disease is given. All compared methods are applied to the 73 genes of this list recovered in the WTCCC dataset (Wellcome Trust Case Control Consortium, 2007).

In total, 8 genes have been significantly detected by at least one method and the corresponding $p$-values after a Bonferroni correction are given in table 2.5 ($p$-values lower than 0.05 are in bold). The MGF-R test is the only method able to detect the 8 reported signals. Regarding the other methods, it can be underlined that the $L^2$-norm test, together with HC and GHC, is able to correctly detect genes DENND1B and MST1 while the Hotelling's t-square test fails at reporting them as significant. Conversely, genes SCAMP3, ZMIZ1 and C10orf55 are detected by the Hotelling's t-square test and not by other methods. These results show the flexibility of the MGF-R test, that has the power to detect a wider range of signals than existing methods. Such flexibility is further enhanced by the capacity for the MGF-R test to correctly detect CARD9, a gene not found by any other method. Our results further demonstrate that MGF-R, which optimizes weights in a broad class of statistics, is obviously more robust than an omnibus strategy limited to the combination of the methods.

# 5 Discussion

The choice of a global testing procedure remains a challenging issue, especially when it is to be used in a wide variety of dependence and association signal patterns, such as for SNP-set association testing for GWAS. Although several statistical procedures consisting in aggregating pointwise test statistics have been proposed, none of them outperforms the others for every possible conjunction of a dependence structure and a distribution of the non-null features of the association signal.

Based on these observations, we proposed a global test statistic defined as an adaptively weighted sum of squared decorrelated pointwise test statistics. Adaptively choosing the weights provides a flexibility that ensures to maintain satisfying power in most situations. Based on simulations, using realistic correlation structures, the results demonstrate the advantage of using MGF-R compared to existing methods. As shown in Section 4, MGF-R is the only method that reaches acceptable power in all situations. Furthermore, MGF-R is shown to be much more robust to various situations than an *ad-hoc* omnibus strategy.

The application to two GWAS datasets confirms the good performance of MGF-R compared to existing methods. Indeed, the flexibility of MGF-R allows the identification of validated genes in very diverse situations while other methods are limited to detect associated genes in specific situations. The greater robustness of MGF-R observed under several simulation designs is thus confirmed by real datasets, in which the signal and dependence configurations are much more diverse than those usually fixed in simulation studies.

Table 2.5: Corrected $p$-values corresponding to genes detected by at least one test (Crohn's disease dataset)

| Gene | Chr | $p$ | minP | $L^2$-norm | HC | GHC | Hotelling | SKAT | Omnibus | MGF-R |
|---|---|---|---|---|---|---|---|---|---|---|
| DENND1B | 1 | 41 | 0.062 | **0.007** | **0.008** | **0.007** | 0.166 | 1 | **0.014** | **0.022** |
| SCAMP3 | 1 | 8 | 0.651 | 1 | 0.633 | 0.679 | **0.019** | 1 | 0.226 | **0.033** |
| MST1 | 3 | 10 | **0.036** | **0.006** | **0.029** | **0.023** | 0.208 | 1 | **0.007** | **0.020** |
| CARD9 | 9 | 7 | 0.123 | 0.369 | 0.191 | 0.185 | 0.494 | 1 | 0.387 | **0.033** |
| ZMIZ1 | 10 | 48 | 0.443 | 0.289 | 0.137 | 0.1172 | **< 0.001** | 1 | 0.081 | **< 0.001** |
| C10orf55 | 10 | 6 | 1 | 1 | 1 | 1 | **0.008** | 1 | 1 | **0.024** |
| PTPN2 | 18 | 22 | **< 0.001** | **< 0.001** | **< 0.001** | **< 0.001** | **0.021** | 1 | **0.014** | **< 0.001** |
| SBNO2 | 19 | 9 | **0.003** | **< 0.001** | **0.003** | **0.011** | **0.007** | 1 | **0.038** | **0.004** |

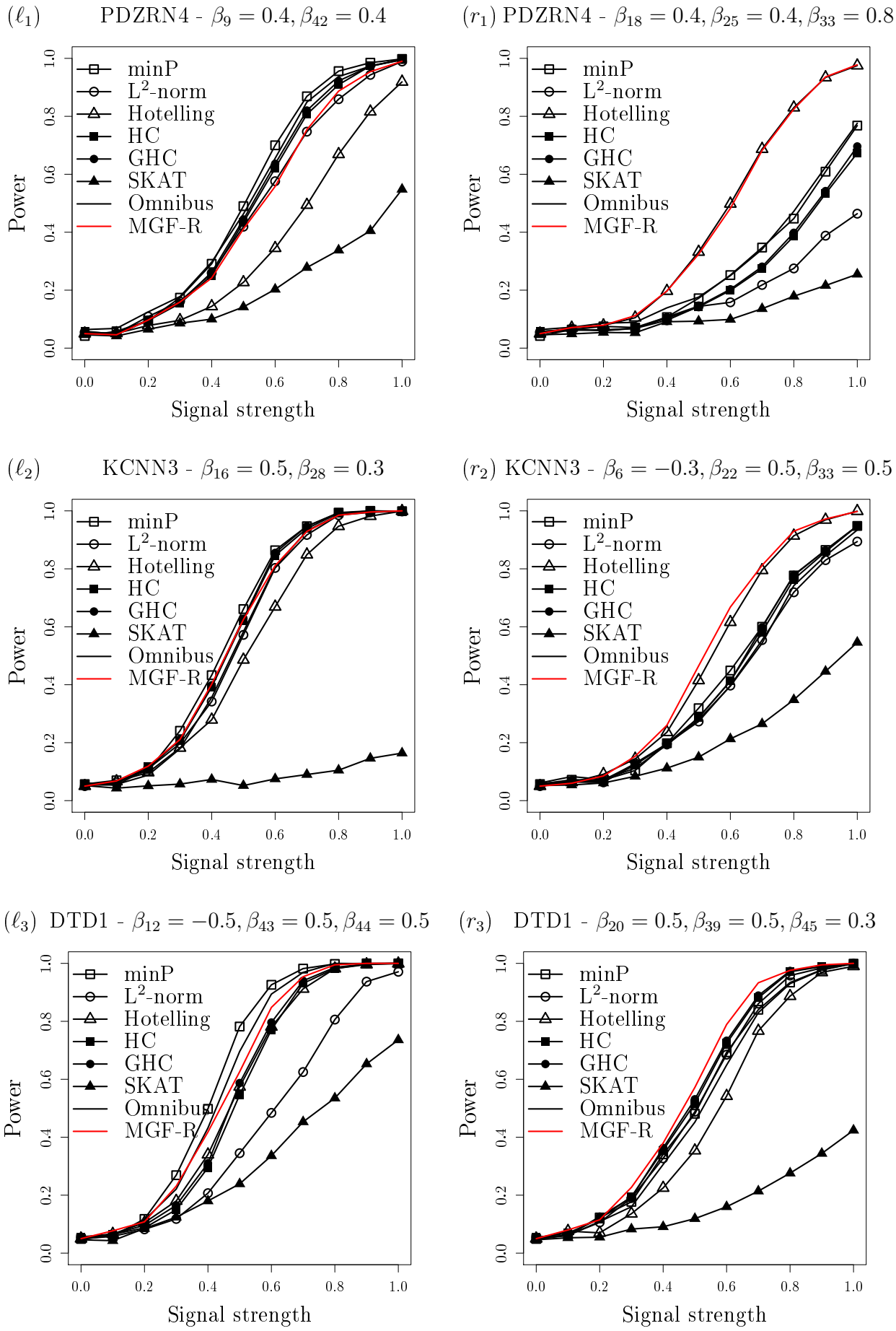


Figure 2.4: Power curves for several scenarios on genes PDZRN4, KCNN3 and DTD1

# 6 Appendix: proof of Theorem 3.1

We first give the main steps of the calculations giving the explicit expression of the difference between cumulant generating functions (equation 2.6). First, the cumulant generating function (CGF) of a random variable distributed according to the $\chi^2_k(\gamma)$ distribution is defined as:

$$\kappa(t) = \frac{\gamma t}{1-2t} - \frac{k}{2}\log(1-2t).$$

Moreover, let $\kappa_X$ be the CGF of a random variable $X$. For any scalar $a \neq 0$, if $\kappa_{aX}$ is the CGF of $aX$, then $\kappa_{aX}(t) = \kappa_X(at)$. Finally, let $X_i, 1 \leq i \leq p$ be $p$ mutually independent random variables, let $\kappa_i$ be the CGF of $X_i$ and let $S = \sum_{i=1}^{p} X_i$. The CGF of $S$ is then defined as $\kappa_S(t) = \sum_{i=1}^{p} \kappa_i(t)$. Using these properties, it can be directly shown that the CGF of a random variable distributed as $\sum_{i=1}^{p} h_i\lambda_i\chi^2_1(\gamma_i)$ (where the $\chi^2_1$ are mutually independent) is:

$$\kappa_{\boldsymbol{\gamma}}(t) = \sum_{i=1}^{p} \frac{h_i\lambda_i\gamma_i t}{1-2th_i\lambda_i} - \frac{1}{2}\sum_{i=1}^{p}\log(1-2h_i\lambda_i t)$$

and if $\gamma_i = 0\ \forall i$,

$$\kappa_{\mathbf{0}}(t) = -\frac{1}{2}\sum_{i=1}^{p}\log(1-2h_i\lambda_i t),$$

hence the result.

The difference between cumulant generating functions is defined as:

$$\mathrm{CGF}(t, \boldsymbol{h}, \boldsymbol{\lambda}, \boldsymbol{\gamma}) - \mathrm{CGF}(t, \boldsymbol{h}, \boldsymbol{\lambda}, \mathbf{0}) \quad = \quad t\sum_{i=1}^{p} \frac{h_i\lambda_i\gamma_i}{1-2th_i\lambda_i}.$$

The optimal coefficients $h_i^*$ are the solution to the following problem:

$$\max_{h_i} t\sum_{i=1}^{p} \frac{h_i\lambda_i\gamma_i}{1-2th_i\lambda_i} \quad \text{under constraint} \quad \sum_{i=1}^{p} h_i = p.$$

The Lagrangian of this problem is:

$$\mathcal{L}(h, \alpha) = t\sum_{i=1}^{p} \frac{h_i\lambda_i\gamma_i}{1-2th_i\lambda_i} - \alpha\left(\sum_{i=1}^{p} h_i - p\right).$$

Deriving the partial derivatives (with respect to each $h_k$ and to $\alpha$) and equating them to 0 yields the following system of $p+1$ equations:

$$\begin{cases} \dfrac{t\lambda_k\gamma_k}{(1-2th_k\lambda_k)^2} - \alpha = 0 \quad \forall k \in \{1,\dots,p\} \\ \displaystyle\sum_{i=1}^{p} h_i - p = 0. \end{cases}$$

The last equation directly gives the constraint $\displaystyle\sum_{i=1}^{p} h_i = p$. The first $p$ equations give:

$$t\lambda_k\gamma_k = \alpha(1-2th_k\lambda_k)^2.$$

Since the difference of cumulant generating functions is defined for $t \leq 1/(2\max_{1\leq i\leq p} h_i\lambda_i)$, $1-2th_k\lambda_k \geq 0\ \forall k$. Then,

$$\begin{aligned} & t\lambda_k\gamma_k &=& \alpha(1-2th_k\lambda_k)^2 \\ \Leftrightarrow\quad & \sqrt{\frac{t\lambda_k\gamma_k}{\alpha}} &=& 1-2th_k\lambda_k \\ \Leftrightarrow\quad & h_k &=& \frac{1}{2t\lambda_k} - \frac{\operatorname{sign}(t)}{2\sqrt{t\alpha}}\sqrt{\frac{\gamma_k}{\lambda_k}}. \end{aligned}$$

Replacing the $h_k$ by this last expression in $\displaystyle\sum_{k=1}^{p} h_k = p$ and solving for $\alpha$ yields

$$\alpha = \frac{\left(\displaystyle\sum_{k=1}^{p}\sqrt{\frac{\gamma_k}{\lambda_k}}\right)^2}{4t\left(p - \dfrac{1}{2t}\displaystyle\sum_{k=1}^{p}\frac{1}{\lambda_k}\right)^2}$$

and thus

$$h_i = \frac{1}{2t\lambda_i} - \operatorname{sign}(t)\left|p - \frac{1}{2t}\sum_{k=1}^{p}\frac{1}{\lambda_k}\right| \frac{\sqrt{\dfrac{\gamma_i}{\lambda_i}}}{\displaystyle\sum_{k=1}^{p}\sqrt{\frac{\gamma_k}{\lambda_k}}}.$$

# Chapter 3

# Detection of Interaction Effects Between a Gene and an Environmental Factor

**Abstract.** The development of genetic association studies and the numerous discoveries that were allowed by these led to the emergence of new problematics. Indeed, genetic regions highlighted by these studies only partially explain the studies phenotypes. It is now commonly accepted that other factors must be considered to improve the understanding of complex diseases. In particular, interaction effects are thought to count for an important part of the so-called missing heritability. This chapter namely aims at discussing the impact of dependence on the performance of testing procedures for the detection of interaction effects between a gene and an environmental factor. Moreover, the MGF-R test proposed in the previous chapter will be extended to this more challenging problem using an appropriate model and a suitable resampling strategy to guarantee the control of the type I error rate. We show through extensive simulations that the MGF-R test maintains its robustness to dependence and its ability to detect a wide range of effects, where other tests may fail.

## Contents

**1 Introduction . . . . . 79**

**2 A general framework for the detection of interaction effects between a gene and an environmental factor . . . 81**

2.1 Definition of a general framework adapted to the detection of interaction effects . . . . . 81

2.2 Estimation of the distribution of a global test statistic under the null hypothesis . . . . . 83

2.3 Existing methods specifically designed for the detection of gene - environmental factor interaction effects . . . . . 87

**3 Assessment of the MGF-R test for the detection of gene - environmental factor interaction effects . . . . . 89**

3.1 Control of the type I error rate . . . . . 89

3.2 Power study . . . . . 89

**4 Discussion . . . . . 94**

# 1 Introduction

The generalized linear model (2.1) considered in Chapter 2 can be suitably modified for various types of applications. Among these applications, the interest for detection of interaction effects in genomics has importantly grown over last years. In this chapter, we focus on the detection of interaction effects between a gene and an environmental factor.

Gene-phenotype association studies allowed the discovery of numerous genetic regions significantly associated to various phenotypes. Nevertheless, these associations explain only very partially the studied phenotypes; this is particularly true for complex phenotypes (Manolio et al., 2009). It is now generally admitted that other effects have to be considered to improve the knowledge and understanding of these phenotypes, namely interaction effects between genetic and environmental factors (Thomas, 2010; Cordell, 2009b; Hunter, 2005). For instance, for a given disease, a gene and an environmental factor share an interaction effect if the disease risk for an individual with a given genetic profile is modified only in case of exposure to the environmental factor. A large number of interaction effects were discovered, namely concerning cancers (Clavel, 2007). Some studies also underline genes having a different effect on obesity or depression, depending on gender or the exposure to a particular event (Reddon et al., 2016; Lesch, 2004; Klengel and Binder, 2013). A significant interaction effect is also detected in Lin et al. (2013). The elements highlighted by these studies show the importance that can be represented by these interaction effects for the understanding of complex phenotypes.

Even if the need to take account of interaction effects between genes and environmental factors is recognized, the choice of a method to detect such effects remain an open issue. Several tests were constructed specifically for detecting these interaction effects. These tests are generally defined as (possibly weighted) sums of squares of univariate scores between the phenotype and the interaction profiles (Lin et al., 2013, 2016; Su et al., 2017). However, these tests do not seem to be adapted to take account of dependence between the univariate test statistics. Indeed, the difficulties associated to dependence handling, highlighted in the former chapter, will also be encountered in the context of gene - environmental factor interaction testing. In particular, testing for the presence of genetic effects within a gene (as in

Chapter 2) or testing for the presence of gene - environment interaction effects leads to a vector of test statistics of identical dimension $p$ (where $p$ is the number of SNPs in the gene) and with approximately the same correlation structure. Moreover, it was shown in the previous chapter that depending on the association signal between the phenotype and the SNPs, and depending on its interplay with the correlation structure, the existing methods could yield unstable performances. Without a priori knowledge about the form of the association signal, any testing strategy could be very powerful as well as it could be completely underpowered. It seems that the shape of the association signal and its interplay with the correlation structure will remain central in the gene - environment testing problem.

This chapter is organized as follows. First, modifications are applied to the generalized linear model considered in Chapter 2 (see Conneely and Boehnke (2007)) to obtain a suitable model for the detection of gene - environment interaction effects. This model is also considered in Lin et al. (2013, 2016); Su et al. (2017). Similarly to Chapter 2, a test statistics vector is then constructed, on which the former global testing methods (minP, $L^2$-norm test, higher criticism, Hotelling's test and the MGF-R test) can be applied. However, a major issue is encountered for the estimation of the distribution of a global test statistic under the null hypothesis. Indeed, it is shown that random permutations of the phenotype, as used previously, are not suitable for the estimation of the null distribution (Bůžková et al., 2011; Coombes and Biernacka, 2018; Dudbridge and Fletcher, 2014; Epstein et al., 2012). Based on Bůžková et al. (2011), a parametric bootstrap procedure is proposed to remedy this problem. Then, in Section 3, a simulation-based comparative study is presented. It aims first at checking that in diverse realistic contexts, the type I error rate is rightfully controlled by the MGF-R test, using the former parametric bootstrap procedure. Then, the power of the MGF-R test for the detection of gene - environment effects will be evaluated and compared to that of existing methods. In particular, it will be compared to usual tests introduced in Chapter 2, but also to the tests that were specifically introduced for the detection of interaction effects. It will namely be demonstrated on simulations that the robustness and flexibility of the MGF-R test are maintained. On the other hand, the performances of the usual tests as well as that of the tests specifically developed for the detection of interaction effects will be shown to be unstable from a situation to another.

# 2 A general framework for the detection of interaction effects between a gene and an environmental factor

## 2.1 Definition of a general framework adapted to the detection of interaction effects

In the former chapter, in which the focus was on the detection of candidate genes possibly associated to a phenotype, the generalized linear model (2.1) was considered. This model is written as:

$$h(\mathbb{E}[Y|\boldsymbol{U}=\boldsymbol{u},\boldsymbol{X}=\boldsymbol{x}])=\boldsymbol{u}'\boldsymbol{\alpha}+\boldsymbol{x}'\boldsymbol{\beta}$$

where $\boldsymbol{x}$ is a genotypic profile of dimension $p$ and $\boldsymbol{u}$ an optional profile of covariates of dimension $q$. Natural modifications can be applied to this model to adapt it to the detection of gene - environmental factor interaction effects. Indeed, let $E$ be the variable modeling an environmental factor of interest. For simplicity, let us suppose that the environmental factor is a binary variable (exposure or not to a condition) or that the corresponding effect can be modeled by a linear term. For a genotypic profile $\boldsymbol{x}$, the corresponding interaction profile is denoted $\boldsymbol{s}=(s_1,\ldots,s_p)'$, with $s_j=e\cdot x_j$, where $e$ is the value taken by the environmental variable. The following generalized linear model can then be considered:

$$h(\mathbb{E}[Y|\boldsymbol{U}=\boldsymbol{u},\boldsymbol{X}=\boldsymbol{x},E=e])=\boldsymbol{u}'\boldsymbol{\beta_U}+\boldsymbol{x}'\boldsymbol{\beta_X}+e\beta_E+\boldsymbol{s}'\boldsymbol{\beta_S}.$$

Denoting $\boldsymbol{W}=(\boldsymbol{U}',\boldsymbol{X}',E)'$, $\boldsymbol{\alpha}=(\boldsymbol{\beta}_{\boldsymbol{U}}',\boldsymbol{\beta}_{\boldsymbol{X}}',\beta_E)'$ and $\boldsymbol{\beta}=\boldsymbol{\beta_S}$, the model can be rewritten as:

$$h(\mathbb{E}[Y|\boldsymbol{U}=\boldsymbol{u},\boldsymbol{X}=\boldsymbol{x},E=e])=\boldsymbol{w}'\boldsymbol{\alpha}+\boldsymbol{s}'\boldsymbol{\beta}, \tag{3.1}$$

which enables us to write the model under a similar form to that of model (2.1). This model is also considered by several authors (Lin et al., 2013, 2016; Su et al., 2017). Consequently, testing the presence of at least one significant interaction effect between the considered gene and the environmental factor is equivalent to testing:

$$\begin{cases} H_0: & \boldsymbol{\beta}=\boldsymbol{0} \\ H_1: & \boldsymbol{\beta}\neq\boldsymbol{0}. \end{cases}$$

In the following, we denote $\mathbb{W}$ the $n \times (p+q+1)$ matrix whose $i$-th row is the vector obtained by concatenating the covariates profile, the genotypic profile and the value of the environmental variable for the $i$-th individual:

$$\mathbb{W} = \left(\begin{array}{ccc|ccc|c} u_{11} & \dots & u_{1q} & x_{11} & \dots & x_{1p} & e_1 \\ \vdots & & \vdots & \vdots & & \vdots & \vdots \\ u_{i1} & \dots & u_{iq} & x_{i1} & \dots & x_{ip} & e_i \\ \vdots & & \vdots & \vdots & & \vdots & \vdots \\ u_{n1} & \dots & u_{nq} & x_{n1} & \dots & x_{np} & e_n \end{array}\right)$$

$$= [\mathbb{U} \mid \mathbb{X} \mid \boldsymbol{E}]$$

where $u_{ij}$ is the value taken by the $i$-th individual for the $j$-th covariable, $x_{ij}$ is the value taken by the $i$-th individual for the $j$-th SNP and $e_i$ is the value of the $i$-th individual for the environmental factor. The first covariable corresponds to the intercept of the model and thus has all its values equal to 1. Similarly, we denote $\mathbb{S}$ the $n \times p$ matrix whose $i$-th row is the interaction profile of the $i$-th individual:

$$\mathbb{S} = \begin{pmatrix} e_1 \cdot x_{11} & \dots & e_1 \cdot x_{1j} & \dots & e_1 \cdot x_{1p} \\ \vdots & & \vdots & & \vdots \\ e_i \cdot x_{i1} & \dots & e_i \cdot x_{ij} & \dots & e_i \cdot x_{ip} \\ \vdots & & \vdots & & \vdots \\ e_n \cdot x_{n1} & \dots & e_n \cdot x_{nj} & \dots & e_n \cdot x_{np} \end{pmatrix}.$$

Finally, we denote $\boldsymbol{S_j}$ the $j$-th column of $\mathbb{S}$. Similarly as in Chapter 2, a vector of individual test statistics $\boldsymbol{\mathcal{Z}} = (Z_1, \dots, Z_p)'$ is considered, with

$$Z_j = \frac{\boldsymbol{S_j'}(\boldsymbol{y} - \hat{\boldsymbol{y}}_{\boldsymbol{0}})}{\sqrt{\widehat{\boldsymbol{\Gamma}}_{j,j}}}$$

where $\hat{\boldsymbol{y}}_{\boldsymbol{0}}$ is the $n$-vector whose $i$-th coordinate equals $h^{-1}(\boldsymbol{w_i'}\hat{\boldsymbol{\alpha}}_{\boldsymbol{0}})$ and $\hat{\boldsymbol{\alpha}}_{\boldsymbol{0}}$ is the estimator of $\boldsymbol{\alpha}$ under $H_0$, and

$$\widehat{\boldsymbol{\Gamma}} = \hat{\sigma}^2_Y(\mathbb{S}'\mathbb{S} - \mathbb{S}'\mathbb{W}(\mathbb{W}'\mathbb{W})^{-1}\mathbb{W}'\mathbb{S}).$$

Let us note that $\boldsymbol{\mathcal{Z}}$ can also be rewritten as:

$$\boldsymbol{\mathcal{Z}} = \mathrm{diag}(\widehat{\boldsymbol{\Gamma}})^{-1/2}\mathbb{S}'(\boldsymbol{y} - \hat{\boldsymbol{y}}_{\boldsymbol{0}}).$$

Analogously to the former chapter, it seems then natural to apply usual tests (namely minP, the $L^2$-norm test, the higher criticism and Hotelling's test) as well as the MGF-R test on the vector $\boldsymbol{\mathcal{Z}}$ and the associated correlation matrix $\widehat{\boldsymbol{\Sigma}} = \text{diag}(\widehat{\boldsymbol{\Gamma}})^{-1/2}\widehat{\boldsymbol{\Gamma}}\,\text{diag}(\widehat{\boldsymbol{\Gamma}})^{-1/2}$.

## 2.2 Estimation of the distribution of a global test statistic under the null hypothesis

### 2.2.1 Estimation of the distribution by parametric bootstrap

We propose to construct a procedure dedicated to the detection of interaction effects between a gene and an environmental factor on a phenotype. This procedure is based on model (3.1) and on the construction of a global test statistic $T(\boldsymbol{\mathcal{Z}})$ obtained by aggregating the coordinates of $\boldsymbol{\mathcal{Z}}$. As previously, it will be mandatory to estimate the distribution of $T(\boldsymbol{\mathcal{Z}})$ under the null hypothesis. Under $H_0$, $\boldsymbol{\beta_S} = \mathbf{0}$. The model under $H_0$ can thus be written as:

$$h(\mathbb{E}[Y|\boldsymbol{U} = \boldsymbol{u}, \boldsymbol{X} = \boldsymbol{x}, E = e]) = \boldsymbol{u}'\boldsymbol{\beta_U} + \boldsymbol{x}'\boldsymbol{\beta_X} + e\boldsymbol{\beta_E}.$$

Consequently, it is underlined here that it will not be possible to estimate the distribution of the test statistic under $H_0$ by using simple random permutations of the vector $\boldsymbol{y}$ containing the values of the phenotype variable. Indeed, a simple permutation of $\boldsymbol{y}$ amounts to considering the hypothesis of simultaneous nullity of all parameters of model (3.1). This hypothesis is more restrictive than $H_0$. In presence of non-negligible marginal effects, using simple permutations would lead to a wrong estimation of the null distribution of $T(\boldsymbol{\mathcal{Z}})$. This point is detailed in Bůžková et al. (2011); Coombes and Biernacka (2018); Dudbridge and Fletcher (2014). It is namely shown on simulations that the control of the type I error rate can be very much affected by ignoring marginal effects. To remedy this problem, a parametric bootstrap procedure based on the estimation of the parameters of the model under the null hypothesis is proposed in Bůžková et al. (2011) to generate phenotypes. Similarly, a biased permutation procedure is proposed in Epstein et al. (2012) to take account of confounding factors, which can be considered as an analogous problem to that considered here.

We propose to adapt the parametric bootstrap procedure of Bůžková et al. (2011) to the present context. Let $\hat{\boldsymbol{\beta}}_{\boldsymbol{U}}^{(\mathbf{0})}$, $\hat{\boldsymbol{\beta}}_{\boldsymbol{X}}^{(\mathbf{0})}$, and $\hat{\beta}_E^{(0)}$ be the estimators of $\boldsymbol{\beta_U}$, $\boldsymbol{\beta_X}$ and $\beta_E$ under the null hypothesis. A phenotype is then generated by randomly sampling according to the distribution of $Y$ under the considered model. For a binary phenotype, the $i$-th coordinate of the simulated phenotype is obtained by randomly sampling according to the $\mathcal{B}(\hat{y}_{0,i})$ distribution, where $\hat{y}_{0,i}$ is the estimated probability for the $i$-th individual to take the value 1 under $H_0$:

$$\hat{y}_{0,i} = \frac{\exp\left(\boldsymbol{u_i'}\hat{\boldsymbol{\beta}}_{\boldsymbol{U}}^{(\mathbf{0})} + \boldsymbol{x_i'}\hat{\boldsymbol{\beta}}_{\boldsymbol{X}}^{(\mathbf{0})} + e_i\hat{\beta}_E^{(0)}\right)}{1 + \exp\left(\boldsymbol{u_i'}\hat{\boldsymbol{\beta}}_{\boldsymbol{U}}^{(\mathbf{0})} + \boldsymbol{x_i'}\hat{\boldsymbol{\beta}}_{\boldsymbol{X}}^{(\mathbf{0})} + e_i\hat{\beta}_E^{(0)}\right)}.$$

Thus, marginal effects inducing a dependence between $Y$ and $\boldsymbol{U}$, $\boldsymbol{X}$ and $E$ are maintained for the simulated phenotype. In the following, only binary phenotypes are considered.

### 2.2.2 Estimation of the marginal effects parameters

It is necessary to estimate the marginal genetic effects to compute the vector of test statistics. However, due to the strong dependence among SNPs of a given gene, estimating these effects can be challenging, due to columns almost identical or collinear. To counteract this problem, the marginal genetic effects are estimated by replacing the genetic profiles by the corresponding principal components profiles scaled to unit variance; in $\mathbb{W}$, $\mathbb{X}$ is replaced by the $n \times k$ matrix $\widetilde{\mathbb{X}}$ whose $j$-th column $\tilde{\boldsymbol{X}}_{\boldsymbol{j}}$ is defined as:

$$\tilde{\boldsymbol{X}}_{\boldsymbol{j}} = \frac{\mathbb{X}\boldsymbol{v_j}}{\sqrt{\lambda_j}}$$

where $\boldsymbol{v_j}$ is the $j$-th eigenvector of the correlation matrix estimated between the columns of $\mathbb{X}$, and $\lambda_j$ is the associated eigenvalue. The number of columns $k$ of the matrix $\tilde{\mathbb{X}}$ is the number of retained principal components. In practice, principal component analysis is usually used to reduce dimensionality and obtain only a few components. Numerous criteria exist to determine how many components should be retained. For instance, it is often recommended to retain components associated to an eigenvalue greater than 1 or components which account for more than a given percentage of the total variance of the data, *e.g.* 5%. We aim at obtaining a stable estimation of the marginal effects, not reducing the dimensionality, contrary to these

usual criteria. Using these cut-offs for the present issue could induce a bias in the estimation of the marginal effects, depending on the structure of the considered gene. Consequently, we keep principal components which account for at least 0.1% of the total variance. Despite being very small, this cut-off enables to obtain a stable estimation of the marginal effects.

A possible alternative approach to estimate these marginal effects is to use regularized regression procedures, namely ridge regression (Hoerl and Kennard, 1970); this choice was made by Lin et al. (2013, 2016); Su et al. (2017). However, this raises several issues. First, the value of a regularization parameter $\kappa$ has to be chosen. Usually, it is selected as the value $\hat{\kappa}$ which optimizes a criterion (*e.g.* misclassification rate for a binary response, mean squared error for a continuous response) through a (generalized) cross-validation procedure, which can be computationally cumbersome if permutations are used to estimate the distribution of a global test statistic. Indeed, in a testing procedure such as in the present context, the cross-validation procedure would have to be applied to every shuffled phenotype, since $\hat{\kappa}$ can be viewed as a random variable that depends on $Y$. Consequently, it has to be selected the same way for every shuffled phenotype to ensure the validity of the distribution of the global test statistic under the null hypothesis. To avoid this, Lin et al. (2013, 2016); Su et al. (2017) derived theoretical null distributions of their test statistics. However, these distributions hold asymptotically. For samples of modest size or genes of great dimension, the finite sample distributions can be far from the asymptotic theoretical ones. Finally, regularized regression procedures consist in estimating a biased vector of regression coefficients. If the selected penalty parameter $\hat{\kappa}$ is great, so is the bias. This could affect the type I error rate and power of testing procedures, and would likely lead to unreliable results.

### 2.2.3 Illustration

We propose to use a few simulated examples of various situations under the null hypothesis to illustrate how the type I error rate of a global testing method can be impacted in presence of marginal effects and to demonstrate that the proposed parametric bootstrap procedure effectively restores the control. A population of 100,000 genetic profiles is generated according to the dependence structure observed on gene

PDZRN4 from the WTCCC data (Wellcome Trust Case Control Consortium, 2007) using R package `GenOrd` (R Development Core Team, 2008; Barbiero and Ferrari, 2015a). Then, an environmental factor is generated for the 100,000 individuals independently from the genetic profiles. A binary factor is considered, corresponding to the exposure to a condition. Finally, the phenotype $Y$ is generated according to the following model:

$$\text{logit}(\mathbb{P}[Y = 1|\boldsymbol{X} = \boldsymbol{x}, E = e]) = \beta_0 + \boldsymbol{x}'\boldsymbol{\beta} + e\beta_E,$$

and a sample is constructed by randomly sampling 1,000 cases and 1,000 controls in the generated population. 1,000 samples are generated by using this procedure to estimate the type I error risk in several different scenarios. The following scenarios are considered:

- no marginal effect: $\beta_E = 0$, $\boldsymbol{\beta} = \mathbf{0}$
- environmental effect, no genetic effect: $\beta_E = 2$, $\boldsymbol{\beta} = \mathbf{0}$
- genetic effect, no environmental effect: $\beta_E = 0$, $\boldsymbol{\beta} \neq \mathbf{0}$
- genetic and environmental effects: $\beta_E = 2$, $\boldsymbol{\beta} \neq \mathbf{0}$.

We compare the type I error rates obtained using the minP procedure using simple permutations and using parametric bootstrap for the estimation of the null distribution under each scenario. The results are summarized in Table 3.1. It can be observed that when the environmental factor has a strong marginal effect, the type I error rate using permutations is overcontrolled. Indeed, the observed type I error rate in the situation where $\beta_E = 2$ and $\boldsymbol{\beta} = \mathbf{0}$ is approximately half the nominal level $\alpha$. Such an overcontrol of the type I error rate would most likely lead to a reduced power under the alternative hypothesis and in presence of marginal effects. The type I error rate obtained using permutations does not seem to be impacted when there are only genetic marginal effects. In any situation, the parametric bootstrap procedure successfully restores the control of the type I error rate.

Table 3.1: Type I error rates of the minP procedure under several scenarios using permutations or parametric bootstrap (confidence intervals in brackets)

| $\beta_E$ | $\boldsymbol{\beta}$ | $\alpha$ | Permutations | Parametric bootstrap |
|---|---|---|---|---|
| 0 | $\mathbf{0}$ | 0.01 | 0.011 [0.004, 0.017] | 0.012 [0.005, 0.019] |
| | | 0.05 | 0.047 [0.034, 0.060] | 0.051 [0.037, 0.065] |
| | | 0.1 | 0.097 [0.079, 0.115] | 0.097 [0.079, 0.115] |
| 2 | $\mathbf{0}$ | 0.01 | 0.005 [0.001, 0.009] | 0.006 [0.001, 0.011] |
| | | 0.05 | 0.022 [0.013, 0.031] | 0.048 [0.035, 0.061] |
| | | 0.1 | 0.064 [0.049, 0.079] | 0.095 [0.077, 0.113] |
| 0 | $\beta_{30} = 0.5$, $\beta_{60} = 0.5$ | 0.01 | 0.005 [0.001, 0.009] | 0.007 [0.002, 0.012] |
| | | 0.05 | 0.050 [0.036, 0.064] | 0.054 [0.040, 0.068] |
| | | 0.1 | 0.095 [0.077, 0.113] | 0.098 [0.080, 0.116] |
| 2 | $\beta_{30} = 0.5$, $\beta_{60} = 0.5$ | 0.01 | 0.006 [0.001, 0.011] | 0.007 [0.002, 0.012] |
| | | 0.05 | 0.035 [0.024, 0.046] | 0.049 [0.036, 0.062] |
| | | 0.1 | 0.088 [0.070, 0.106] | 0.111 [0.092, 0.130] |

## 2.3 Existing methods specifically designed for the detection of gene - environmental factor interaction effects

Several methods were specifically designed for the detection of interaction effects between a gene and an environmental factor. These methods are briefly introduced in this section. All of them are based on model (3.1) and on an estimation of the marginal effects by using ridge logistic regression, as stated in Section 2.2.2.

GESAT (Gene - Environment Set Association Test), proposed by Lin et al. (2013), is based on the statistic $(\boldsymbol{y} - \hat{\boldsymbol{y}}_{\mathbf{0}})'\mathbb{S}\mathbb{S}'(\boldsymbol{y} - \hat{\boldsymbol{y}}_{\mathbf{0}})$. The $p$-value is calculated using the theoretical distribution of the test statistic, which is that of a linear combination of independent random variables distributed according to a $\chi^2$ distribution with one degree of freedom.

iSKAT (interaction Sequence Kernel Association Test) was proposed by Lin et al. (2016). It is based on statistics defined as $Q_\rho = (\boldsymbol{y} - \hat{\boldsymbol{y}}_{\mathbf{0}})'\mathbb{S}\boldsymbol{W}\boldsymbol{R}_{\boldsymbol{\rho}}\boldsymbol{W}\mathbb{S}'(\boldsymbol{y} - \hat{\boldsymbol{y}}_{\mathbf{0}})$ where $\boldsymbol{R}_{\boldsymbol{\rho}} = \rho\mathbf{1}_p\mathbf{1}_p' + (1-\rho)\mathbf{I}_p$ and $\boldsymbol{W}$ is a diagonal matrix. The $p$-value $p_\rho$ corresponding

to $Q_\rho$ is calculated using the theoretical distribution of $Q_\rho$, which is distributed as a linear combination of independent variables distributed according to a $\chi^2$ distribution with one degree of freedom. A sequence of values between 0 and 1 is considered for the parameter $\rho$, and the statistic $Q = \min_{0\leq\rho\leq 1} p_\rho$ is calculated. The final $p$-value is the $p$-value associated to $Q$.

Several tests named MiSTi (Mixed effects Score Tests for interaction) were proposed by Su et al. (2017). These tests are based on statistics defined as $Q_1 = (\boldsymbol{y} - \hat{\boldsymbol{y}}_{\boldsymbol{0}})'\mathbb{S}\mathbb{S}'(\boldsymbol{y} - \hat{\boldsymbol{y}}_{\boldsymbol{0}})$ and $Q_2 = (\boldsymbol{y} - \hat{\boldsymbol{y}}_{\boldsymbol{0}})'(\mathbb{B} \bullet \boldsymbol{E})\boldsymbol{V}^{-1}(\mathbb{B} \bullet \boldsymbol{E})'(\boldsymbol{y} - \hat{\boldsymbol{y}}_{\boldsymbol{0}})$ where $\boldsymbol{V}$ is the estimated variance-covariance matrix of $(\mathbb{B} \bullet \boldsymbol{E})'(\boldsymbol{y} - \hat{\boldsymbol{y}}_{\boldsymbol{0}})$, $\mathbb{B}$ is a $n \times R$ matrix and $\mathbb{B} \bullet \boldsymbol{E}$ is obtained by computing the element-wise product of each column of $\mathbb{B}$ and $\boldsymbol{E}$, $\boldsymbol{E}$ being the vector of values taken by the environmental factor. More details about the definition of $\mathbb{B}$ are given in Su et al. (2017). $Q_1$ and $Q_2$ can then be combined using several approaches. A first one consists in computing the linear combination $T_\rho = \rho Q_1 + (1-\rho)Q_2$ by choosing the value of $\rho$ (taken on a scale of values between 0 and 1) minimizing the associated $p$-value. Two other approaches are based on a combination of the $p$-values associated to $Q_1$ and $Q_2$, either by Tippett's method (consisting in considering the smallest of the two $p$-values) or by an adaptive combination of the $p$-values. These three approaches are respectively named oMiSTi, aMiSTi and fMiSTi.

The different statistics described in this section are defined as quadratic forms of the vector of univariate interaction scores $\mathbb{S}'(\boldsymbol{y} - \hat{\boldsymbol{y}}_{\boldsymbol{0}})$. Interestingly, the standard $L^2$-norm of the vector $\boldsymbol{\mathcal{Z}}$ previously introduced can be written in the present interaction context as: $(\boldsymbol{y} - \hat{\boldsymbol{y}}_{\boldsymbol{0}})'\mathbb{S}\,\mathrm{diag}(\widehat{\boldsymbol{\Gamma}})^{-1}\mathbb{S}'(\boldsymbol{y} - \hat{\boldsymbol{y}}_{\boldsymbol{0}})$, which is also a quadratic form of $\mathbb{S}'(\boldsymbol{y} - \hat{\boldsymbol{y}}_{\boldsymbol{0}})$. However, the univariate interaction scores are scaled to unit variance in the expression of this statistic, which makes it different from those introduced in this section. In particular, the expression of the GESAT statistic is very similar to that of the former $L^2$-norm statistic. However, they are distinguished by the presence of $\mathrm{diag}(\widehat{\boldsymbol{\Gamma}})^{-1}$ in the expression of the $L^2$-norm statistic. The MGF-R test proposed in Chapter 2 can also be written as a quadratic form of $\mathbb{S}'(\boldsymbol{y} - \hat{\boldsymbol{y}}_{\boldsymbol{0}})$. Indeed, the MGF-R statistic is defined as $\sum_{j=1}^{p} h_j[Z_j^*]^2 = (\boldsymbol{y} - \hat{\boldsymbol{y}}_{\boldsymbol{0}})'\mathbb{S}\,\mathrm{diag}(\widehat{\boldsymbol{\Gamma}})^{-1/2}\boldsymbol{P}\boldsymbol{D}_{\boldsymbol{h}}\boldsymbol{P}'\,\mathrm{diag}(\widehat{\boldsymbol{\Gamma}})^{-1/2}\mathbb{S}'(\boldsymbol{y} - \hat{\boldsymbol{y}}_{\boldsymbol{0}})$ where the $h_j$ are the weights, $\boldsymbol{D}_{\boldsymbol{h}}$ is the diagonal matrix whose diagonal entries are the $h_j$ and $\boldsymbol{P}$ is the matrix of eigenvectors of $\widehat{\boldsymbol{\Sigma}}$. Consequently, it is interesting to extend the MGF-R test to

the present context and to assess its performances for the detection of interaction effects.

# 3 Assessment of the MGF-R test for the detection of gene - environmental factor interaction effects

In this section, we assess the performances of the previously proposed MGF-R test.

## 3.1 Control of the type I error rate

We first check that the type I error rate is rightfully controlled by the MGF-R test using the parametric bootstrap procedure previously described. The same simulation procedure as that used in Section 2.2.3 is used. Results obtained on genes PDZRN4, KCNN3 and DTD1 are presented in tables 3.2, 3.3 and 3.4, for several values of the nominal level $\alpha$. In every situation, the type I error rate is rightfully controlled by the MGF-R test, using the parametric bootstrap procedure previously introduced. This was also observed for the other global tests introduced in Chapter 2 (minP, HC, $L^2$-norm and Hotelling's test). It can be remarked that in presence of marginal effects, other tests can undercontrol or overcontrol the risk. In particular, if there is an environmental effect, the iSKAT test can slightly overcontrol the risk, as well as the GESAT test if there are genetic effects. Moreover, in presence of both genetic and environmental effects, GESAT undercontrols the risk. For greater values of $\beta_E$, an increase of the difference between the nominal level and the observed level of GESAT was observed. Consequently, using this test appears to be unsafe, since the control of the type I error rate is not guaranteed.

## 3.2 Power study

Now, the power of the MGF-R test is assessed in the context of detection of interaction effects between a gene and an environmental factor. The power of MGF-R will be compared to that of minP, the $L^2$-norm test, the higher criticism and Hotelling's

Table 3.2: Estimated type I error rates of the MGF-R test and the tests introduced in Section 2.3 under several scenarios on gene PDZRN4

| $\beta_E$ | $\boldsymbol{\beta}$ | $\alpha$ | GESAT | iSKAT | oMiSTi | aMiSTi | fMiSTi | MGF-R |
|---|---|---|---|---|---|---|---|---|
| 0 | $\mathbf{0}$ | 0.01 | 0.013 [0.006, 0.020] | 0.009 [0.003, 0.015] | 0.012 [0.005, 0.019] | 0.012 [0.005, 0.019] | 0.010 [0.004, 0.016] | 0.010 [0.004, 0.016] |
| | | 0.05 | 0.053 [0.039, 0.067] | 0.046 [0.033, 0.059] | 0.055 [0.041, 0.069] | 0.056 [0.042, 0.070] | 0.052 [0.038, 0.066] | 0.049 [0.036, 0.062] |
| | | 0.1 | 0.108 [0.089, 0.127] | 0.088 [0.070, 0.106] | 0.111 [0.092, 0.130] | 0.113 [0.093, 0.133] | 0.114 [0.094, 0.134] | 0.105 [0.086, 0.124] |
| 2 | $\mathbf{0}$ | 0.01 | 0.008 [0.002, 0.014] | 0.003 [0, 0.006] | 0.010 [0.004, 0.016] | 0.010 [0.004, 0.016] | 0.007 [0.002, 0.012] | 0.011 [0.005, 0.017] |
| | | 0.05 | 0.047 [0.034, 0.060] | 0.029 [0.019, 0.039] | 0.055 [0.041, 0.069] | 0.053 [0.039, 0.067] | 0.063 [0.048, 0.078] | 0.052 [0.038, 0.066] |
| | | 0.1 | 0.083 [0.066, 0.100] | 0.079 [0.062, 0.096] | 0.116 [0.096, 0.136] | 0.120 [0.100, 0.140] | 0.118 [0.098, 0.138] | 0.094 [0.076, 0.112] |
| 0 | $\beta_{30} = 0.5$, $\beta_{60} = 0.5$ | 0.01 | 0.003 [0, 0.006] | 0.009 [0.003, 0.015] | 0.006 [0.001, 0.011] | 0.007 [0.002, 0.012] | 0.008 [0.002, 0.014] | 0.012 [0.005, 0.019] |
| | | 0.05 | 0.026 [0.016, 0.036] | 0.049 [0.036, 0.062] | 0.038 [0.026, 0.050] | 0.041 [0.029, 0.053] | 0.039 [0.027, 0.051] | 0.045 [0.032, 0.058] |
| | | 0.1 | 0.068 [0.052, 0.084] | 0.104 [0.085, 0.123] | 0.082 [0.065, 0.099] | 0.080 [0.063, 0.097] | 0.083 [0.066, 0.100] | 0.084 [0.067, 0.101] |
| 2 | $\beta_{30} = 0.5$, $\beta_{60} = 0.5$ | 0.01 | 0.016 [0.008, 0.024] | 0.012 [0.005, 0.019] | 0.006 [0.001, 0.011] | 0.006 [0.001, 0.011] | 0.005 [0.001, 0.009] | 0.010 [0.004, 0.016] |
| | | 0.05 | 0.076 [0.060, 0.092] | 0.053 [0.039, 0.067] | 0.053 [0.039, 0.067] | 0.050 [0.036, 0.064] | 0.050 [0.036, 0.064] | 0.046 [0.033, 0.059] |
| | | 0.1 | 0.141 [0.119, 0.163] | 0.114 [0.094, 0.134] | 0.105 [0.086, 0.124] | 0.100 [0.081, 0.119] | 0.110 [0.091, 0.129] | 0.111 [0.092, 0.130] |

Table 3.3: Estimated type I error rates of the MGF-R test and the tests introduced in Section 2.3 under several scenarios on gene KCNN3

| $\beta_E$ | $\boldsymbol{\beta}$ | $\alpha$ | GESAT | iSKAT | oMiSTi | aMiSTi | fMiSTi | MGF-R |
|---|---|---|---|---|---|---|---|---|
| 0 | $\mathbf{0}$ | 0.01 | 0.011 [0.005, 0.017] | 0.008 [0.002, 0.014] | 0.007 [0.002, 0.012] | 0.007 [0.002, 0.012] | 0.006 [0.001, 0.011] | 0.009 [0.003, 0.015] |
| | | 0.05 | 0.057 [0.043, 0.071] | 0.043 [0.030, 0.056] | 0.044 [0.031, 0.057] | 0.044 [0.031, 0.057] | 0.053 [0.039, 0.067] | 0.052 [0.038, 0.066] |
| | | 0.1 | 0.112 [0.092, 0.132] | 0.089 [0.071, 0.107] | 0.095 [0.077, 0.113] | 0.093 [0.075, 0.111] | 0.099 [0.080, 0.118] | 0.092 [0.074, 0.110] |
| 2 | $\mathbf{0}$ | 0.01 | 0.003 [0, 0.006] | 0.007 [0.002, 0.012] | 0.010 [0.004, 0.016] | 0.009 [0.003, 0.015] | 0.008 [0.002, 0.014] | 0.011 [0.005, 0.017] |
| | | 0.05 | 0.029 [0.019, 0.039] | 0.025 [0.015, 0.035] | 0.040 [0.028, 0.052] | 0.040 [0.028, 0.052] | 0.044 [0.031, 0.057] | 0.049 [0.036, 0.062] |
| | | 0.1 | 0.065 [0.050, 0.080] | 0.061 [0.046, 0.076] | 0.090 [0.072, 0.108] | 0.088 [0.070, 0.106] | 0.084 [0.067, 0.101] | 0.103 [0.084, 0.122] |
| 0 | $\beta_{10} = 0.5$, $\beta_{30} = 0.5$ | 0.01 | 0.009 [0.003, 0.015] | 0.013 [0.006, 0.020] | 0.011 [0.005, 0.017] | 0.011 [0.005, 0.017] | 0.011 [0.005, 0.017] | 0.012 [0.005, 0.019] |
| | | 0.05 | 0.049 [0.036, 0.062] | 0.045 [0.032, 0.058] | 0.058 [0.044, 0.072] | 0.054 [0.040, 0.068] | 0.052 [0.038, 0.066] | 0.047 [0.034, 0.060] |
| | | 0.1 | 0.097 [0.079, 0.115] | 0.104 [0.085, 0.123] | 0.107 [0.088, 0.126] | 0.111 [0.092, 0.130] | 0.101 [0.082, 0.120] | 0.095 [0.077, 0.113] |
| 2 | $\beta_{10} = 0.5$, $\beta_{30} = 0.5$ | 0.01 | 0.016 [0.008, 0.024] | 0.008 [0.002, 0.014] | 0.010 [0.004, 0.016] | 0.010 [0.004, 0.016] | 0.013 [0.006, 0.020] | 0.013 [0.006, 0.020] |
| | | 0.05 | 0.079 [0.062, 0.096] | 0.037 [0.025, 0.049] | 0.061 [0.046, 0.076] | 0.061 [0.046, 0.076] | 0.051 [0.037, 0.065] | 0.055 [0.041, 0.069] |
| | | 0.1 | 0.135 [0.114, 0.156] | 0.087 [0.070, 0.104] | 0.118 [0.098, 0.138] | 0.118 [0.098, 0.138] | 0.114 [0.094, 0.134] | 0.104 [0.085, 0.123] |

Table 3.4: Estimated type I error rates of the MGF-R test and the tests introduced in Section 2.3 under several scenarios on gene DTD1

| $\beta_E$ | $\boldsymbol{\beta}$ | $\alpha$ | GESAT | iSKAT | oMiSTi | aMiSTi | fMiSTi | MGF-R |
|---|---|---|---|---|---|---|---|---|
| 0 | $\mathbf{0}$ | 0.01 | 0.012 [0.005, 0.019] | 0.014 [0.007, 0.021] | 0.019 [0.011, 0.027] | 0.019 [0.011, 0.027] | 0.017 [0.009, 0.025] | 0.014 [0.007, 0.021] |
| | | 0.05 | 0.052 [0.038, 0.066] | 0.051 [0.037, 0.065] | 0.053 [0.039, 0.067] | 0.054 [0.04, 0.068] | 0.047 [0.034, 0.06] | 0.044 [0.031, 0.057] |
| | | 0.1 | 0.095 [0.077, 0.113] | 0.096 [0.078, 0.114] | 0.105 [0.086, 0.124] | 0.105 [0.086, 0.124] | 0.106 [0.087, 0.125] | 0.100 [0.081, 0.119] |
| 2 | $\mathbf{0}$ | 0.01 | 0.005 [0.001, 0.009] | 0.009 [0.003, 0.015] | 0.013 [0.006, 0.020] | 0.012 [0.005, 0.019] | 0.009 [0.003, 0.015] | 0.010 [0.004, 0.016] |
| | | 0.05 | 0.039 [0.027, 0.051] | 0.043 [0.030, 0.056] | 0.053 [0.039, 0.067] | 0.049 [0.036, 0.062] | 0.043 [0.030, 0.056] | 0.056 [0.042, 0.070] |
| | | 0.1 | 0.093 [0.075, 0.111] | 0.081 [0.064, 0.098] | 0.096 [0.078, 0.114] | 0.094 [0.076, 0.112] | 0.091 [0.073, 0.109] | 0.109 [0.090, 0.128] |
| 0 | $\beta_{20} = 0.5$, $\beta_{40} = 0.5$ | 0.01 | 0.009 [0.003, 0.015] | 0.012 [0.005, 0.019] | 0.006 [0.001, 0.011] | 0.006 [0.001, 0.011] | 0.007 [0.002, 0.012] | 0.009 [0.003, 0.015] |
| | | 0.05 | 0.052 [0.038, 0.066] | 0.053 [0.039, 0.067] | 0.055 [0.041, 0.069] | 0.051 [0.037, 0.065] | 0.061 [0.046, 0.076] | 0.045 [0.032, 0.058] |
| | | 0.1 | 0.098 [0.080, 0.116] | 0.094 [0.076, 0.112] | 0.109 [0.090, 0.128] | 0.112 [0.092, 0.132] | 0.118 [0.098, 0.138] | 0.093 [0.075, 0.111] |
| 2 | $\beta_{20} = 0.5$, $\beta_{40} = 0.5$ | 0.01 | 0.022 [0.013, 0.031] | 0.013 [0.006, 0.020] | 0.012 [0.005, 0.019] | 0.012 [0.005, 0.019] | 0.007 [0.002, 0.012] | 0.013 [0.006, 0.020] |
| | | 0.05 | 0.090 [0.072, 0.108] | 0.038 [0.026, 0.050] | 0.050 [0.036, 0.064] | 0.048 [0.035, 0.061] | 0.046 [0.033, 0.059] | 0.049 [0.036, 0.062] |
| | | 0.1 | 0.144 [0.122, 0.166] | 0.092 [0.074, 0.110] | 0.100 [0.081, 0.119] | 0.103 [0.084, 0.122] | 0.110 [0.091, 0.129] | 0.099 [0.080, 0.118] |

test, as well as to the tests specifically designed for the detection of interaction effects, described in Section 2.3. To estimate the power, a simulation procedure analogous to that used in the previous section is used. However, here, the phenotype is generated according to the following model:

$$\text{logit}(\mathbb{P}[Y = 1|\boldsymbol{X} = \boldsymbol{x}, E = e]) = \beta_0 + \boldsymbol{s}'\boldsymbol{\beta},$$

where $\boldsymbol{s} = (x_1 e, \ldots, x_p e)'$. Several scenarios will be chosen by changing the coordinates of $\boldsymbol{\beta}$. The considered scenarios are adapted from those that were defined in Chapter 2 and used to compare the tests in the context of detection of gene - phenotype associations. More precisely, the vectors of parameters that were used to generate phenotypes associated to genetic profiles are used here to generate phenotypes associated to interaction profiles. Only the maximal values of the coefficients are modified in order to obtain a power near 1 for the most powerful tests for these values.

The power curves under the different scenarios are represented on Figure 3.1. For usual tests (minP, $L^2$-norm higher criticism and Hotelling's test), the observed differences in those scenarios are similar to the differences observed in Chapter 2. Moreover, the ranking of the MGF-R test is similar to that observed in the context of gene - phenotype associations detection. Under these scenarios, the test appears flexible and robust to the signal/dependence configuration, and its loss of power is always limited. The performances of the tests designed for the detection of interaction effects appear to be very inconstant. iSKAT and GESAT generally have weak performances in the considered scenarios. GESAT can be compared to usual tests only on gene KCNN3, and iSKAT is almost systematically the least powerful of all. Quite surprisingly, under scenario ($r_1$), the power of GESAT (and that of iSKAT, to a lesser extent), considerably decreases when the signal strength gets greater than 0.7. The three MiSTi tests have very similar and inconstant performances. Under scenario ($\ell_1$), these tests are among the most powerful ones. Under sceario ($r_1$), their performances are far superior to those of the usual tests, but clearly inferior to that of the MGF-R test. Under scenario ($\ell_2$), these tests have great power; on the contrary, under scenario ($r_2$), they are less powerful than usual tests. In both scenarios on gene DTD1, they are among the least powerful tests, with iSKAT and GESAT. In any case, it can be observed that the MGF-R test provides a satisfying power. Choosing a test rather than another would yield weak power in at least one

scenario. The good performances and the flexibility of the MGF-R test, observed in Chapter 2, are again highlighted here.

# 4 Discussion

In this chapter, the MGF-R test was adapted and applied to the detection of interaction effects between genes and an environmental factor. A suitable framework was first introduced in order to construct a vector of test statistics on which the MGF-R test could be applied.

An extensive simulation study was then realized in order to check if the properties of the test that were observed in Chapter 2 could still be found in the interaction context. It was thus noted that by using a parametric bootstrap procedure, the type I error rate was controlled by the MGF-R test even in presence of marginal effects. On the other hand, the GESAT test, whose $p$-value is obtained by using the theoretical asymptotic distribution of the test statistic, can undercontrol the type I error rate. This is particularly true when there are strong marginal effects.

Moreover, the power of the test for the detection of interaction effects was evaluated. Various scenarios were used, based on real dependence structures observed in the WTCCC data (Wellcome Trust Case Control Consortium, 2007). The results show that the MGF-R test keeps a generally satisfying power and adapts well to the signal/dependence configuration. Similarly as in Chapter 2, it is observed that minP, the $L^2$-norm test, the higher criticism and Hotelling's test all have inconstant performances from a situation to another. Furthermore, the test was compared to GESAT (Lin et al., 2013), iSKAT (Lin et al., 2016) and MiSTi (Su et al., 2017), which were specifically designed for the detection of gene - environmental factor effects. It was observed that these tests were all strongly affected by the dependence structure and the support of the signal. Under some scenarios, the power of these tests was indeed very weak. Thus, the MGF-R test remains flexible and robust, as observed in the context of gene - phenotype association detection.

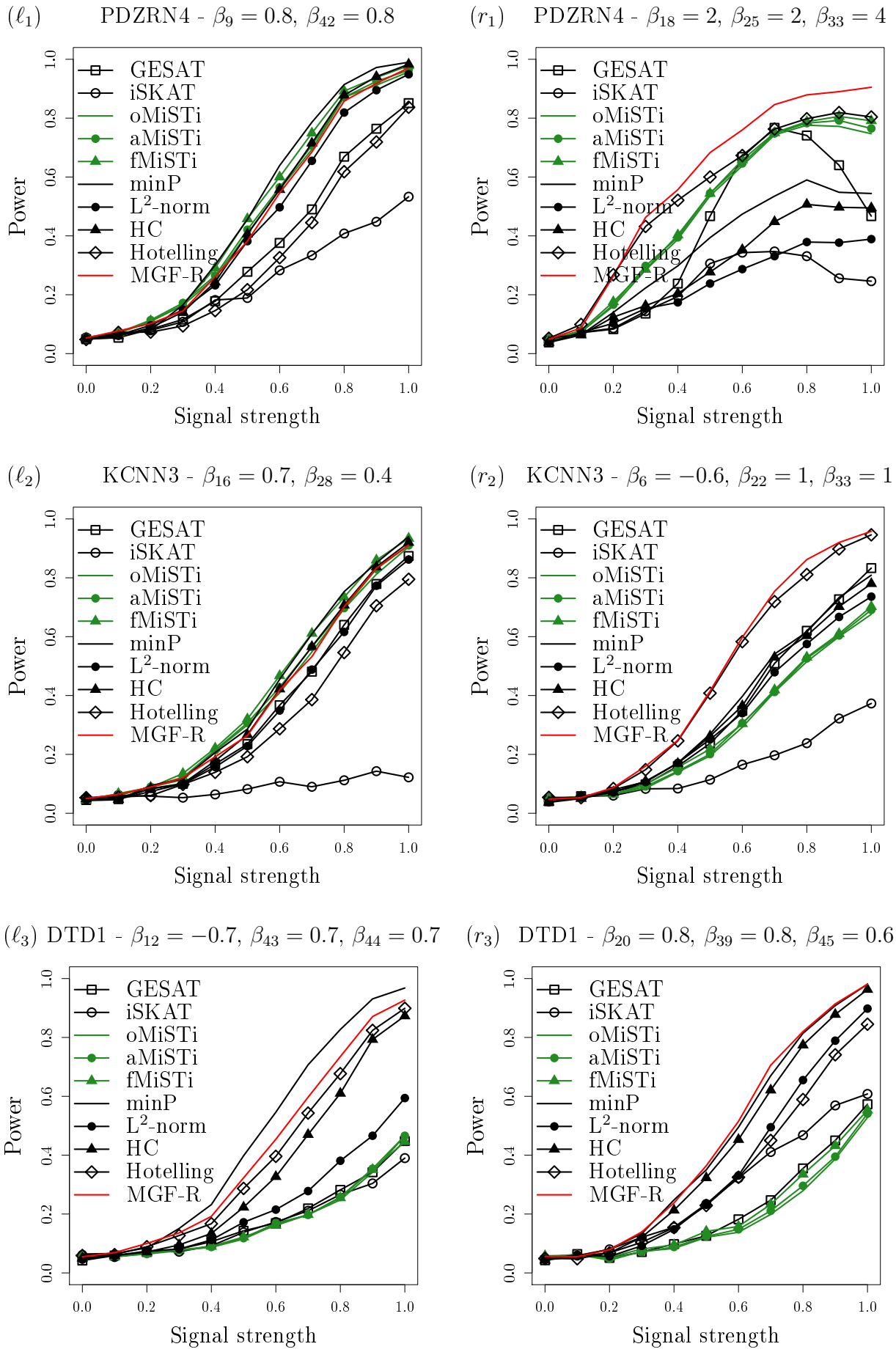


Figure 3.1: Power curves for the detection of gene - environment interaction effects under several scenarios on genes PDZRN4, KCNN3 and DTD1

# Chapter 4

# Detection of Interaction Effects Between Two Genes

**Abstract.** Genetic association studies led to the discovery of numerous regions significantly associated to phenotypes. Nevertheless, it has been observed that these numerous discoveries only account for a low part of the complexity of the studied diseases. It is now broadly assumed that more complex biological mechanisms have to be considered. Among these mechanisms, gene - gene interaction effects are thought to play a major role. Developing methods dedicated to the detection of such effects reveals to be a challenging problem, due to the great complexity of the studied effects. Several methods were proposed over the last few years, constructed on different definitions of the interaction effect. In the following, a generalized linear model suitable for the detection of gene - gene interaction effects is proposed. Then, a vector of test statistics is derived from this model. It will be exposed that its dependence structure has a very particular structure, of which advantage can be taken for its estimation. Finally, we propose to apply global testing methods on this vector. Under several situations, the formerly proposed MGF-R test appears to be among the most powerful ones, with the higher criticism test.

# Contents

**1 Introduction . . . . . 99**

**2 Gene - gene interaction effects modeling . . . . . 101**

2.1 A generalized linear model for the detection of gene - gene interaction effects . . . . . 101

2.2 Estimation of the distribution of a global test statistic under the null hypothesis . . . . . 103

2.3 Estimation of the marginal effects parameters . . . . . 104

**3 Properties of the correlation matrix of the test statistics vector . . . . . 105**

3.1 Modelization of the dependence structure . . . . . 105

3.2 Eigendecomposition . . . . . 106

3.3 Sparsity and relative weakness of the correlation coefficients 107

**4 Comparison of global testing methods for gene - gene interaction effects detection . . . . . 111**

4.1 Simulation procedure . . . . . 114

4.2 Results . . . . . 115

**5 Extension of the proposed model for other interaction effects . . . . . 117**

5.1 Dummy coding - based SNP - SNP interaction model . . . 119

5.2 Construction of a dummy coding - based gene - gene interaction model . . . . . 120

5.3 Correlation structure of the test statistics vector . . . . . 121

5.4 Comparison of the continuous and dummy codings . . . . 123

**6 Discussion . . . . . 127**

# 1 Introduction

As stated in the previous chapter, the generalized linear model (2.1) introduced in Chapter 2 is very flexible and can be used in diverse contexts. It was shown in Chapter 3 that this model could be suitably modified for the detection of gene - environment interaction effects. We now focus on the more challenging problem of gene - gene interaction effects detection, which requires deeper modifications of model (2.1).

Gene - phenotype association studies allowed the discovery of numerous genetic regions significantly associated to various phenotypes. Nevertheless, these associations explain only very partially the studied phenotypes; this is particularly true for complex phenotypes (Manolio et al., 2009). It is now widely admitted that other biological mechanisms must be taken into account to improve the understanding of the phenotypes. Among these mechanisms, gene - gene interaction effects are thought to be of major importance (Zuk et al., 2012; Woo et al., 2017; De et al., 2015).

Interaction effects between genes were initially considered at the SNP scale. Several tests were thus constructed to detect the presence of a significant interaction effect between two SNPs. For binary phenotypes, interaction test statistics are generally based on a comparison of the joint distribution of the two studied SNPs between cases and controls, namely by using generalized linear models (Cordell, 2002, 2009a). Numerous definitions of interaction were introduced, all of them being based on different biological or statistical hypotheses (Hu et al., 2014; Cordell, 2002). A different test statistic is associated to each definition, hence a great number of test statistics for testing the presence of a SNP - SNP interaction effect, and a variable power of each test depending on the nature of the interaction effect (Hu et al., 2014; Yu et al., 2015).

Researchers are now more and more interested in detecting interaction effects at the gene scale, *i.e.* considering a pair of genes and testing the presence of a significant interaction effect between these genes on a given phenotype. The issue of the nature of the interaction effect formerly evoked is also encountered in the problem of detection of interaction effects at the gene scale (Emily, 2018). This led to the

development of numerous methods, each corresponding to a specific definition of the interaction effect (Li et al., 2009a; Peng et al., 2010; Zhang et al., 2013; Rajapakse et al., 2012; Li et al., 2015).

The great diversity of existing methods for the detection of gene - gene interaction effects illustrates several difficulties. The first one concerns the modelization of the interaction effect, which directly depends on the underlying biological or statistical hypotheses. Then, interaction effects testing demands great computational resources. Indeed, searching for the presence of interaction effects between two genes can be considered as searching for the presence of interaction effects between any possible pair of SNPs which can be constructed from these two genes. From a statistical point of view, a natural modelization for searching interaction effects would be based on a generalized linear model similar to model (2.1). In the following, such a model is proposed, by adapting model (2.1) to the problem of gene - gene interaction effects detection. Consequently, a suitable framework for signal detection will be introduced. In this framework, usual global testing methods will be applied. Nevertheless, for the particular problem considered in this chapter, it can be seen that the correlation matrix of the test statistics vector has a very particular shape. Indeed, the dependence between test statistics here is inherited from the two genes which were used to compute them. However, it will also be shown that it is possible to take advantage of this particular structure. It will be demonstrated that the performances of the previously introduced signal detection methods and of the existing approaches for gene - gene interaction effects detection are inconstant, except for the higher criticism test, which seems quite robust. The previously proposed MGF-R test will also be included in the comparisons.

The outline of this chapter is the following. In Section 2, a global model extending model (2.1) is introduced. Similarly as in Chapter 2, a test statistics vector is derived. Then, Section 3 is dedicated to the study of the correlation matrix of the test statistics. Indeed, this correlation matrix has a very particular structure. An estimation method is given for this correlation matrix, which is used to derive its inverse or eigenvalue decomposition for a much lower computational cost. In Section 4, the global testing methods introduced in Chapter 2 and methods especially designed for the detection of gene - gene interaction effects are compared under several scenarios. As a possible extension, another global model for gene - gene

interaction effects is introduced in Section 5, based on a different coding of the SNPs. The chapter is finally ended by a discussion.

# 2 Gene - gene interaction effects modeling

## 2.1 A generalized linear model for the detection of gene - gene interaction effects

In Chapter 2, in which we were interested in detecting candidate genes possibly associated to a phenotype, the following generalized linear model was considered (Conneely and Boehnke, 2007):

$$h(\mathbb{E}[Y|\boldsymbol{U}=\boldsymbol{u},\boldsymbol{X}=\boldsymbol{x}])=\boldsymbol{u}'\boldsymbol{\alpha}+\boldsymbol{x}'\boldsymbol{\beta}$$

where $\boldsymbol{x}$ is a $p$-dimensional genotypic profile and $\boldsymbol{u}$ an optional $q$-dimensional covariates profile. Natural modifications can be applied to this model to adapt it to the problem of gene - gene interaction effects detection. Indeed, let $\boldsymbol{X^{(1)}}=(X_1^{(1)},\ldots,X_{p_1}^{(1)})'$ and $\boldsymbol{X^{(2)}}=(X_1^{(2)},\ldots,X_{p_2}^{(2)})'$ be two genes of dimension $p_1$ and $p_2$ respectively. The previous generalized linear model can then be extended to the problem of gene - gene interaction effects detection by considering the following $(p_1p_2)$-dimensional vector:

$$\boldsymbol{S}=(X_1^{(1)}X_1^{(2)},\ldots,X_1^{(1)}X_{p_2}^{(2)},\ldots,X_i^{(1)}X_1^{(2)},\ldots,X_i^{(1)}X_{p_2}^{(2)},\ldots,X_{p_1}^{(1)}X_1^{(2)},\ldots,X_{p_1}^{(1)}X_{p_2}^{(2)})',$$

which can be rewritten as:

$$\boldsymbol{S}=\boldsymbol{X^{(1)}}\otimes\boldsymbol{X^{(2)}}$$

where $\otimes$ denotes the Kronecker product between two vectors or matrices. This yields the following model:

$$h(\mathbb{E}[Y|\boldsymbol{U}=\boldsymbol{u},\boldsymbol{X^{(1)}}=\boldsymbol{x^{(1)}},\boldsymbol{X^{(2)}}=\boldsymbol{x^{(2)}}])=\boldsymbol{u}'\boldsymbol{\beta_U}+\boldsymbol{x^{(1)\prime}}\boldsymbol{\beta_{(1)}}+\boldsymbol{x^{(2)\prime}}\boldsymbol{\beta_{(2)}}+\boldsymbol{s}'\boldsymbol{\beta_S}. \tag{4.1}$$

Let us denote $\boldsymbol{W}=(\boldsymbol{U}',\boldsymbol{X^{(1)\prime}},\boldsymbol{X^{(2)\prime}})'$, $\boldsymbol{\alpha}=(\boldsymbol{\beta}_{\boldsymbol{U}}',\boldsymbol{\beta}_{\boldsymbol{(1)}}',\boldsymbol{\beta}_{\boldsymbol{(2)}}')'$ and $\boldsymbol{\beta}=\boldsymbol{\beta_S}$. Model (4.1) can then be written under a form similar to model (2.1), as follows:

$$h(\mathbb{E}[Y|\boldsymbol{X^{(1)}}=\boldsymbol{x^{(1)}},\boldsymbol{X^{(2)}}=\boldsymbol{x^{(2)}},\boldsymbol{U}=\boldsymbol{u}])=\boldsymbol{w}'\boldsymbol{\alpha}+\boldsymbol{s}'\boldsymbol{\beta}.$$

Testing for the presence of at least one interaction effect then amounts to testing for the global nullity of $\boldsymbol{\beta}$:

$$\begin{cases} H_0 : & \boldsymbol{\beta} = \mathbf{0} \\ H_1 : & \boldsymbol{\beta} \neq \mathbf{0}. \end{cases} \tag{4.2}$$

In the following, we denote $\mathbb{W}$ the matrix whose $i$-th row is obtained by concatening the covariates profile and the genotypic profiles of the $i$-th individual. Each row of $\mathbb{W}$ is a vector of dimension $(q + p_1 + p_2)$ as follows:

$$\mathbb{W} = \left(\begin{array}{ccc|ccc|ccc} u_{11} & \dots & u_{1q} & x_{11}^{(1)} & \dots & x_{1p_1}^{(1)} & x_{11}^{(2)} & \dots & x_{1p_2}^{(2)} \\ \vdots & & \vdots & \vdots & & \vdots & \vdots & & \vdots \\ u_{i1} & \dots & u_{iq} & x_{i1}^{(1)} & \dots & x_{ip_1}^{(1)} & x_{i1}^{(2)} & \dots & x_{ip_2}^{(2)} \\ \vdots & & \vdots & \vdots & & \vdots & \vdots & & \vdots \\ u_{n1} & \dots & u_{nq} & x_{n1}^{(1)} & \dots & x_{np_1}^{(1)} & x_{n1}^{(2)} & \dots & x_{np_2}^{(2)} \end{array}\right)$$

$$= [\, \mathbb{U} \mid \mathbb{X}^{(1)} \mid \mathbb{X}^{(2)} \,]$$

where $u_{ij}$ is the value taken by the $i$-th individual for the $j$-th covariate and $x_{ij}^{(\ell)}$ is the value taken by the $i$-th individual for the $j$-th SNP of the $\ell$-th gene. Similarly, we denote $\mathbb{S}$ the matrix whose $i$-th row is the interaction profile of the $i$-th individual:

$$\mathbb{S} = \begin{pmatrix} (\boldsymbol{x_1^{(1)}} \otimes \boldsymbol{x_1^{(2)}})' \\ \vdots \\ (\boldsymbol{x_i^{(1)}} \otimes \boldsymbol{x_i^{(2)}})' \\ \vdots \\ (\boldsymbol{x_n^{(1)}} \otimes \boldsymbol{x_n^{(2)}})' \end{pmatrix}$$

where $\boldsymbol{x_i^{(\ell)}} = (x_{i1}^{(\ell)}, \dots, x_{ij}^{(\ell)}, \dots, x_{ip_\ell}^{(\ell)})'$.

We now denote $\boldsymbol{S_j}$ the $j$-th column of $\mathbb{S}$. Similarly as in the previous chapters, a vector of test statistics $\boldsymbol{\mathcal{Z}}$ is introduced for testing (4.2). The coordinates of $\boldsymbol{\mathcal{Z}}$ are denoted $Z_1, \dots, Z_p$ with $p = p_1 p_2$. The $j$-th coordinate of $\boldsymbol{\mathcal{Z}}$ is defined by:

$$Z_j = \frac{\boldsymbol{S_j'}(\boldsymbol{y} - \hat{\boldsymbol{y}}_{\mathbf{0}})}{\sqrt{\widehat{\boldsymbol{\Gamma}}_{j,j}}}$$

where $\hat{\boldsymbol{y}}_{\mathbf{0}}$ is the $n$-vector whose $i$-th coordinate is $h^{-1}(\boldsymbol{w_i'}\hat{\boldsymbol{\alpha}}_{\mathbf{0}})$ and $\hat{\boldsymbol{\alpha}}_{\mathbf{0}}$ is the estimate of $\boldsymbol{\alpha}$ under $H_0$. $\widehat{\boldsymbol{\Gamma}}$ is the estimated variance-covariance matrix of $\mathbb{S}'(\boldsymbol{y} - \hat{\boldsymbol{y}}_{\mathbf{0}})$; following

Conneely and Boehnke (2007), this estimator is defined as:

$$\widehat{\mathbf{\Gamma}} = \hat{\sigma}_Y^2(\mathbb{S}'\mathbb{S} - \mathbb{S}'\mathbb{W}(\mathbb{W}'\mathbb{W})^{-1}\mathbb{W}'\mathbb{S}),$$

with $\hat{\sigma}_Y^2 = \frac{1}{n}(\boldsymbol{y} - \hat{\boldsymbol{y}}_{\mathbf{0}})'(\boldsymbol{y} - \hat{\boldsymbol{y}}_{\mathbf{0}})$. To compute test (4.2), global tests introduced in Chapter 2 can then be applied to $\boldsymbol{\mathcal{Z}}$.

## 2.2 Estimation of the distribution of a global test statistic under the null hypothesis

We propose to construct a procedure dedicated to the detection of interaction effects between two genes on a phenotype. This procedure is based on model (4.1) and on the construction of a global test statistic $T(\boldsymbol{\mathcal{Z}})$ obtained by aggregating the coordinates of $\boldsymbol{\mathcal{Z}}$. As previously, it will be mandatory to estimate the distribution of $T(\boldsymbol{\mathcal{Z}})$ under the null hypothesis. Under $H_0$, $\boldsymbol{\beta_S} = \mathbf{0}$; the model under $H_0$ can thus be written as:

$$h(\mathbb{E}[Y|\boldsymbol{U} = \boldsymbol{u}, \boldsymbol{X^{(1)}} = \boldsymbol{x^{(1)}}, \boldsymbol{X^{(2)}} = \boldsymbol{x^{(2)}}]) = \boldsymbol{u}'\boldsymbol{\beta}_{\boldsymbol{U}} + \boldsymbol{x^{(1)\prime}}\boldsymbol{\beta_{(1)}} + \boldsymbol{x^{(2)\prime}}\boldsymbol{\beta_{(2)}}.$$

Consequently, as in Chapter 3, it is underlined here that it will not be possible to estimate the distribution of the test statistic under $H_0$ by using simple random permutations of the vector $\boldsymbol{y}$ containing the values of the phenotype variable. Indeed, a simple permutation of $\boldsymbol{y}$ amounts to considering the hypothesis of simultaneous nullity of all parameters of model (4.1). This hypothesis is more restrictive than $H_0$. In presence of non-negligible marginal effects, using simple permutations would lead to a wrong estimation of the null distribution of $T(\boldsymbol{\mathcal{Z}})$ (Bůžková et al., 2011; Coombes and Biernacka, 2018; Dudbridge and Fletcher, 2014). It is namely shown by Bůžková et al. (2011) on simulations that the control of the type I error rate can be very much affected by ignoring the marginal effects. To remedy this problem, as in the previous chapter, we propose to adapt the parametric bootstrap procedure of Bůžková et al. (2011) to the present context.

Let $\hat{\boldsymbol{\beta}}_{\boldsymbol{U}}^{(\mathbf{0})}$, $\hat{\boldsymbol{\beta}}_{\boldsymbol{(1)}}^{(\mathbf{0})}$, and $\hat{\boldsymbol{\beta}}_{\boldsymbol{(1)}}^{(\mathbf{0})}$ be the estimators of $\boldsymbol{\beta}_{\boldsymbol{U}}$, $\boldsymbol{\beta_{(1)}}$ and $\boldsymbol{\beta_{(2)}}$ under the null hypothesis. A phenotype is then generated by randomly sampling according to the estimated distribution of $Y$ under the considered model. For a binary phenotype,

the $i$-th coordinate of the simulated phenotype is obtained by randomly sampling according to the $\mathcal{B}(\hat{y}_{0,i})$ distribution, where $\hat{y}_{0,i}$ is the estimated probability for the $i$-th individual to take the value 1 under $H_0$:

$$\hat{y}_{0,i} = \frac{\exp\left(\boldsymbol{u}_i'\hat{\boldsymbol{\beta}}_{\boldsymbol{U}}^{(\mathbf{0})} + \boldsymbol{x}_i^{(\mathbf{1})\prime}\hat{\boldsymbol{\beta}}_{(\mathbf{1})}^{(\mathbf{0})} + \boldsymbol{x}_i^{(\mathbf{2})\prime}\hat{\boldsymbol{\beta}}_{(\mathbf{2})}^{(\mathbf{0})}\right)}{1 + \exp\left(\boldsymbol{u}_i'\hat{\boldsymbol{\beta}}_{\boldsymbol{U}}^{(\mathbf{0})} + \boldsymbol{x}_i^{(\mathbf{1})\prime}\hat{\boldsymbol{\beta}}_{(\mathbf{1})}^{(\mathbf{0})} + \boldsymbol{x}_i^{(\mathbf{2})\prime}\hat{\boldsymbol{\beta}}_{(\mathbf{2})}^{(\mathbf{0})}\right)}.$$

Thus, marginal effects inducing a dependence between $Y$ and $\boldsymbol{U}$, $\boldsymbol{X}^{(\mathbf{1})}$ and $\boldsymbol{X}^{(\mathbf{2})}$ are maintained for the simulated phenotype. In the following, only binary phenotypes are considered.

## 2.3 Estimation of the marginal effects parameters

It is necessary to estimate the marginal genetic effects to compute the vector of test statistics. Nevertheless, due to the strong within-gene dependence structure, estimating these effects can be challenging. To counteract this problem, we propose to replace the genetic profiles by the corresponding principal components profiles. Consequently, in $\mathbb{W} = [\ \mathbb{U} \mid \mathbb{X}^{(1)} \mid \mathbb{X}^{(2)}\ ]$, $\mathbb{X}^{(\ell)}$ is replaced by the $n \times k_\ell$ matrix $\tilde{\mathbb{X}}^{(\ell)}$ whose $j$-th column $\tilde{\boldsymbol{X}}_{\boldsymbol{j}}^{(\boldsymbol{\ell})}$ is defined as

$$\tilde{\boldsymbol{X}}_{\boldsymbol{j}}^{(\boldsymbol{\ell})} = \frac{\mathbb{X}^{(\ell)}\boldsymbol{v}_{\boldsymbol{j}}^{(\boldsymbol{\ell})}}{\sqrt{\lambda_j^{(\ell)}}}$$

where $\boldsymbol{v}_{\boldsymbol{j}}^{(\boldsymbol{\ell})}$ is the $j$-th eigenvector of the estimated correlation matrix of the columns of $\mathbb{X}^{(\ell)}$, and $\lambda_j^{(\ell)}$ is the associated eigenvalue. The number of columns $k_\ell$ of the matrix $\tilde{\mathbb{X}}^{(\ell)}$ is the number of retained principal components. We aim at obtaining a stable estimation of the marginal effects but not at reducing the dimensionality of the vector of parameters. As in Section 2.2.2 of Chapter 3, we thus propose to keep principal components which account for at least 0.1% of the total variance for each gene, which enables to obtain a stable estimation of the parameters without reducing the dimensionality too importantly. Indeed, greater cut-offs could lead to remove too many principal components, which could then yield a biased estimation of the marginal effects. Then, the type I error rate would not be rightfully controlled.

In the remainder of this chapter, for simplicity, we suppose there are no covariates in the model (except for the intercept term).

# 3 Properties of the correlation matrix of the test statistics vector

## 3.1 Modelization of the dependence structure

We now focus on the dependence structure of the test statistics vector. It can be observed that the estimated correlation matrix $\widehat{\mathbf{\Sigma}}$ of $\boldsymbol{\mathcal{Z}}$ has a very particular structure, since it is inherited from both genes used in the model. Let us highlight this point with an example.

Five SNPs were selected in the PDZRN4 gene and in the KCNN3 gene, used in the previous chapters. A sample of 1,000 cases and 1,000 controls was then generated without any effect between the phenotype and the genes. The estimated correlation matrix $\widehat{\mathbf{\Sigma}}$ of the test statistics vector was then computed using the formula previously given and adapted from that given in Conneely and Boehnke (2007):

$$\widehat{\mathbf{\Gamma}} = \hat{\sigma}^2_Y(\mathbb{S}'\mathbb{S} - \mathbb{S}'\mathbb{W}(\mathbb{W}'\mathbb{W})^{-1}\mathbb{W}'\mathbb{S}).$$

$\widehat{\mathbf{\Sigma}}$ is then obtained as $\widehat{\mathbf{\Sigma}} = \text{diag}(\widehat{\mathbf{\Gamma}})^{-1/2}\widehat{\mathbf{\Gamma}}\,\text{diag}(\widehat{\mathbf{\Gamma}})^{-1/2}$. The correlation matrices of each gene and the correlation matrix of the test statistics vector are represented on Figure 4.1.

The striking observation on these figures is that the estimated correlation matrix $\widehat{\mathbf{\Sigma}}$ of the test statistics vector has a very particular structure. The structure seems to be inherited from those of both genes included in the model. Let us denote $\widehat{\mathbf{\Sigma}}_\mathbf{1}$ and $\widehat{\mathbf{\Sigma}}_\mathbf{2}$ the estimated correlation matrices corresponding to each gene. Along the diagonal, one can observe blocks corresponding exactly to $\widehat{\mathbf{\Sigma}}_\mathbf{2}$. Aside from these blocks, the shape of $\widehat{\mathbf{\Sigma}}_\mathbf{2}$ is recognized everywhere in $\widehat{\mathbf{\Sigma}}$. Nevertheless, it can be observed that the values within the block are modified. For instance, taking the block of $\widehat{\mathbf{\Sigma}}$ composed of the rows 1 to 5 and columns 11 to 15, the structure is exactly that of $\widehat{\mathbf{\Sigma}}_\mathbf{2}$, but the coefficients are negative, contrary to the initial $\widehat{\mathbf{\Sigma}}_\mathbf{2}$. This can be considered as the third top block in $\widehat{\mathbf{\Sigma}}$ (starting from the left). Furthermore, the third value of the first row of $\widehat{\mathbf{\Sigma}}_\mathbf{1}$ is strongly negative. It can be seen that the crossing between this value and $\widehat{\mathbf{\Sigma}}_\mathbf{2}$ made the considered block have the structure of $\widehat{\mathbf{\Sigma}}_\mathbf{2}$, but with negative values. This observation can be transposed to any block in

$\widehat{\mathbf{\Sigma}}$. Finally, it seems that $\mathbf{\Sigma}$ can be reasonably estimated by the following matrix:

$$\begin{pmatrix} \hat{\sigma}_{11}^{(1)}\widehat{\mathbf{\Sigma}}_{\mathbf{2}} & \dots & \hat{\sigma}_{1j}^{(1)}\widehat{\mathbf{\Sigma}}_{\mathbf{2}} & \dots & \hat{\sigma}_{1p_1}^{(1)}\widehat{\mathbf{\Sigma}}_{\mathbf{2}} \\ \vdots & & \vdots & & \vdots \\ \hat{\sigma}_{i1}^{(1)}\widehat{\mathbf{\Sigma}}_{\mathbf{2}} & \dots & \hat{\sigma}_{ij}^{(1)}\widehat{\mathbf{\Sigma}}_{\mathbf{2}} & \dots & \hat{\sigma}_{ip_1}^{(1)}\widehat{\mathbf{\Sigma}}_{\mathbf{2}} \\ \vdots & & \vdots & & \vdots \\ \hat{\sigma}_{p_1 1}^{(1)}\widehat{\mathbf{\Sigma}}_{\mathbf{2}} & \dots & \hat{\sigma}_{p_1 j}^{(1)}\widehat{\mathbf{\Sigma}}_{\mathbf{2}} & \dots & \hat{\sigma}_{p_1 p_1}^{(1)}\widehat{\mathbf{\Sigma}}_{\mathbf{2}} \end{pmatrix}$$

where $\hat{\sigma}_{ij}^{(1)}$ is the $(i,j)$-th term of $\widehat{\mathbf{\Sigma}}_{\mathbf{1}}$. It can be seen that the above matrix is the Kronecker product of $\widehat{\mathbf{\Sigma}}_{\mathbf{1}}$ and $\widehat{\mathbf{\Sigma}}_{\mathbf{2}}$; consequently, it seems $\mathbf{\Sigma}$ can be reasonably estimated by $\widehat{\mathbf{\Sigma}}_{\mathbf{1}} \otimes \widehat{\mathbf{\Sigma}}_{\mathbf{2}}$.

## 3.2 Eigendecomposition

The former estimation method offers a clear advantage for estimating the inverse or the eigendecomposition of $\mathbf{\Sigma}$, which are necessary to compute Hotelling's statistic and the MGF-R test. Indeed, let us introduce the eigendecompositions of $\widehat{\mathbf{\Sigma}}_{\mathbf{1}}$ and $\widehat{\mathbf{\Sigma}}_{\mathbf{2}}$ as follows:

$$\begin{aligned} \widehat{\mathbf{\Sigma}}_{\mathbf{1}} &= \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}', \quad \boldsymbol{V} = [\, \boldsymbol{v_1} \mid \dots \mid \boldsymbol{v_{p_1}} \,], \quad \boldsymbol{\Lambda} = \operatorname{diag}(\lambda_1, \dots, \lambda_{p_1}) \\ \widehat{\mathbf{\Sigma}}_{\mathbf{2}} &= \boldsymbol{W}\boldsymbol{\Omega}\boldsymbol{W}', \quad \boldsymbol{W} = [\, \boldsymbol{w_1} \mid \dots \mid \boldsymbol{w_{p_2}} \,], \quad \boldsymbol{\Omega} = \operatorname{diag}(\omega_1, \dots, \omega_{p_2}). \end{aligned}$$

Then, the eigendecomposition of $\widehat{\mathbf{\Sigma}}_{\mathbf{1}} \otimes \widehat{\mathbf{\Sigma}}_{\mathbf{2}}$ is $\boldsymbol{Q}\boldsymbol{\Theta}\boldsymbol{Q}'$, where

$$\begin{aligned} \boldsymbol{Q} &= [\, \boldsymbol{v_1} \otimes \boldsymbol{w_1} \mid \dots \mid \boldsymbol{v_1} \otimes \boldsymbol{w_{p_2}} \mid \dots \mid \boldsymbol{v_i} \otimes \boldsymbol{w_1} \mid \dots \mid \boldsymbol{v_i} \otimes \boldsymbol{w_{p_2}} \mid \\ &\qquad \dots \mid \boldsymbol{v_{p_1}} \otimes \boldsymbol{w_1} \mid \dots \mid \boldsymbol{v_{p_1}} \otimes \boldsymbol{w_{p_2}} \,], \\ \boldsymbol{\Theta} &= \boldsymbol{\Lambda} \otimes \boldsymbol{\Omega}. \end{aligned}$$

See for example Horn and Johnson (1994), theorem 4.2.12. Consequently, the eigendecomposition of $\widehat{\mathbf{\Sigma}}_{\mathbf{1}} \otimes \widehat{\mathbf{\Sigma}}_{\mathbf{2}}$, which is a $(p_1 p_2) \times (p_1 p_2)$ matrix, can be obtained through the eigendecompositions of a $p_1 \times p_1$ matrix and of a $p_2 \times p_2$ matrix. This provides a major computational gain. Indeed, to derive the eigendecomposition of $\widehat{\mathbf{\Sigma}}$, $p = p_1 p_2$ eigenvalues must be derived, and so must $p$ eigenvectors of dimension $p$, which is critical due to the great dimension $p$. On the other hand, only $p_1$ (respectively $p_2$) eigenvalues and $p_1$ (respectively $p_2$) eigenvectors of dimension $p_1$ (respectively $p_2$) are required for the eigendecomposition of $\widehat{\mathbf{\Sigma}}_{\mathbf{1}}$ (respectively $\widehat{\mathbf{\Sigma}}_{\mathbf{2}}$). In the following,

when the inverse or the eigendecomposition of $\mathbf{\Sigma}$ is needed, it is estimated by using the former result, assuming $\mathbf{\Sigma}$ can be estimated by $\widehat{\mathbf{\Sigma}}_\mathbf{1} \otimes \widehat{\mathbf{\Sigma}}_\mathbf{2}$.

To give an example of how efficient the proposed method is, we generate a sample using genes PDZRN4 and KCNN3, the phenotype being generated under the null hypothesis. Then, the eigendecomposition of the estimated correlation matrix of the test statistics vector is computed, first using the direct formula and then using the method described above. The first method took around 24 seconds, while the second one required only 0.02 second approximately. The computational gain is far from negligible. Moreover, we expect the estimation to be much more stable. Indeed, the direct formula requires the estimation of $p_1p_2(p_1p_2-1)/2$ extradiagonal terms, whereas the estimation method proposed in this section only requires $p_1(p_1-1)/2+p_2(p_2-1)/2$ extradiagonal terms. In the particular case where $p_1 = p_2$, the number of extradiagonal terms to be estimated by using the direct formula is $p_1^2(p_1^2-1)/2$, against only $p_1(p_1-1)$ by using the above proposed method. In the following, only the eigenvectors associated to an eigenvalue greater than 1 are used, in order to avoid using noisy eigenvectors.

## 3.3 Sparsity and relative weakness of the correlation coefficients

It can be observed that the values of the correlation coefficients in $\mathbf{\Sigma}$ will be weaker than those in $\mathbf{\Sigma_1}$ and $\mathbf{\Sigma_2}$, and that the proportion of weak values will be greater in $\mathbf{\Sigma}$. Indeed, let us suppose that the decomposition $\mathbf{\Sigma} = \mathbf{\Sigma_1} \otimes \mathbf{\Sigma_2}$ is true. We further assume that $\mathbf{\Sigma_1}$ contains $a_1p_1^2$ terms which absolute value is lower than $\varepsilon \in [0,1]$ and $(b_1-a_1)p_1^2$ terms which absolute value is in the interval $[\varepsilon, \sqrt{\varepsilon}]$, with $a_1 \in [0,1], b_1 \in [a_1,1]$. Similarly, we assume that $\mathbf{\Sigma_2}$ contains $a_2p_2^2$ terms which absolute value is lower than $\varepsilon$ and $(b_2-a_2)p_2^2$ terms which absolute value is in the interval $[\varepsilon, \sqrt{\varepsilon}]$, with $a_2 \in [0,1], b_2 \in [a_2,1]$. Let us denote $\sigma_{ij}^{(\ell)}$ the $(i,j)$-th term of $\mathbf{\Sigma_\ell}$ and $\sigma_{ij}$ the $(i,j)$-th term of $\mathbf{\Sigma}$. The following points can then be observed:

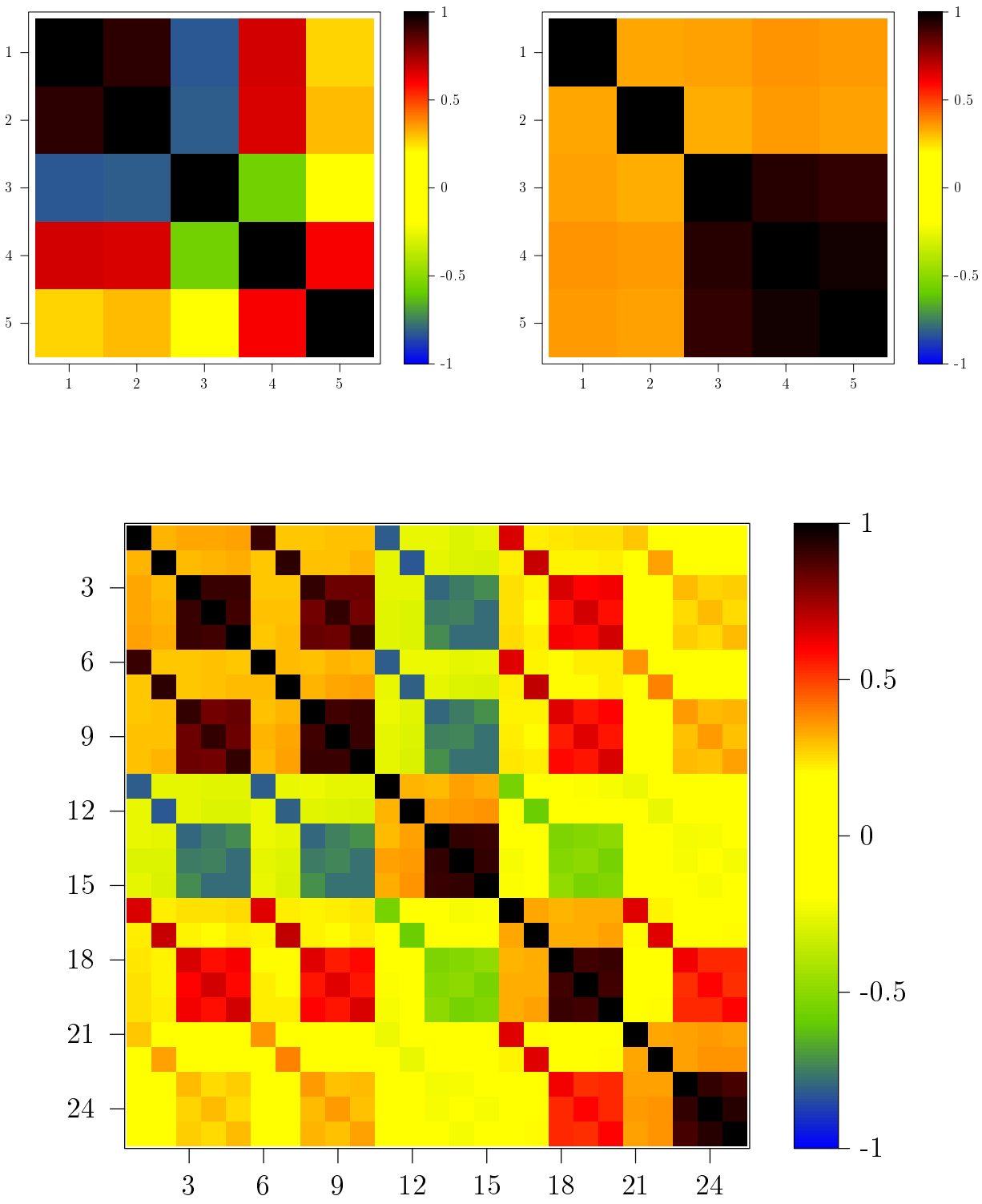


Figure 4.1: Upper part: correlation matrices of 5 SNPs from genes PDZRN4 and KCNN3. Lower part: corresponding correlation matrix for the test statistics vector

- if $|\sigma_{ij}^{(1)}| \leq \varepsilon$, then all the $p_2^2$ terms of $\sigma_{ij}^{(1)}\mathbf{\Sigma_2}$ are lower than $\varepsilon$ in absolute value, thus $a_1 p_1^2 p_2$ terms are lower than $\varepsilon$ in $\mathbf{\Sigma}$ in absolute value
- if $\varepsilon \leq |\sigma_{ij}^{(1)}| \leq \sqrt{\varepsilon}$, then the $(k,\ell)$-th term of $\sigma_{ij}^{(1)}\mathbf{\Sigma_2}$ is lower than $\varepsilon$ in absolute value if $|\sigma_{k\ell}^{(2)}| \leq \varepsilon$ ($a_2 p_2^2$ terms) or if $\varepsilon \leq |\sigma_{k\ell}^{(2)}| \leq \sqrt{\varepsilon}$ ($(b_2 - a_2)p_2^2$ terms), thus $(b_1 - a_1)b_2 p_2^2$ terms are lower than $\varepsilon$ in $\mathbf{\Sigma}$ in absolute value
- if $|\sigma_{ij}^{(1)}| \geq \sqrt{\varepsilon}$, then the $(k,\ell)$-th term of $\sigma_{ij}^{(1)}\mathbf{\Sigma_2}$ is lower than $\varepsilon$ in absolute value if $|\sigma_{k\ell}^{(2)}| \leq \varepsilon$ ($a_2 p_2^2$ terms), thus $(1 - b_1)a_2 p_1^2 p_2^2$ terms are lower than $\varepsilon$ in $\mathbf{\Sigma}$ in absolute value.

More terms in $\mathbf{\Sigma}$ might be lower than $\varepsilon$, but only the certain cases are enumerated above. Finally, we get that $\mathbf{\Sigma}$ contains at least $(a_1+(b_1-a_1)b_2+(1-b_1)a_2)p_1^2p_2^2$ terms lower than $\varepsilon$ in absolute value; the proportion of terms lower than $\varepsilon$ in absolute value in $\mathbf{\Sigma}$ is greater than $\hat{\psi} = a_1 + (b_1 - a_1)b_2 + (1 - b_1)a_2$. It can be straightforwardly observed that $\hat{\psi} \geq a_1$. Moreover, by rearranging the terms, we get $\hat{\psi} = a_2 + (b_2 - a_2)b_1 + (1 - b_2)a_1$, and thus $\hat{\psi} \geq a_2$. Consequently, the proportion of coefficients lower than $\varepsilon$ in absolute value is greater in $\mathbf{\Sigma}$ than in both $\mathbf{\Sigma_1}$ and $\mathbf{\Sigma_2}$.

For several pairs of genes, we compare the lower bound $\hat{\psi}$ to the proportions $a_1$ and $a_2$ on a scale of values for $\varepsilon$. We also compare the actual proportion $\psi$ directly calculated on $\mathbf{\Sigma}$ to assess the quality of the lower bound derived above. The proportions $a_1$ and $a_2$ are directly calculated on the correlation matrices and displayed as functions of $\varepsilon$, as well as the proportion $\psi$ for $\mathbf{\Sigma}$ and the lower bound $\hat{\psi}$ derived above. The results are displayed on Figure 4.2. It can be observed that the lower bound $\hat{\psi}$ is close to the true value $\psi$, for any value of $\varepsilon$. We further observe that in any case and for any value of $\varepsilon$, $\psi$ (or $\hat{\psi}$) is obviously greater than both $a_1$ and $a_2$. In particular, it can be much greater than $\min(a_1, a_2)$, for example on the KCNN3/DTD1 pair. Consequently, the global dependence among the test statistics will be weaker than what was observed in Chapter 2.

We give another graphic representation of the lower bound $\hat{\psi}$ on the proportions of terms $\sigma_{ij}$ such that $|\sigma_{ij}| \leq \varepsilon$, considering arbitrary correlation matrices $\mathbf{\Sigma_1}$ and $\mathbf{\Sigma_2}$. The only assumption in the following is that the respective proportions in these matrices are equal: $a_1 = a_2 = a$ and $b_1 = b_2 = b$, $b \geq a$. Then $\hat{\psi} = 2a + b^2 - 2ab$. $\hat{\psi}$ is represented as a function of $a$ and $b$ on Figure 4.3 as an heatmap with level

curves. This figure gives indications on how $\hat{\psi}$ evolves with $a$ and $b$, for any value of $\varepsilon$. It can be observed that the proportion of coefficients lower than $\varepsilon$ in absolute value in $\mathbf{\Sigma}$ grows quickly with $a$ and $b$. For instance, let us suppose that $\mathbf{\Sigma_1}$ and $\mathbf{\Sigma_2}$ both contain a proportion $a = 0.2$ of terms lower than $\varepsilon$ in absolute value and a proportion $b = 0.6$ of terms lower than $\sqrt{\varepsilon}$ in absolute value. Then, the proportion of terms in $\mathbf{\Sigma}$ lower than $\varepsilon$ is greater than $\hat{\psi} \simeq 0.5$. In this example, if we take $\varepsilon$ as a weak value, *e.g.* $\sqrt{\varepsilon} = 0.3$, then more than half of the coefficients of $\mathbf{\Sigma}$ are lower than $0.3^2$. Such coefficients are close enough to 0 to be considered negligible, thus showing that $\mathbf{\Sigma}$ is sparser than $\mathbf{\Sigma_1}$ and $\mathbf{\Sigma_2}$.

As a conclusion, the correlation matrix $\mathbf{\Sigma}$ is sparser and has weaker correlation coefficients than both $\mathbf{\Sigma_1}$ and $\mathbf{\Sigma_2}$. Indeed, since all the elements of $\mathbf{\Sigma}$ are products of the elements of the two other matrices, the elements of $\mathbf{\Sigma}$ have obviously lower values, since the elements of $\mathbf{\Sigma_1}$ and $\mathbf{\Sigma_2}$ belong to $[-1, 1]$. This and the great dimension of the test statistics vector will undoubtedly impact the performances of the global testing methods. The dimension of the test statistics vector is several dozens of times greater than what it used to be in the previous chapters, which might advantage the higher criticism test. Indeed, it was designed for large scale problems, and its optimality is obtained asymptotically with respect to the dimension of the vector (Donoho and Jin, 2004; Arias-Castro et al., 2011). Moreover, Arias-Castro et al. (2011) showed the optimality of the HC in a linear model $y = \boldsymbol{X}\boldsymbol{b} + \varepsilon$ where $\boldsymbol{b}$ is sparse and the explanatory variables are weakly correlated. They state that the explanatory variables are weakly correlated if their correlation matrix $\boldsymbol{C}_{p\times p}$ fulfills the two following conditions:

- $\forall i \neq j, |c_{ij}| \leq 1 - \dfrac{1}{\log(p)}$
- $|\{j, |c_{ij}| > \gamma\}| \leq \Delta$

where $\gamma$ and $\Delta$ are given values. Then, still assuming that $\mathbf{\Sigma} = \mathbf{\Sigma_1} \otimes \mathbf{\Sigma_2}$, and denoting the $(s,t)$-th term of $\mathbf{\Sigma}$ as $\sigma_{st} = \sigma^{(1)}_{ij}\sigma^{(2)}_{k\ell}$, $s = p_2 i - p_2 + k$ and $t = p_2 j - p_2 + \ell$, $\mathbf{\Sigma}$ corresponds to a weakly correlated design if

- $\forall i \neq j, |\sigma^{(1)}_{ij}\sigma^{(2)}_{k\ell}| \leq 1 - \dfrac{1}{\log(p_1 p_2)}$

- $|\{j, |\sigma_{ij}^{(1)}\sigma_{ij}^{(2)}| > \gamma\}| \leq \Delta$.

It can be seen these assumptions are less restrictive than the initial ones. Consequently, under the assumption that $\mathbf{\Sigma} = \mathbf{\Sigma_1} \otimes \mathbf{\Sigma_2}$, $\mathbf{\Sigma}$ can fulfill the two above assumptions with weaker assumptions on $\mathbf{\Sigma_1}$ or $\mathbf{\Sigma_2}$. However, how will the particular shape of the correlation matrix impact the methods is unsure. In the introduction of this thesis, it was stated that taking dependence explicitly into account by decorrelating the test statistics could improve the detection power as well as it could deter it. Nevertheless, this was demonstrated only by considering very simple situations in which a few dozens of pointwise test statistics were involved. In the present chapter, due to the great dimension of the test statistics vector (several hundreds or even thousands of pointwise test statistics) and to its correlation structure, taking dependence explictly into account might not be as profitable as it previously was.

# 4 Comparison of global testing methods for gene - gene interaction effects detection

In this section, the power of the MGF-R test is compared to that of the higher criticism, minP, $L^2$-norm and Hotelling's tests for gene - gene interaction effects. Each global test is applied to the test statistics vector previously introduced. The comparisons will also include the following methods, which were dedicated to the detection of gene - gene interactions: PCA (Li et al., 2009a), CCA (Peng et al., 2010), CLD (Rajapakse et al., 2012) and Aggregator (Emily, 2016).

The CCA and CLD methods are based on a comparison of the covariance or linkage disequilibrium structures of the genes between cases and controls. The PCA method is a likelihood-ratio test under a logistic model based on the principal components corresponding to the two genes. Finally, the Aggregator method is based on the construction of the $p_1p_2$ SNP - SNP interaction models for each pair of SNPs $(X_i^{(1)}, X_j^{(2)})$ as follows (Cordell et al., 2001; Cordell, 2002, 2009a; Musameh et al., 2015; Emily, 2016; Ueki and Cordell, 2012):

$$h(\mathbb{E}[Y|X_i^{(1)} = x_1, X_j^{(2)} = x_2]) = \beta_0^{(i,j)} + \beta_1^{(i,j)}x_1 + \beta_2^{(i,j)}x_2 + \beta_3^{(i,j)}x_1x_2. \tag{4.3}$$

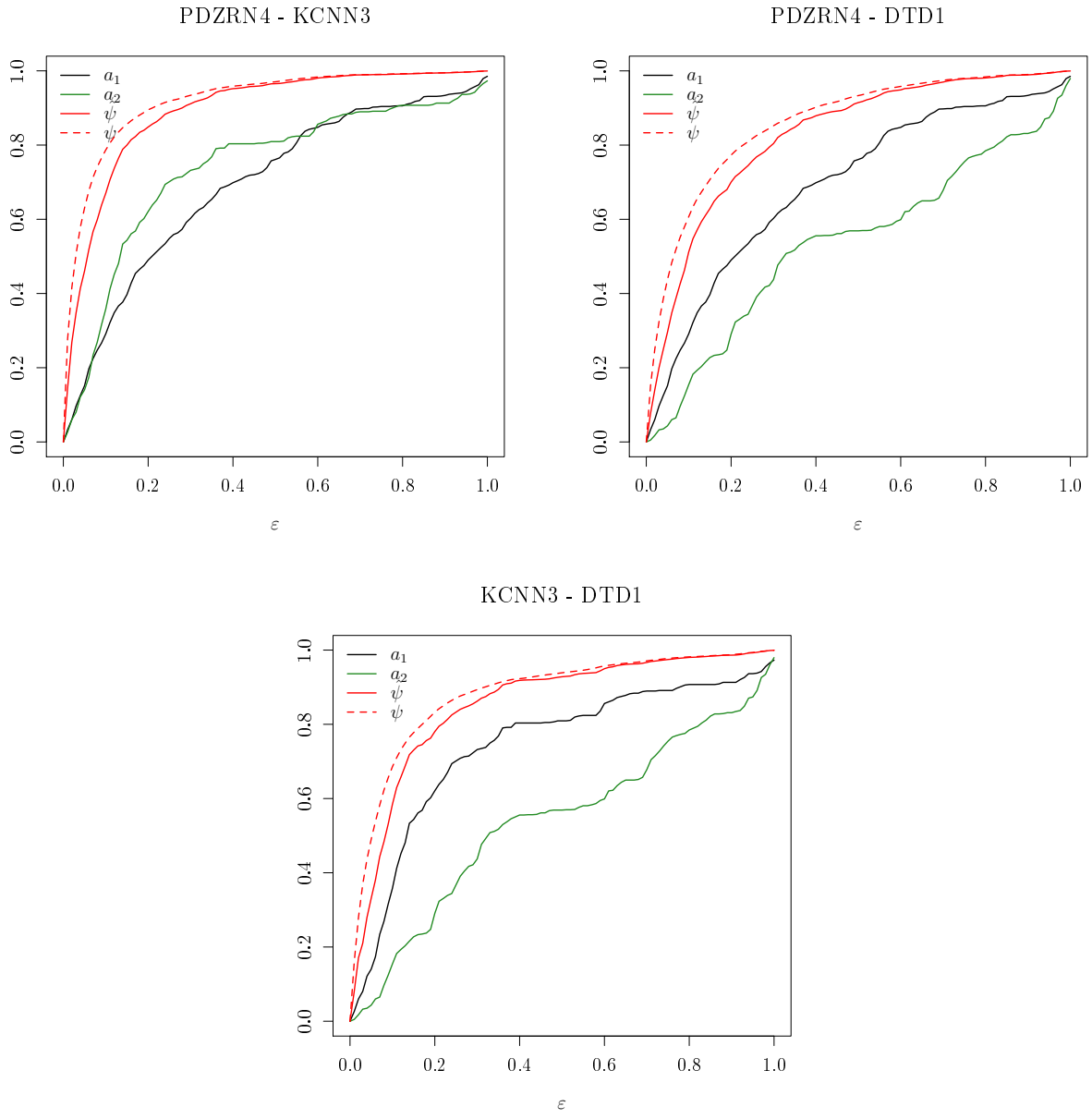


Figure 4.2: Proportion $\psi$ of coefficients of $\mathbf{\Sigma}$ lower than $\varepsilon$ in absolute value compared to proportions $a_1$ and $a_2$ of coefficients of $\mathbf{\Sigma_1}$ and $\mathbf{\Sigma_2}$ lower than $\varepsilon$ in absolute value considering three different pairs of genes (see Section 4.1 in Chapter 2)

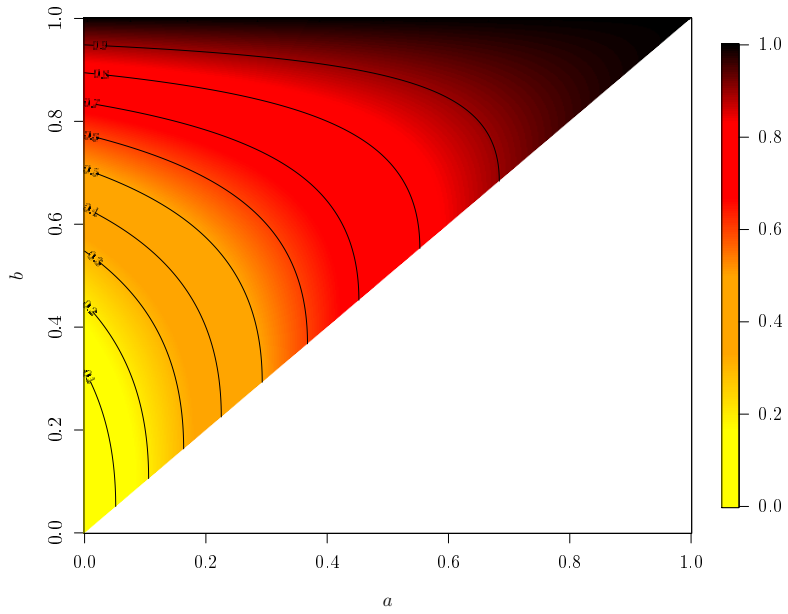


Figure 4.3: Heatmap of the lower bound $\hat{\psi}$ as a function of $a$ and $b$

The Wald statistic for the significance test of the interaction effect is then considered, defined as $T^{(i,j)} = \hat{\beta}_3^{(i,j)} / \sqrt{\widehat{\text{var}}(\hat{\beta}_3^{(i,j)})}$. Finally, the vector of size $p_1 p_2$ containing all the $T^{(i,j)}$ statistics for all pairs of SNPs is constructed, and a minP procedure is applied to it. If no marginal genetic effects are present, this method is expected to have similar performances to those of the minP procedure applied to the vector of score statistics $\boldsymbol{\mathcal{Z}}$ introduced in Section 2.1. Indeed, univariate Wald statistics are asymptotically equivalent to score statistics, and for both methods, the global statistic is defined as the greatest pointwise test statistic. The main difference is that the vector of score statistics $\boldsymbol{\mathcal{Z}}$ is constructed by taking account of the marginal effects of all SNPs at once, whereas the Aggregator method takes account of the marginal effects of each pair of SNPs one by one in the computation of the Wald statistic. Therefore, the performances could be different in presence of marginal effects.

## 4.1 Simulation procedure

The simulation procedure used to compare the tests is as follows. We first consider two genes; for each, a matrix of 100,000 rows is generated, the columns having the same dependence structure and marginal distributions as the observed gene. For a profile $\boldsymbol{X^{(1)}}$ corresponding to the first gene and a profile $\boldsymbol{X^{(2)}}$ corresponding to the second gene, the phenotype is then generated according to the following model:

$$\text{logit}(\mathbb{P}[Y = 1|\boldsymbol{X^{(1)}} = \boldsymbol{x^{(1)}}, \boldsymbol{X^{(2)}} = \boldsymbol{x^{(2)}}]) = \mu + \boldsymbol{x^{(1)\prime}}\boldsymbol{\alpha} + \boldsymbol{x^{(2)\prime}}\boldsymbol{\beta} + \boldsymbol{s}'\boldsymbol{\gamma}$$

where $\boldsymbol{S}$ is the interaction profile corresponding to $\boldsymbol{X^{(1)}}$ and $\boldsymbol{X^{(2)}}$. $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are the vectors of parameters for the marginal effects corresponding to genes 1 and 2, respectively, and $\boldsymbol{\gamma}$ is the vector of interaction effects parameters. The PDZRN4 and KCCN3 genes previously introduced will be used in this simulation study.

First, to assess the control of the type I error rate of the MGF-R test in the context of gene - gene interaction testing, $\boldsymbol{\gamma}$ is naturally set to $\mathbf{0}$. Several different values for $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ will be considered to ensure that the type I error rate is controlled in the presence of marginal effects. 1,000 samples containing 1,000 cases and 1,000 controls will be generated under each scenario to estimate the type I error rate.

To estimate the power of the methods under the alternative hypothesis, no marginal effects will be considered, *i.e.* $\boldsymbol{\alpha} = \boldsymbol{\beta} = \mathbf{0}$, as in Larson et al. (2014); Emily (2016). Therefore, the interaction profile can be rewritten as:

$$\boldsymbol{s} = (s_{1,1}, \ldots, s_{1,p_2}, \ldots, s_{i,1}, \ldots, s_{i,p_2}, \ldots, s_{p_1,1}, \ldots, s_{p_1,p_2})'$$

where $s_{i,j} = x_i^{(1)} x_j^{(2)}$, $x_k^{(\ell)}$ being the value of the $k$-th SNP of the $\ell$-th gene. $\boldsymbol{\gamma}$ can be similarly rewritten, $\gamma_{i,j}$ being the interaction parameter corresponding to the $i$-th SNP of gene 1 and the $j$-th SNP of gene 2. Two sets of causative SNPs are then selected, one from each gene, denoted $\mathcal{I} = \{i_1, \ldots, i_K\}$ and $\mathcal{J} = \{j_1, \ldots, j_K\}$, respectively (a given SNP can appear several times in a set). The sets contain the same number $K$ of causal SNPs, so that the $\ell$-th SNP $i_\ell$ from $\mathcal{I}$ has an interaction effect with the $\ell$-th SNP $j_\ell$ from $\mathcal{J}$. Let us denote $\mathcal{D} = \{(i_1, j_1), \ldots, (i_K, j_K)\}$ the set containing the pairs of the SNPs having an interaction effect. The simulation model can thus be rewritten as:

$$\text{logit}(\mathbb{P}[Y = 1|\boldsymbol{X^{(1)}} = \boldsymbol{x^{(1)}}, \boldsymbol{X^{(2)}} = \boldsymbol{x^{(2)}}]) = \mu + \sum_{(i,j)\in\mathcal{D}} \gamma_{i,j} s_{i,j}.$$

Once the phenotype is simulated, a sample is constructed by randomly sampling 1,000 cases and 1,000 controls. For a given scenario (*i.e.* two sets $\mathcal{I}$ and $\mathcal{J}$ of causal SNPs and corresponding coefficients $\gamma_{i,j}$), this process is repeated to obtain 500 samples, which are then used to estimate the power of each test. The simulation scenarios considered in the following are summarized in Table 4.1.

Table 4.1: Simulation scenarios (causal pairs and corresponding effects) for genes PDZRN4 and KCNN3 (scenarios 1 and 2) and for genes PDZRN4 and DTD1 (scenarios 3 and 4)

| Scenario | $\mathcal{I}$ | $\mathcal{J}$ | Effects |
|---|---|---|---|
| 1 | {10,23,30} | {3,18,7} | $\gamma_{10,3} = 0.6$, $\gamma_{23,18} = 0.2$, $\gamma_{30,7} = -0.6$ |
| 2 | {10,20,30} | {5,10,15} | $\gamma_{10,5} = 0.4$, $\gamma_{20,10} = 0.4$, $\gamma_{30,15} = 0.4$ |
| 3 | {5,25} | {15,16} | $\gamma_{5,15} = 0.5$, $\gamma_{25,16} = 0.5$ |
| 4 | {10,30} | {5,8} | $\gamma_{10,5} = 0.3$, $\gamma_{30,8} = 0.3$ |

## 4.2 Results

### 4.2.1 Assessment of the type I error rate

The empirical type I error rates obtained under several scenarios are given in Table 4.2. In any situation and for several values of the nominal level $\alpha$, the MGF-R test accurately controls the type I error rate. Despite the great dimension of the test statistics vector and the usage of the estimation method previously introduced for the correlation matrix, the MGF-R test remains valid in the gene - gene interaction testing context.

### 4.2.2 Power study

The results corresponding to the scenarios mentioned in Table 4.1 are displayed on Figure 4.4. It can be observed that, depending on the scenario, the different tests perform variably. We first compare the global tests that were introduced in Chapter 2; the PCA, CLD, Aggregator and CCA methods will then be included in

Table 4.2: Estimated type I error rates of the MGF-R test under several scenarios for gene - gene interaction testing (confidence intervals between brackets)

| $\boldsymbol{\alpha}$ | $\boldsymbol{\beta}$ | $\alpha$ | Type I error rate |
|---|---|---|---|
| **0** | **0** | 0.01 | 0.015 [0.007, 0.023] |
| | | 0.05 | 0.057 [0.043, 0.071] |
| | | 0.1 | 0.102 [0.083, 0.121] |
| $\alpha_{10} = 0.5$, $\alpha_{20} = 0.5$ | **0** | 0.01 | 0.013 [0.006, 0.020] |
| | | 0.05 | 0.053 [0.039, 0.067] |
| | | 0.1 | 0.103 [0.084, 0.122] |
| **0** | $\beta_{10} = 0.5$, $\beta_{20} = 0.5$ | 0.01 | 0.010 [0.004, 0.016] |
| | | 0.05 | 0.055 [0.041, 0.069] |
| | | 0.1 | 0.106 [0.087, 0.125] |
| $\alpha_{10} = 0.5$, $\alpha_{20} = 0.5$ | $\beta_{10} = 0.5$, $\beta_{20} = 0.5$ | 0.01 | 0.015 [0.007, 0.023] |
| | | 0.05 | 0.058 [0.044, 0.072] |
| | | 0.1 | 0.114 [0.094, 0.134] |

the comparison. Under scenario 1, the minP method is the least powerful, whereas Hotelling's test and the MGF-R test are the most powerful ones. The $L^2$-norm test and HC are equally powerful, between minP and the MGF-R test. Under scenario 2, minP and Hotelling's test are the least performing methods, whereas the three other tests are the most powerful ones. Under scenario 3, the $L^2$-norm test has obviously the weakest performance. The minP method is the most powerful one, whereas the MGF-R test is slightly less powerful, and performs similarly to HC. Finally, under scenario 4, Hotelling's test is far less powerful than the other ones. The minP method is obviously more powerful than Hotelling's test, but approximately equivalently less powerful than HC and the $L^2$-norm test, which are clearly the most powerful ones, followed by the MGF-R test.

For any scenario, the CCA method appears to be completely powerless and the CLD method has very weak power. The power of the Aggregator method is the same as that of the minP method. As previously mentioned, both methods are expected to be close when no marginal genetic effects are present, as in the present simulation setting. However, the two methods are not scrictly equivalent

and could yield different results under different scenarios, namely if marginal effects were present. The performances of the PCA method can vary importantly from a situation to another. It is indeed the most powerful test under scenario 1, but has weak power under scenario 4, compared to other tests.

As in Chapter 2, the MGF-R test appears to perform well in a wide range of situations. However, its performances are seriously affected under scenario 3, where Hotelling's test is clearly the least powerful one. Furthermore, the higher criticism also seems to be quite robust and often performs well. The robustness of the HC might be explained by two main reasons. First, the optimality of the HC, both under independence (Donoho and Jin, 2004) or under weak dependence (Arias-Castro et al., 2011), is asymptotic with respect to the dimensionality of the test statistics vector. In the context of gene - gene interaction effects testing, the test statistics vector is of dimension $p_1 p_2$. Consequently, the dimension grows very quickly to several hundreds as the number of SNPs in each gene increases. Then, even if the correlation matrix of the test statistics vector has a very particular structure, it is much sparser than the correlation matrices of each gene. As a consequence, it is more likely to correspond to a weakly correlated design, under which optimality results were derived in Arias-Castro et al. (2011). Nevertheless, the results of Arias-Castro et al. (2011) were derived in the quite different analysis of variance context and cannot be directly tranposed to the present context.

# 5 Extension of the proposed model for other interaction effects

The generalized linear model proposed in this chapter is based on a continuous coding of the SNPs. Consequently, global tests based on this model are powerful for detecting interaction effects when the trend of the effects for a pair of SNPs is approximately linear with respect to the total number of copies of the minor allele for the pair. Nevertheless, numerous other interaction effects can be assumed, whose trend can be highly nonlinear (Li and Reich, 2000; Li et al., 2015; Emily, 2016). Under such effects, the test statistics vector previously introduced might not be able to provide sufficient power.

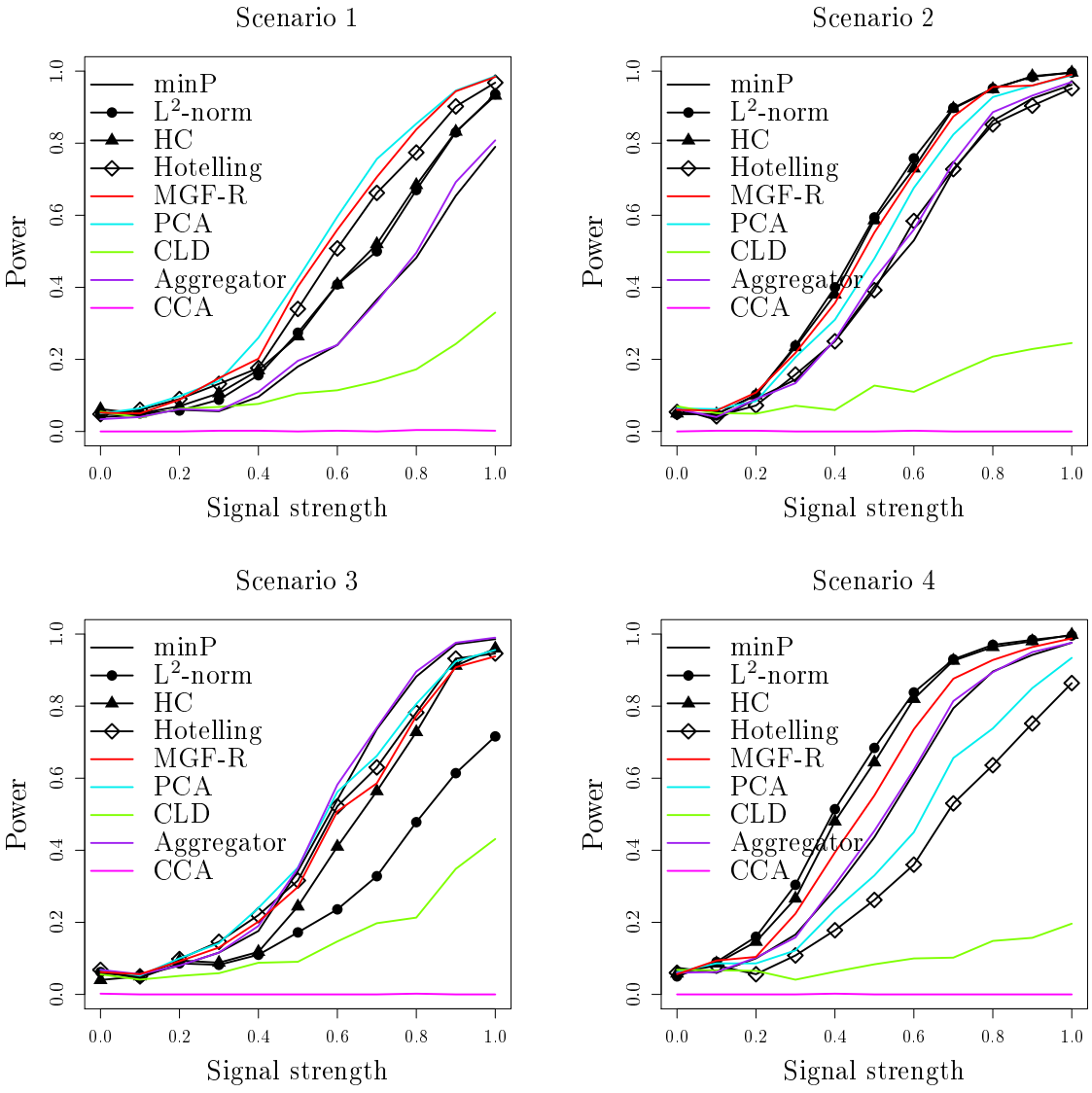


Figure 4.4: Power curves for the detection of gene - gene interaction effects under scenarios 1 to 4 (see Table 4.1)

## 5.1 Dummy coding - based SNP - SNP interaction model

In this chapter, a generalized linear model for the detection of gene - gene interaction effects was proposed. It aimed at extending the SNP - SNP interaction model (4.3) which was written as:

$$h(\mathbb{E}[Y|X_1 = x_1, X_2 = x_2]) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2.$$

Since this model is based on a continuous coding of the SNPs, the model proposed in this chapter is based on the same coding. However, the following model was also used to study SNP - SNP interaction effects (Yu et al., 2015; Ueki and Cordell, 2012; Ueki, 2014; Emily, 2012):

$$\begin{aligned} h(\mathbb{E}[Y|X_1 = x_1, X_2 = x_2]) &= \alpha + \sum_{k=1}^{2} \beta_k \mathbf{1}_{\{x_1=k\}} + \sum_{\ell=1}^{2} \gamma_\ell \mathbf{1}_{\{x_2=\ell\}} \\ &\quad + \sum_{k=1}^{2}\sum_{\ell=1}^{2} \delta_{k\ell} \mathbf{1}_{\{x_1=k, x_2=\ell\}}. \end{aligned} \tag{4.4}$$

These two models are compared in VanderWeele and Laird (2011), and intermediate models are also studied. The first one is simpler by needing the estimation of only four parameters, whereas the second one requires the estimation of nine parameters. However, the second one is more flexible, and would allow to detect a wider range of types of interaction effects than the first. Indeed, using a continuous coding as in model (4.3) enables to construct very powerful tests if the effect has an approximately linear trend with respect to the total number of copies of the minor allele for the two SNPs considered in the model. On the other hand, if the trend is far from linear, using a dummy coding as in model (4.4) can yield greater power. A major difference between these models is that a significance test for the interaction effect is based on a statistic with one degree of freedom under the first model, against four degrees of freedom under the second. Therefore, as an extension to the global model introduced in this chapter, a global model aiming at extending model (4.4) can also be introduced.

## 5.2 Construction of a dummy coding - based gene - gene interaction model

To extend model (4.4), we first consider the dummy coding versions of $\boldsymbol{X^{(1)}}$ and $\boldsymbol{X^{(2)}}$:

$$\boldsymbol{A^{(\ell)}} = (\mathbf{1}_{\{X_1^{(\ell)}=1\}}, \ldots, \mathbf{1}_{\{X_j^{(\ell)}=1\}}, \ldots, \mathbf{1}_{\{X_{p_\ell}^{(1)}=1\}})',$$
$$\boldsymbol{B^{(\ell)}} = (\mathbf{1}_{\{X_1^{(\ell)}=2\}}, \ldots, \mathbf{1}_{\{X_j^{(\ell)}=2\}}, \ldots, \mathbf{1}_{\{X_{p_\ell}^{(1)}=2\}})',$$

where $\ell = 1, 2$ denotes the gene number. Similarly as for the construction of model (4.1), model (4.4) can be extended at the gene scale by using the following $(4p_1p_2)$-dimensional vector:

$$\boldsymbol{S} = \begin{pmatrix} \boldsymbol{A^{(1)}} \\ \boldsymbol{B^{(1)}} \end{pmatrix} \otimes \begin{pmatrix} \boldsymbol{A^{(2)}} \\ \boldsymbol{B^{(2)}} \end{pmatrix},$$

which yields the following model:

$$\begin{aligned} h(\mathbb{E}[Y|\boldsymbol{U}=\boldsymbol{u}, \boldsymbol{X^{(1)}}=\boldsymbol{x^{(1)}}, \boldsymbol{X^{(2)}}=\boldsymbol{x^{(2)}}]) &= \boldsymbol{u}'\boldsymbol{\beta_U} + (\boldsymbol{a^{(1)\prime}}, \boldsymbol{b^{(1)\prime}})'\boldsymbol{\beta_{(1)}} \\ &\quad + (\boldsymbol{a^{(2)\prime}}, \boldsymbol{b^{(2)\prime}})'\boldsymbol{\beta_{(2)}} + \boldsymbol{s}'\boldsymbol{\beta_S}. \end{aligned} \tag{4.5}$$

By denoting $\boldsymbol{W} = (\boldsymbol{U}', \boldsymbol{A^{(1)\prime}}, \boldsymbol{B^{(1)\prime}}, \boldsymbol{A^{(2)\prime}}, \boldsymbol{B^{(2)\prime}})'$, $\boldsymbol{\alpha} = (\boldsymbol{\beta}_{\boldsymbol{U}}', \boldsymbol{\beta}_{\boldsymbol{(1)}}', \boldsymbol{\beta}_{\boldsymbol{(2)}}')'$ and $\boldsymbol{\beta} = \boldsymbol{\beta_S}$, this model can be rewritten as:

$$h(\mathbb{E}[Y|\boldsymbol{X^{(1)}}=\boldsymbol{x^{(1)}}, \boldsymbol{X^{(2)}}=\boldsymbol{x^{(2)}}, \boldsymbol{U}=\boldsymbol{u}]) = \boldsymbol{w}'\boldsymbol{\alpha} + \boldsymbol{s}'\boldsymbol{\beta}.$$

The test statistics vector $\boldsymbol{\mathcal{Z}}$ for testing $H_0 : \boldsymbol{\beta} = \mathbf{0}$ is then straightforwardly obtained by adapting the procedure used under model (4.1). We first introduce the matrix $\mathbb{W}$ whose $i$-th row is obtained by concatening the covariates profile and the genotypic profiles of the $i$-th individual:

$$\mathbb{W} = \left(\begin{array}{ccc|ccc|ccc|ccc|ccc} u_{11} & \ldots & u_{1q} & a_{11}^{(1)} & \ldots & a_{1p_1}^{(1)} & b_{11}^{(1)} & \ldots & b_{1p_1}^{(1)} & a_{11}^{(2)} & \ldots & a_{1p_2}^{(2)} & b_{11}^{(2)} & \ldots & b_{1p_2}^{(2)} \\ \vdots & & \vdots & \vdots & & \vdots & \vdots & & \vdots & \vdots & & \vdots & \vdots & & \vdots \\ u_{i1} & \ldots & u_{iq} & a_{i1}^{(1)} & \ldots & a_{ip_1}^{(1)} & b_{i1}^{(1)} & \ldots & b_{ip_1}^{(1)} & a_{i1}^{(2)} & \ldots & a_{ip_2}^{(2)} & b_{i1}^{(2)} & \ldots & b_{ip_2}^{(2)} \\ \vdots & & \vdots & \vdots & & \vdots & \vdots & & \vdots & \vdots & & \vdots & \vdots & & \vdots \\ u_{n1} & \ldots & u_{nq} & a_{n1}^{(1)} & \ldots & a_{np_1}^{(1)} & b_{n1}^{(1)} & \ldots & b_{np_1}^{(1)} & a_{n1}^{(2)} & \ldots & a_{np_2}^{(2)} & b_{n1}^{(2)} & \ldots & b_{np_2}^{(2)} \end{array}\right)$$

$$= [\, \mathbb{U} \mid \mathbb{A}^{(1)} \mid \mathbb{B}^{(1)} \mid \mathbb{A}^{(2)} \mid \mathbb{B}^{(2)} \,]$$

where $a_{ij}^{(\ell)} = \mathbf{1}_{\{x_{ij}^{(\ell)}=1\}}$ and $b_{ij}^{(\ell)} = \mathbf{1}_{\{x_{ij}^{(\ell)}=2\}}$. Then, the matrix $\mathbb{S}$ of interaction profiles

is introduced:

$$\mathbb{S} = \begin{pmatrix} (\boldsymbol{a_1^{(1)\prime}}, \boldsymbol{b_1^{(1)\prime}}) \otimes (\boldsymbol{a_1^{(2)\prime}}, \boldsymbol{b_1^{(2)\prime}}) \\ \vdots \\ (\boldsymbol{a_i^{(1)\prime}}, \boldsymbol{b_i^{(1)\prime}}) \otimes (\boldsymbol{a_i^{(2)\prime}}, \boldsymbol{b_i^{(2)\prime}}) \\ \vdots \\ (\boldsymbol{a_n^{(1)\prime}}, \boldsymbol{b_n^{(1)\prime}}) \otimes (\boldsymbol{a_n^{(2)\prime}}, \boldsymbol{b_n^{(2)\prime}}) \end{pmatrix}$$

where $\boldsymbol{a_i^{(\ell)}} = (a_{i1}^{(\ell)}, \ldots, a_{ip_\ell}^{(\ell)})'$ and $\boldsymbol{b_i^{(\ell)}} = (b_{i1}^{(\ell)}, \ldots, b_{ip_\ell}^{(\ell)})'$.

To estimate the marginal effects parameters in model (4.1), the marginal profiles were replaced by the corresponding principal components profiles (see Section 2.3). Similarly, to estimate the marginal effects parameters in model (4.5) the four matrices $\mathbb{A}^{(\ell)}$ and $\mathbb{B}^{(\ell)}$ are replaced by constructing the corresponding matrices of principal components profiles.

## 5.3 Correlation structure of the test statistics vector

We now examine the estimated correlation matrix of the test statistics vector using the dummy coding to ensure that it admits a similar decomposition as that of the test statistics vector using the continuous coding. As previously, we display the estimated correlation matrices of two genes and the estimated correlation matrix of the corresponding gene - gene interaction test statistics vector on Figure 4.5. The genes used are the same as those used for Figure 4.1. However, the correlation matrices of the dummy variables of each gene are considered here. Judging by this example, the correlation matrix of the dummy coding - based test statistics vector seems to admit a Kronecker decomposition, like the correlation matrix of the continuous coding - based test statistics vector: $\widehat{\boldsymbol{\Sigma}} = \widehat{\boldsymbol{\Sigma}}_{\mathbf{1}} \otimes \widehat{\boldsymbol{\Sigma}}_{\mathbf{2}}$ where $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{i}}$ is the correlation matrix of the dummy variables constructed from the SNPs of gene $i$. Therefore, this decomposition provides the same advantage if the eigendecomposition or the inverse of $\widehat{\boldsymbol{\Sigma}}$ is needed. The advantage is even greater here, since $\widehat{\boldsymbol{\Sigma}}$ is a $(4p_1p_2) \times (4p_1p_2)$ matrix, whereas it is a $(p_1p_2) \times (p_1p_2)$ matrix when using the continuous coding.

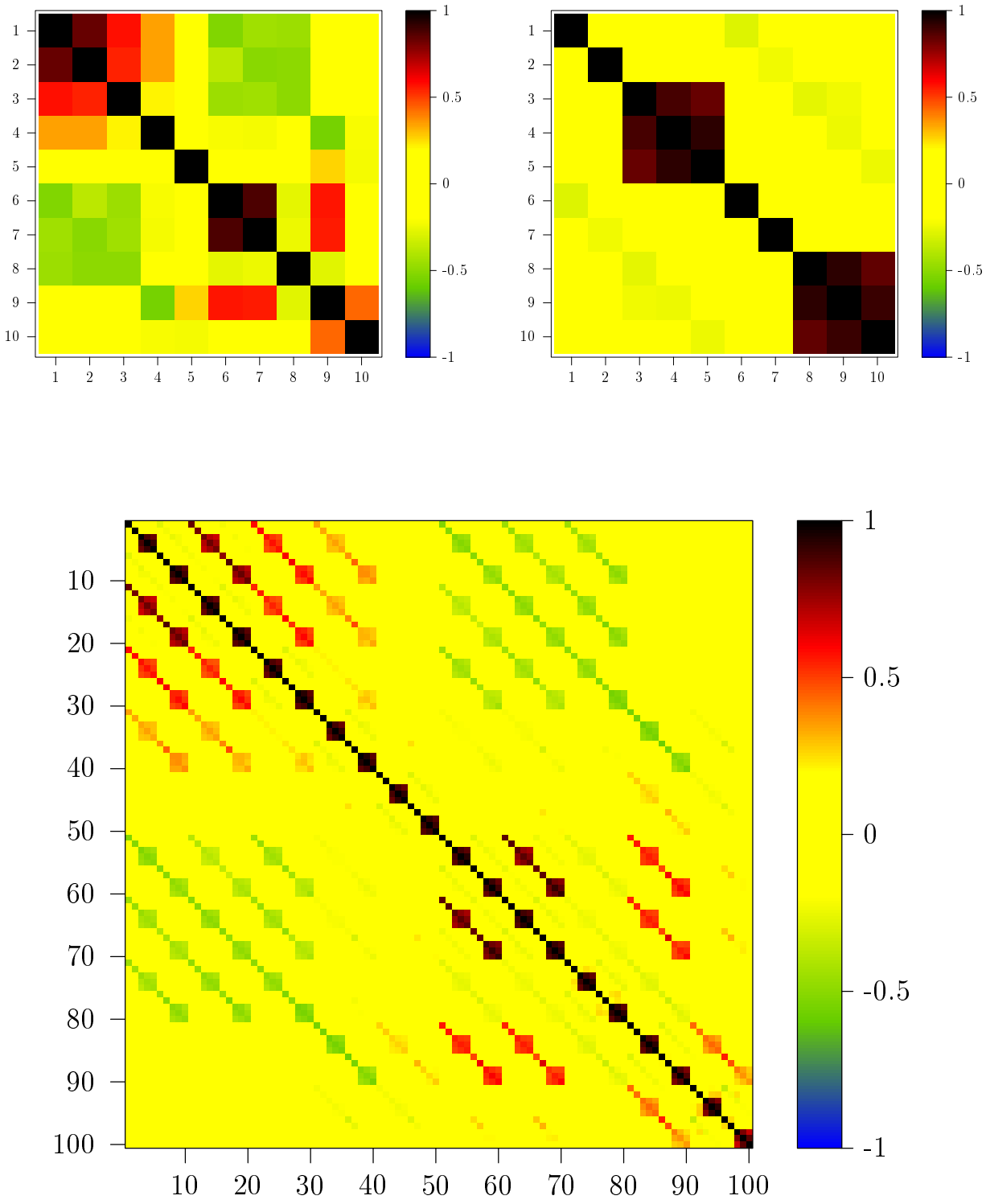


Figure 4.5: Upper part: correlation matrices of 5 SNPs from genes PDZRN4 and KCNN3. Lower part: corresponding correlation matrix for the test statistics vector (dummy coding)

## 5.4 Comparison of the continuous and dummy codings

### 5.4.1 Power comparison of the continuous and dummy codings

We now compare the minP, HC, $L^2$-norm, Hotelling and MGF-R tests using the continuous or the dummy coding - based test statistics vectors. First, we assess the performances of these tests using the dummy coding under scenario 1 in Table 4.1. The values of the coefficients are multiplied by 2 so that the most powerful methods reach a power of 1 for the maximal signal strength. The results are displayed on Figure 4.6. Even if the gaps between the curves are clearly greater than they were when using the continuous coding (see Figure 4.4), the ranking of the methods is the same on both figures. However, for a given method, the continuous coding yields greater power. Indeed, the maximal values of the coefficients had to be multiplied by 2 to reach a power of 1 with the dummy coding. This was expected, since the trend of the interaction effects is completely linear. Indeed, using the continuous coding can be seen as assuming that the trend is linear. If the assumption is fulfilled, this coding is likely to yield greater power than the dummy coding.

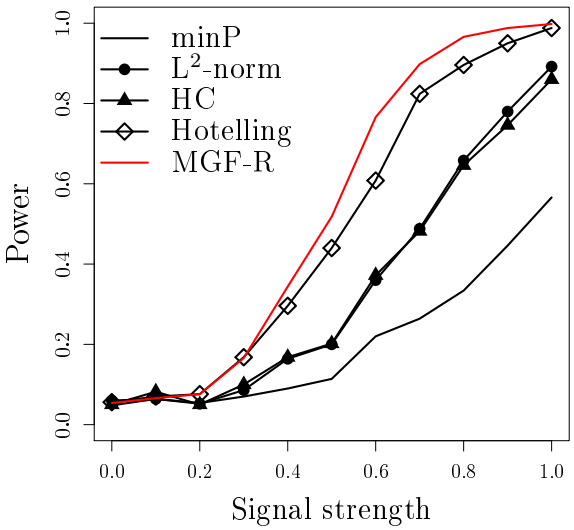


Figure 4.6: Power curves under scenario 1 (see Table 4.1) using the dummy coding - based tests

The dummy coding - based vector is naturally disadvantaged because of its

dimension, which is four times that of the continuous coding - based vector. However, for more general interaction effects, if the trend is far from linear, a test based on the continuous coding can be completely unable to detect the effect. On the contrary, since no specific assumption is made on the interaction effect when using the dummy coding, a test based on this coding might be more powerful. To illustrate this, a simulation scenario is considered by applying several modifications to simulation scenario 1, in order to generate phenotypes under a recessive - recessive interaction model (see the supplementary material of Emily (2016)). First, the simulation model previously used is modified. The phenotype was previously generated according to the following model:

$$\text{logit}(\mathbb{P}[Y = 1|\boldsymbol{X^{(1)}} = \boldsymbol{x^{(1)}}, \boldsymbol{X^{(2)}} = \boldsymbol{x^{(2)}}]) = \mu + \sum_{(i,j)\in\mathcal{D}} \gamma_{i,j} s_{i,j}$$

where $s_{ij} = x_i^{(1)} x_j^{(2)}$ (see Section 4.1). We now consider the same simulation model, but with $s_{ij} = x_i^{(1)} x_j^{(2)} \mathbf{1}_{\{x_i^{(1)}=2, x_j^{(2)}=2\}}$; the disease risk of an individual is affected only if, for a given causative pair, two copies of the minor allele are carried for both SNPs. The values of the coefficients used in scenario 1 are multiplied by 3 to obtain a maximal power close to 1 for the most powerful methods. The results for the minP, HC, $L^2$-norm, Hotelling and MGF-R tests with both codings are represented on figure 4.7. For each test, the dummy coding - based version is clearly more powerful than the continuous coding - based version. The gain in power for the HC, $L^2$-norm and MGF-R tests is clearly important and demonstrates the interest of considering the dummy coding in gene - gene interaction testing. Quite surprisingly, Hotelling's test is almost the most powerful among the continuous coding - based tests, whereas it is the least powerful one among the dummy coding - based tests. It is also quite surprising to note that HC and the $L^2$-norm tests, which are clearly less powerful than Hotelling's test when using the continuous coding, are the most powerful ones when using the dummy coding. For a given coding, the MGF-R test is among the most powerful ones.

### 5.4.2 Omnibus testing method

The former results indicate that depending on the true interaction effects, the power obtained by one global testing method can be very different when using one vector

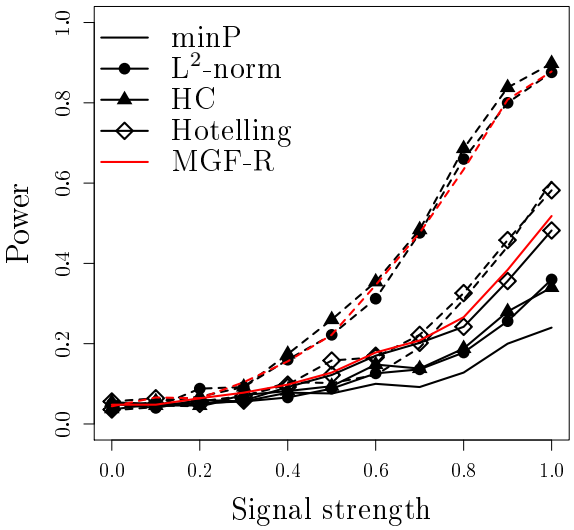


Figure 4.7: Power curves under a simulation scenario involving recessive - recessive effects; solid lines correspond to the continuous coding, dashed lines to the dummy coding

or the other. Without a priori knowledge of these effects, choosing one vector or the other could yield only very low power. To counteract this problem, an omnibus combining method based on both vectors can be considered. First, let $\boldsymbol{q}$ be the vector containing the $p$-values obtained by some chosen global tests, and $\boldsymbol{q_k^{(0)}}$, $1 \leq k \leq N$ the vectors of $p$-values obtained by the same global tests using simulated phenotypes under the null hypothesis, where $N$ is the number of simulated phenotypes (*e.g.* $N = 1,000$). Then, let $\varphi$ be a function, which takes as argument the vector $\boldsymbol{q}$ and returns an associated combined $p$-value. The final $p$-value of the procedure is defined as:

$$p = \frac{1}{N} \sum_{k=1}^{N} \mathbf{1}_{\{\varphi(\boldsymbol{q_k^{(0)}}) \leq \varphi(\boldsymbol{q})\}}.$$

In the following, we define $\varphi$ as the Simes' combining method (Simes, 1986):

$$\varphi(\boldsymbol{q}) = \min_{1 \leq i \leq r} \frac{r q_{(i)}}{i}$$

where $r$ is the number of tests to be combined (*i.e.* the number of coordinates of $\boldsymbol{q}$) and $q_{(i)}$ is the $i$-th order statistic of $\boldsymbol{q}$. To reduce the computational time, the $p$-values (or corresponding statistics) simulated under the null hypothesis to obtain

the global $p$-values at the first stage are also used at the second stage in the vectors $\boldsymbol{q}_{\boldsymbol{k}}^{(\mathbf{0})}$ to compute the final $p$-value.

We propose to consider an omnibus method based on the two versions of the minP, $L^2$-norm, HC, and Hotelling's tests, and an omnibus method based on the two versions of the MGF-R test. The results under the scenario 1 previously considered (see Table 4.1), using additive - additive effects first, and recessive - recessive effects then, are represented on Figure 4.8. As previously seen, when additive - additive effects are considered, tests based on the continuous coding are much more powerful. On the other hand, they are clearly underpowered when recessive - recessive effects are considered. It can be seen that in both situations, the omnibus combining methods are very powerful. They are naturally not totally as powerful as the best methods for each type of effect, but maintain great power for both types of effects. This demonstrates that combining several global tests based on the two vectors introduced in this chapter can be a suitable approach to detect a wide range of interaction effects.

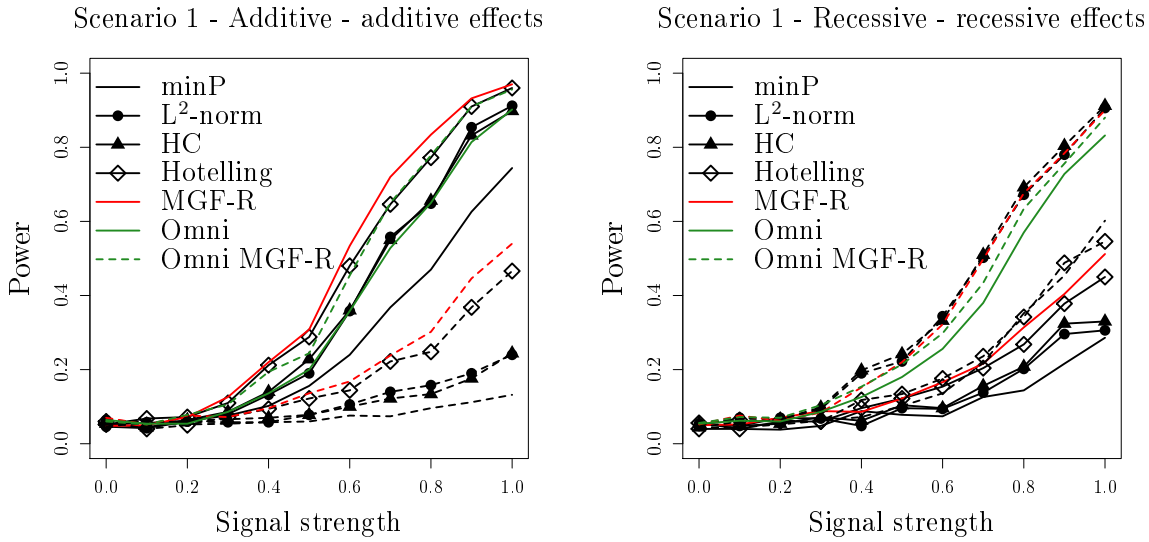


Figure 4.8: Power curves for the omnibus combining methods under scenario 1 (with additive - additive effects on the left, recessive - recessive effects on the right). For black and red lines, solid lines correspond to the continuous coding, dashed lines to the dummy coding. "Omni" stands for the omnibus test based on the two versions of the minP, $L^2$-norm, HC, and Hotelling's tests, and "Omni MGF-R" stands for the omnibus test based on the two versions of the MGF-R test

# 6 Discussion

In this chapter, a suitable framework for the detection of gene - gene interaction effects was introduced. A test statistics vector was then derived, on which global testing methods compared in Chapter 2 could be applied. The global model used in Chapter 2 is naturally extended to the gene - gene interaction detection context. However, the straightforward extension of the former model to the gene - gene interaction detection context was not previously proposed, even if a similar framework was considered in the construction of the Aggregator method (Emily, 2018). Nevertheless, Emily (2018) proposed to consider each possible pair of SNPs, one at a time, whereas the model proposed in this chapter includes all pairs of SNPs at once.

Interestingly, in the particular context of this chapter, the correlation matrix of the test statistics vector has a very special structure. Dependence among test statistics is indeed inherited from the dependence of the two genes that are examinated. An estimation method was introduced using the Kronecker product of the correlation matrices corresponding to the two genes. Under the assumption that the genes are independent, this estimation seems rather reasonable and provides great computational gain. Nevertheless, the quality of this estimation might be impacted if non-negligible dependence is present between the genes.

Based on this approximation, the MGF-R test introduced in Chapter 2 was adapted to the gene - gene interaction context by taking advantage of the former approximation of the correlation matrix. As in Chapter 2, the test was shown to perform well in diverse situations. However, the higher criticism also often performs well. This is most likely due to the large dimension of the test statistics vector. Indeed, the optimality of the higher criticism, under independence (Donoho and Jin, 2004) or under weak dependence (Arias-Castro et al., 2011), is asymptotic with respect to the dimension of the test statistics vector.

The test statistics vector introduced in this chapter was derived using a continuous coding of the SNPs first. This coding is often used, namely for computational and modelization simplicity reasons (Li et al., 2009a; Emily, 2016; Cordell et al., 2001; Cordell, 2002, 2009a; Ueki and Cordell, 2012). Nevertheless, for interaction effects with a nonlinear trend, using this test statististics vector might yield low

power. Therefore, another test statistics vector was derived as a possible extension, using a categorical coding of the SNPs (Yu et al., 2015; Ueki and Cordell, 2012; Ueki, 2014; Emily, 2012). Nevertheless, it seems difficult to choose which vector to use without a priori knowledge about the effects. A feasible approach to account both for dependence among test statistics and for the possible diversity of interaction effects would be to compute both vectors and apply several global tests on both. Then, the global $p$-values would be combined in a single one. It was demonstrated that this approach would maintain great power for very diverse types of interaction effects. However, this raises the question of the choice of the combining strategy for the combination of the two global tests. For instance, Simes' method (Simes, 1986) or Fisher's method (Fisher, 1925) could be used, among numerous alternatives.

# Chapter 5

# Adaptive Handling of Dependence for Regression Modeling

**Abstract.** How to handle dependence in large or high dimensional regression modeling remains an open issue. The strong belief issuing from our complete knowledge of well-proven methods in low-dimensional settings that dependence should not be ignored is not necessarily true in high dimension. To address this point, we introduce a new class of prediction scores defined as linear combinations of a same random vector. This includes the naive regression score designed to be optimal under independence across explanatory variables and the Ordinary Least Squares regression score that, on the contrary, fully accounts for dependence by a preliminary whitening of the explanatory variables, but also Ridge and Partial Least Squares regression scores, offering intermediate ways of dealing with dependence. The former general framework enables a theoretical comparative study of the prediction performance of the former methods. The main general conclusion of this comparative study is that the best handling of dependence depends on the interplay between the structure of conditional dependence across explanatory variables and the pattern of the association signal. We also derive the closed form expression of the prediction score with the best prediction performance within the proposed class, leading to an adaptive handling of dependence. Finally, it is demonstrated through simulation studies and using benchmark datasets that this predictor outperforms existing methods in various settings.

## Contents


**1 Introduction . . . . . 131**

**2 Optimal versus naive linear prediction . . . . . 134**

2.1 Optimal linear prediction . . . . . 135

2.2 Naive linear prediction . . . . . 136

2.3 Prediction performance of naive linear prediction . . . . . 138

2.4 Ordinary Least Squares and naive linear prediction in high dimension . . . . . 139

2.5 Illustrative comparative study in high dimension . . . . . 140

**3 A new class of prediction scores . . . . . 144**

3.1 Introduction of a new class of prediction scores . . . . . 144

3.2 $\mathcal{L}$ contains Ridge predictions scores . . . . . 145

3.3 $\mathcal{L}$ contains Partial Least Squares (PLS) prediction scores . 146

3.4 Alternative prediction scores . . . . . 147

**4 Optimal prediction within $\mathcal{L}$ . . . . . 149**

4.1 Closed-form expression of the best predictor . . . . . 150

4.2 Estimation of the optimal predictor within $\mathcal{L}$ . . . . . 151

**5 Comparative study . . . . . 154**

5.1 Simulation study . . . . . 154

5.2 Performance comparisons on datasets . . . . . 157

**6 Discussion . . . . . 160**

**7 Appendix: proof of Theorem 2.1 . . . . . 161**

**8 Appendix: proof of Theorem 3.1 . . . . . 163**

**9 Appendix: Lemma 9.1 . . . . . 164**

**10 Appendix: proof of Theorem 4.1 . . . . . 166**

# 1 Introduction

Regression modeling in a prediction perspective is well-studied and proven in standard designs where the number $n$ of observations exceeds the number $p$ of explanatory variables. Thus, it has created strong beliefs about good practice of regression rules that are yet not necessarily true in high-dimensional designs. The question of ignoring dependence or not when fitting prediction models perfectly illustrates this point. Let us consider the usual normal setting for regression model where the response $Y$ is real-valued, and let us assume that the explanatory variables $\boldsymbol{X} = (X_1, \ldots, X_p)'$ are uncorrelated, conditionally on the response. Then, the linear predictor with largest squared correlation with the response is just a linear combination of the coordinates of the vector $\tilde{\boldsymbol{X}} = \boldsymbol{D}_{\boldsymbol{\sigma}}^{-1}(\boldsymbol{X} - \boldsymbol{\mu}_{\boldsymbol{x}})$ of scaled explanatory variables, where $\boldsymbol{D}_{\boldsymbol{\sigma}}$ is the $p \times p$ diagonal matrix whose vector of diagonal entries is the $p$-vector $\boldsymbol{\sigma}$ of conditional standard deviations of the explanatory variables and $\boldsymbol{\mu}_{\boldsymbol{x}} = \mathbb{E}[\boldsymbol{X}]$:

$$L(\tilde{\boldsymbol{X}}) \quad = \quad \tilde{\boldsymbol{X}}' \tilde{\boldsymbol{\sigma}}_{\boldsymbol{xy}}, \tag{5.1}$$

where $\tilde{\boldsymbol{\sigma}}_{\boldsymbol{xy}}$ is the $p$-vector of covariances between $Y$ and $\tilde{\boldsymbol{X}}$. The former $\tilde{\boldsymbol{\sigma}}_{\boldsymbol{xy}}$ can be viewed as the vector of one-to-one association parameters between each of the scaled explanatory variables and the response.

In situations where the explanatory variables are no longer independent, with positive $p \times p$ conditional correlation matrix $\boldsymbol{C}$, then there exists matrices $\boldsymbol{W}$ such that $\boldsymbol{C}^{-1} = \boldsymbol{W}'\boldsymbol{W}$ or equivalently $\boldsymbol{W}\boldsymbol{C}\boldsymbol{W}' = \mathbf{I}_p$. Such matrices are either called decorrelation or whitening matrices, since the variance-covariance matrix of the whitened explanatory variables $\boldsymbol{Z} = \boldsymbol{W}\tilde{\boldsymbol{X}}$ is the identity matrix $\mathbf{I}_p$ (Kessy et al., 2018). The optimal linear predictor then takes the form of expression (5.1) where $\tilde{\boldsymbol{X}}$ is replaced by $\boldsymbol{Z}$:

$$L(\boldsymbol{Z}) \quad = \quad \boldsymbol{Z}' \boldsymbol{\sigma}_{\boldsymbol{zy}}, \tag{5.2}$$

where $\boldsymbol{\sigma}_{\boldsymbol{zy}} = \boldsymbol{W}\tilde{\boldsymbol{\sigma}}_{\boldsymbol{xy}}$ is the $p$-vector of covariances between $Y$ and $\boldsymbol{Z}$. Therefore, still in this purely probabilistic framework where the parameters of the joint distribution of $\boldsymbol{X}$ and $Y$ are supposed to be known, applying the optimal regression rule under independence on the whitened explanatory variables turns out to be the best way to handle dependence.

The same idea also holds for the two-group classification issue, where $Y$ can only take the values 0 and 1, under the standard assumption of a mixture of normal distributions for $\boldsymbol{X}$ with equal variance-covariance for the two components of the mixture. Indeed, if furthermore conditional independence across explanatory variables is assumed, the linear Bayes classifier, minimizing the probability of misclassification, is also an affine transformation of the linear prediction score $L(\tilde{\boldsymbol{X}}) = \tilde{\boldsymbol{X}}'\tilde{\boldsymbol{\delta}}_{\boldsymbol{xy}}$, where $\tilde{\boldsymbol{X}} = \boldsymbol{D}_{\boldsymbol{\sigma}}^{-1}(\boldsymbol{X} - \boldsymbol{\mu})$ with $\boldsymbol{\mu} = [\mathbb{E}(\boldsymbol{X} \mid Y = 1) + \mathbb{E}(\boldsymbol{X} \mid Y = 0)]/2$ and $\tilde{\boldsymbol{\delta}}_{\boldsymbol{xy}} = \mathbb{E}(\tilde{\boldsymbol{X}} \mid Y = 1) - \mathbb{E}(\tilde{\boldsymbol{X}} \mid Y = 0)$ stands for the vector of association parameters between the scaled explanatory variables and the response. Analogously with the situation of a real-valued response, if it is now assumed that the conditional variance-covariance of $\tilde{\boldsymbol{X}}$ is $\boldsymbol{C} \neq \mathbf{I}_p$, then the optimal linear score is, up to an affine transformation:

$$L(\boldsymbol{Z}) = \boldsymbol{Z}'\boldsymbol{\delta}_{\boldsymbol{zy}}, \tag{5.3}$$

with $\boldsymbol{Z} = \boldsymbol{W}\boldsymbol{X}$ and $\boldsymbol{\delta}_{\boldsymbol{zy}} = \boldsymbol{W}\tilde{\boldsymbol{\delta}}_{\boldsymbol{xy}}$, $\boldsymbol{W}'\boldsymbol{W} = \boldsymbol{C}^{-1}$. One general conclusion is that, when the explanatory variables are known to be mutually dependent, which is generally the case, optimal prediction requires to take explicitly into account this dependence and the suitable way to do that is by whitening the explanatory variables. Consistently, Ordinary Least Squares regression (OLS) and Linear Discriminant Analysis (LDA), which can both be viewed as empirical counterparts of (5.2) and (5.3) respectively, are gold standard methods for regression and classification.

In high-dimensional settings, plugging-in the sample estimate $\hat{\boldsymbol{C}}$ of $\boldsymbol{C}$ in expressions (5.2) and (5.3) is no longer possible, essentially because those expressions involve the inversion of $\boldsymbol{C}$, and $\hat{\boldsymbol{C}}$ is not full-rank. Since OLS and LDA prediction scores also result from the optimization of a least-squares goodness-of-fit criterion, a very popular and numerically efficient approach to circumvent this issue is to modify the objective function of the fitting algorithm by adding a regularization term. Least Absolute Shrinkage and Selection Operator (Lasso, Tibshirani (1996)) and Ridge regression (Hoerl and Kennard, 1970) or classification are based on this idea with a regularization term defined as proportional to the sum of absolute regression coefficients and the sum of squared regression coefficients, respectively. Many variants of the former regularized estimation methods exist, for example combining the two kinds of penalty terms or accounting for a group structure among variables. In a regression context with a real-valued response, we just mention here an alterna-

tive shrinkage-based regression method introduced by Opgen-Rhein and Strimmer (2007b) (see also Zuber and Strimmer (2011)), in which both the sample estimates of the covariance matrix and of the covariances between the explanatory variables and the response are replaced with James-Stein type estimates (Schäfer and Strimmer, 2005). A similar approach is introduced in the LDA framework by Ahdesmäki and Strimmer (2010). Still within the OLS and LDA framework, a rank-reduced estimation of the linear prediction score can also lead to excellent prediction performance. Partial Least Squares (PLS) regression (Wold et al. (1983, 1984)) and Discriminant Analysis (Boulesteix (2004); Gottfries et al. (1995); Barker and Rayens (2003)), consisting in a preliminary extraction of latent variables that summarize the explanatory variables in a low-dimensional kernel, are among the most widely used methods for high-dimensional regression and classification.

Consistently with the normal framework introduced at the beginning of the present section, the list of high-dimensional regression and classification methods above focuses on leading procedures estimating a linear regression score. However, the variety of approaches used to accommodate high-dimension, including tree-based prediction and neural networks, has generated a much larger list of possible methods, with unclear recommendations about which should be preferred in which circumstances (see Chapter 3 of Hastie et al. (2009) for a review of high-dimensional regression methods). Many comparative studies highlight the impact of the pattern of the vector of association parameters, namely its fraction of zeros, and the amount of correlation across explanatory variables on the prediction performance (see for example Tibshirani (1996); Chong and Jun (2005); Wold et al. (1984); Ahmad et al. (2006)). Despite some theoretical advances on those points, it is finally often recommended in practice to compare methods by cross-validated evaluation of their prediction performance (Krstajic et al., 2014). Moreover, most of these methods involve one or more hyper-parameters, either shrinkage parameters or numbers of latent variables, which optimization can be challenging and deter the prediction performance.

Essentially for classification issues in high dimension, naive options consisting in ignoring dependence have been tried in some of these comparative studies (Dudoit et al., 2002), with surprisingly good prediction performance. The so-called naive Bayes classifier is also compared with LDA on a theoretical basis in Bickel and

Levina (2004) and it turns out to show very good prediction performance in large or high dimensional settings. Starting from this observation, we introduce a similar naive linear regression score and demonstrate that both the OLS score and the naive prediction score are just different linear combinations of a same random vector $\boldsymbol{\xi}(\boldsymbol{Z})$. The class $\mathcal{L} = \{\boldsymbol{h}'\boldsymbol{\xi}(\boldsymbol{Z}),\ \boldsymbol{h}'\boldsymbol{h} = 1\}$ of all possible linear combinations of $\boldsymbol{\xi}(\boldsymbol{Z})$ is introduced for a more flexible handling of dependence. Since Ridge and PLS regression scores form subclasses of $\mathcal{L}$, $\mathcal{L}$ turns out to define a general framework for a formal comparison of a large scope of high-dimensional regression methods. Moreover, we propose a closed-form expression for the prediction score within $\mathcal{L}$ having the largest squared correlation with the response.

This chapter is organized as follows. Section 2 focuses on a theoretical comparison of the prediction performance of the naive linear predictor, designed to be optimal under independence and the OLS prediction score in the standard linear regression framework. Extending the results obtained by Bickel and Levina (2004) in a two-group classification context, sharp bounds for the relative efficiency of the naive predictor are given under assumption of an arbitrary dependence structure, with explicit expressions for the association parameters corresponding to the lowest and largest efficiency. In Section 3, based on the former comparative study, the class $\mathcal{L}$ of prediction scores, previously mentioned above, is introduced to extend the handling of dependence to more flexible approaches than just ignorance or a complete whitening of the explanatory variables. We show that both Ridge and PLS predictors belong to $\mathcal{L}$. A closed-form expression of the optimal predictor within $\mathcal{L}$ is derived in Section 4 and compared to existing methods through a simulation study and using two benchmark datasets in Section 5. Finally, a discussion ends this chapter.

## 2 Optimal versus naive linear prediction

The $n \ll p$ paradigm has raised fundamental questions about the extension of well-proven estimation methods in regular $n > p$ designs, such as least-squares or more generally maximum-likelihood estimation. One crucial point in those discussions is the ignorance or not of the dependence across explanatory variables. As in Bickel

and Levina (2004) for two-group classification using Linear Discriminant Analysis, let us first discuss this issue in the standard linear regression framework.

## 2.1 Optimal linear prediction

Hereafter, it is assumed that $(\boldsymbol{X}', Y)'$ is normally distributed with mean $(\boldsymbol{\mu}_{\boldsymbol{x}}', \mu_y)'$ and a positive variance-covariance matrix, with $\mathrm{Var}(\boldsymbol{X}) = \boldsymbol{\Sigma}_{\boldsymbol{x}}$, $\mathrm{Cov}(\boldsymbol{X}, Y) = \boldsymbol{\sigma}_{\boldsymbol{xy}}$ and $\mathrm{var}(Y) = \sigma_y^2$. In the present multivariate normal set-up, the unconditional dependence of the explanatory variables is captured by the variance-covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{x}}$. A part of the former dependence is due to the one-to-one association between each of the explanatory variables and the response, whereas a complementary part can be viewed as intrinsic dependence conditionally on the response. Indeed, if $\boldsymbol{\Sigma} = \mathrm{Var}(\boldsymbol{X}|Y)$ stands for the conditional variance-covariance matrix of the explanatory variables, the following relationship holds between conditional and unconditional dependence: $\boldsymbol{\Sigma}_{\boldsymbol{x}} = \boldsymbol{\Sigma} + \boldsymbol{\sigma}_{\boldsymbol{xy}}\boldsymbol{\sigma}_{\boldsymbol{xy}}'/\sigma_y^2$. In order to disentangle properly the association parameters and the intrinsic dependence across explanatory variables, the parameters $(\boldsymbol{\Sigma}, \boldsymbol{\sigma}_{\boldsymbol{xy}}, \sigma_y^2)$ will be preferred hereafter to the natural variance parameters $(\boldsymbol{\Sigma}_{\boldsymbol{x}}, \boldsymbol{\sigma}_{\boldsymbol{xy}}, \sigma_y^2)$ of the joint distribution.

Within the class $\{L_{\boldsymbol{\ell}}(\boldsymbol{X}) = \ell_0 + \boldsymbol{\ell}'\boldsymbol{X}, \ell_0 \in \mathbb{R}, \boldsymbol{\ell} \in \mathbb{R}^p\}$ of linear predictors, the so-called Best Linear Unbiased Predictor (BLUP) gives the closed-form expression of the predictor minimizing the Mean Squared Error of Prediction (MSEP):

$$\begin{aligned} L_{\mathrm{opt}}(\boldsymbol{X}) &= \mu_y + (\boldsymbol{X} - \boldsymbol{\mu}_{\boldsymbol{x}})'\boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1}\boldsymbol{\sigma}_{\boldsymbol{xy}}, \\ &= \mu_y + \left(1 + \frac{\boldsymbol{\sigma}_{\boldsymbol{xy}}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\sigma}_{\boldsymbol{xy}}}{\sigma_y^2}\right)^{-1}(\boldsymbol{X} - \boldsymbol{\mu}_{\boldsymbol{x}})'\boldsymbol{\Sigma}^{-1}\boldsymbol{\sigma}_{\boldsymbol{xy}}. \end{aligned}$$

Let us reformulate $L_{\mathrm{opt}}(\boldsymbol{X})$ by introducing the conditional correlation matrix $\boldsymbol{C} = \boldsymbol{D}_{\boldsymbol{\sigma}}^{-1}\boldsymbol{\Sigma}\boldsymbol{D}_{\boldsymbol{\sigma}}^{-1}$ of the explanatory variables, where $\boldsymbol{D}_{\boldsymbol{\sigma}}$ is the $p \times p$ diagonal matrix whose diagonal entries are the conditional standard deviations $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_p)'$, the vector $\tilde{\boldsymbol{X}} = \boldsymbol{D}_{\boldsymbol{\sigma}}^{-1}(\boldsymbol{X} - \boldsymbol{\mu}_{\boldsymbol{x}})$ of scaled explanatory variables and the $p$-vector $\tilde{\boldsymbol{\sigma}}_{\boldsymbol{xy}} = \boldsymbol{D}_{\boldsymbol{\sigma}}^{-1}\boldsymbol{\sigma}_{\boldsymbol{xy}}$ of covariances between $\tilde{\boldsymbol{X}}$ and $Y$:

$$L_{\mathrm{opt}}(\tilde{\boldsymbol{X}}) = \mu_y + \left(1 + \frac{\tilde{\boldsymbol{\sigma}}_{\boldsymbol{xy}}'\boldsymbol{C}^{-1}\tilde{\boldsymbol{\sigma}}_{\boldsymbol{xy}}}{\sigma_y^2}\right)^{-1}\tilde{\boldsymbol{X}}'\boldsymbol{C}^{-1}\tilde{\boldsymbol{\sigma}}_{\boldsymbol{xy}}. \tag{5.4}$$

Still within the class of linear predictors, it is straightforwardly proved that $L_{\mathrm{opt}}(\tilde{\boldsymbol{X}})$

also has the largest squared correlation $\mathrm{R}^2_{\mathrm{opt}}$ with $Y$, where:

$$\mathrm{R}^2_{\mathrm{opt}} = \frac{\tilde{\boldsymbol{\sigma}}'_{\boldsymbol{xy}} \boldsymbol{C}^{-1} \tilde{\boldsymbol{\sigma}}_{\boldsymbol{xy}}}{\sigma_y^2 + \tilde{\boldsymbol{\sigma}}'_{\boldsymbol{xy}} \boldsymbol{C}^{-1} \tilde{\boldsymbol{\sigma}}_{\boldsymbol{xy}}}.$$

## 2.2 Naive linear prediction

Similarly as in the Linear Discriminant Analysis framework (Bickel and Levina, 2004; Dudoit et al., 2002), a naive linear prediction score is obtained by replacing $\boldsymbol{C}$ in expression (5.4) by $\mathbf{I}_p$, which amounts to ignoring the conditional dependence between explanatory variables:

$$L_{\mathrm{N}}(\tilde{\boldsymbol{X}}) = \mu_y + \left(1 + \frac{\tilde{\boldsymbol{\sigma}}'_{\boldsymbol{xy}} \tilde{\boldsymbol{\sigma}}_{\boldsymbol{xy}}}{\sigma_y^2}\right)^{-1} \tilde{\boldsymbol{X}}' \tilde{\boldsymbol{\sigma}}_{\boldsymbol{xy}}. \tag{5.5}$$

For a straightforward comparison of the naive and optimal linear predictors, the squared correlation $\mathrm{R}^2_{\mathrm{N}}$ between $L_{\mathrm{N}}(\boldsymbol{Z})$ and the response is expressed as a function of $\mathrm{R}^2_{\mathrm{opt}}$:

$$\begin{aligned} \mathrm{R}^2_{\mathrm{N}} &= \frac{(\tilde{\boldsymbol{\sigma}}'_{\boldsymbol{xy}} \tilde{\boldsymbol{\sigma}}_{\boldsymbol{xy}})^2}{\sigma_y^2 \tilde{\boldsymbol{\sigma}}'_{\boldsymbol{xy}} \boldsymbol{C} \tilde{\boldsymbol{\sigma}}_{\boldsymbol{xy}} + (\tilde{\boldsymbol{\sigma}}'_{\boldsymbol{xy}} \tilde{\boldsymbol{\sigma}}_{\boldsymbol{xy}})^2} \\ &= \mathrm{R}^2_{\mathrm{opt}} \frac{f(\mathrm{R}^2_{\mathrm{opt}}) + 1}{f(\mathrm{R}^2_{\mathrm{opt}}) + g(\tilde{\boldsymbol{\sigma}}_{\boldsymbol{xy}}, \boldsymbol{C})}, \end{aligned} \tag{5.6}$$

where $f(x) = x/(1-x)$ and $g(\tilde{\boldsymbol{\sigma}}_{\boldsymbol{xy}}, \boldsymbol{C}) = \tilde{\boldsymbol{\sigma}}'_{\boldsymbol{xy}} \boldsymbol{C} \tilde{\boldsymbol{\sigma}}_{\boldsymbol{xy}} \tilde{\boldsymbol{\sigma}}'_{\boldsymbol{xy}} \boldsymbol{C}^{-1} \tilde{\boldsymbol{\sigma}}_{\boldsymbol{xy}} / (\tilde{\boldsymbol{\sigma}}'_{\boldsymbol{xy}} \tilde{\boldsymbol{\sigma}}_{\boldsymbol{xy}})^2$. Analogously with Bickel and Levina (2004) for the comparison of the classification performance of the linear Bayes and naive Bayes classifiers, the Kantorovitch inequality (Bickel and Levina, 2004) provides an upper bound for $g(\tilde{\boldsymbol{\sigma}}_{\boldsymbol{xy}}, \boldsymbol{C})$ over all possible $\tilde{\boldsymbol{\sigma}}_{\boldsymbol{xy}}$:

$$g(\tilde{\boldsymbol{\sigma}}_{\boldsymbol{xy}}, \boldsymbol{C}) \leq \frac{1}{4} \frac{\{1 + \tau(\boldsymbol{C})\}^2}{\tau(\boldsymbol{C})} = g_{\max}(\boldsymbol{C}),$$

where $\tau(\boldsymbol{C}) = \lambda_{\max}(\boldsymbol{C}) / \lambda_{\min}(\boldsymbol{C})$ is the ratio between the largest and smallest eigenvalues of $\boldsymbol{C}$. This quantity $\tau(\boldsymbol{C})$ can be interpreted as a measure of the amount of conditional dependence between explanatory variables: when it is close to 1, $\boldsymbol{C}$ is itself close to the identity matrix whereas a large $\tau(\boldsymbol{C})$ means that there exists a linear combination of the scaled explanatory variables which concentrates a large

part of the variability of those explanatory variables. Finally, a lower bound for the squared correlation between the naive predictor and the response is deduced:

$$R^2_{opt} \frac{f(R^2_{opt}) + 1}{f(R^2_{opt}) + g_{max}(\boldsymbol{C})} \leq R^2_N \leq R^2_{opt}. \tag{5.7}$$

Figure 5.1 displays the lower bound for $R^2_N$ given in expression (5.7) as a function of $R^2_{opt}$ for a range of values of $\tau(\boldsymbol{C})$ going from 1 to 100. It clearly shows that ignoring dependence can strongly deter the prediction performance of the naive predictor and that the potential loss increases with the dependence of the explanatory variables.

Even if the former conclusion is generally true over all possible association patterns between $\boldsymbol{X}$ and $Y$, it is important to keep in mind that, whatever the dependence pattern in $\boldsymbol{C}$, $\tilde{\boldsymbol{\sigma}}_{\boldsymbol{xy}}$ may be such that $R^2_N$ is actually close to $R^2_{opt}$. Moreover, the lower bound given in (5.7) may not be reached for any vector $\tilde{\boldsymbol{\sigma}}_{\boldsymbol{xy}}$, as demonstrated further in the present section.

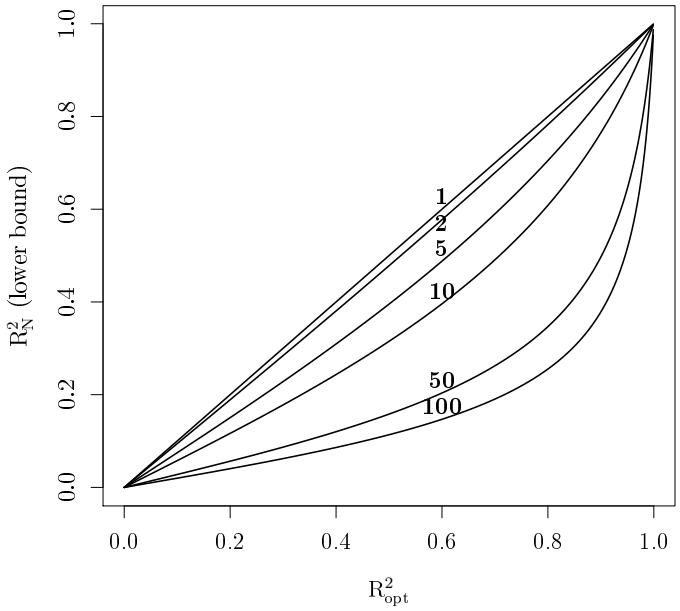


Figure 5.1: Lower bound for $R^2_N$ as a function of $R^2_{opt}$, for different values of $\tau(\boldsymbol{C})$.

## 2.3 Prediction performance of naive linear prediction

The singular value decomposition of $\boldsymbol{C} = \boldsymbol{U}\boldsymbol{D}_{\boldsymbol{\lambda}}\boldsymbol{U}'$ is now introduced, where $\boldsymbol{D}_{\boldsymbol{\lambda}}$ is the $p \times p$ diagonal matrix whose diagonal entries are the eigenvalues $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_p)'$ of $\boldsymbol{C}$ and $\boldsymbol{U}$ is the $p \times p$ matrix of corresponding eigenvectors, with $\boldsymbol{U}\boldsymbol{U}' = \mathbf{I}_p$. In the following, the vector of whitened explanatory variables $\boldsymbol{Z} = \boldsymbol{U}'\tilde{\boldsymbol{X}}$, with $\mathrm{Var}(\boldsymbol{Z}) = \boldsymbol{U}'\boldsymbol{C}\boldsymbol{U} = \boldsymbol{D}_{\boldsymbol{\lambda}}$, will be used instead of the previously considered vector $\tilde{\boldsymbol{X}}$. Accordingly, the notation $\boldsymbol{\gamma} = \boldsymbol{U}'\tilde{\boldsymbol{\sigma}}_{\boldsymbol{xy}}$ will be used hereafter for the covariance between the response and $\boldsymbol{Z}$. Note that the two prediction scores, previously referred to as $L_{\mathrm{opt}}(\tilde{\boldsymbol{X}})$ and $L_{\mathrm{N}}(\tilde{\boldsymbol{X}})$, will accordingly be referred to as $L_{\mathrm{opt}}(\boldsymbol{Z})$ and $L_{\mathrm{N}}(\boldsymbol{Z})$.

Sharp upper and lower bounds for $\mathrm{R}^2_{\mathrm{N}}$ over all possible vectors $\boldsymbol{\gamma}$ of association parameters are given in the following Theorem, with explicit values of $\boldsymbol{\gamma}$ for which these bounds are reached.

**Theorem 2.1** *Let $\boldsymbol{v}(\boldsymbol{\lambda})$ denote the eigenvector associated to the only positive eigenvalue of the matrix $\boldsymbol{\lambda}\boldsymbol{\lambda}^{-1\prime} + \boldsymbol{\lambda}^{-1}\boldsymbol{\lambda}'$ with $\boldsymbol{v}'(\boldsymbol{\lambda})\boldsymbol{v}(\boldsymbol{\lambda}) = 1$, where $\boldsymbol{\lambda}^{-1} = (1/\lambda_1, \ldots, 1/\lambda_p)'$. Then,*

$$R^2_{opt}\frac{f(R^2_{opt}) + 1}{f(R^2_{opt}) + g_{\max}(\boldsymbol{\lambda})} \leq R^2_N \leq R^2_{opt}, \tag{5.8}$$

*where $g_{\max}(\boldsymbol{\lambda}) = \boldsymbol{v}(\boldsymbol{\lambda})'\boldsymbol{\lambda}.\boldsymbol{v}(\boldsymbol{\lambda})'\boldsymbol{\lambda}^{-1}$.*

*If the coordinates of $\boldsymbol{\gamma}$ are the square-roots of the coordinates of $\boldsymbol{v}(\boldsymbol{\lambda})$, then $R^2_N$ reaches its lower limit. On the contrary, for any vector $\boldsymbol{\gamma}$ with only one nonzero coordinate, $R^2_N$ reaches its upper limit.*

*Proof*: see Section 7.

It is deduced from Theorem 2.1 and from its proof that the potential loss of prediction performance induced by ignoring dependence among explanatory variables is determined by the interplay between the patterns of association and dependence through $g(\tilde{\boldsymbol{\sigma}}_{\boldsymbol{xy}}, \boldsymbol{C}) \coloneqq g(\boldsymbol{\gamma}, \boldsymbol{\lambda})$. Up to now, the optimal and naive linear prediction scores have been introduced in a purely probabilistic framework, where all parameters of the joint distribution of $(\boldsymbol{X}', Y)'$ are supposed to be known. In the following,

we introduce the sample counterparts of $L_{\mathrm{N}}(\boldsymbol{Z})$ and $L_{\mathrm{opt}}(\boldsymbol{Z})$ and show that, in large or high-dimensional contexts, naive linear prediction can actually markedly outperform optimal linear prediction.

## 2.4 Ordinary Least Squares and naive linear prediction in high dimension

It is now supposed that the training sample contains $n \geq 2$ independent observations $(\boldsymbol{X}_{\boldsymbol{i}}', Y_i)'$ of $(\boldsymbol{X}', Y)'$, where the sample size $n$ can be smaller than $p$. Hereafter, $(\bar{\boldsymbol{X}}, \bar{Y})$ will denote the sample estimates of $(\boldsymbol{\mu_x}, \mu_y)$ and $(\boldsymbol{S_x}, \boldsymbol{s_{xy}}, s_y^2)$ the sample estimates of $(\boldsymbol{\Sigma_x}, \boldsymbol{\sigma_{xy}}, \sigma_y^2)$. Consistently, $\boldsymbol{S} = \boldsymbol{S_x} - (\boldsymbol{s_{xy}}\boldsymbol{s_{xy}}')/s_y^2$ will stand for the estimate of $\boldsymbol{\Sigma}$ and $\hat{\boldsymbol{C}} = \boldsymbol{D_s}^{-1}\boldsymbol{S}\boldsymbol{D_s}^{-1}$ for the estimate of $\boldsymbol{C}$, where $\boldsymbol{D_s}$ is the $p \times p$ diagonal matrix which diagonal entries are the sample conditional standard deviations $\boldsymbol{s} = (s_1, \ldots, s_p)'$ of the explanatory variables. The Singular Value Decomposition of $\hat{\boldsymbol{C}}$ is also introduced here: $\hat{\boldsymbol{C}} = \hat{\boldsymbol{U}}\boldsymbol{D}_{\hat{\boldsymbol{\lambda}}}\hat{\boldsymbol{U}}'$, where $\boldsymbol{D}_{\hat{\boldsymbol{\lambda}}}$ is the $p \times p$ diagonal matrix whose diagonal entries are the eigenvalues $\hat{\boldsymbol{\lambda}} = (\hat{\lambda}_1, \ldots, \hat{\lambda}_p)'$ of $\hat{\boldsymbol{C}}$ and $\hat{\boldsymbol{U}}$ is the $p \times p$ matrix of corresponding eigenvectors, with $\hat{\boldsymbol{U}}\hat{\boldsymbol{U}}' = \mathbf{I}_p$. It is also assumed that the rank $q$ of $\hat{\boldsymbol{C}}$ can be smaller than $p$.

The sample counterpart of $L_{\mathrm{opt}}(\boldsymbol{Z})$ is the Ordinary Least Squares (OLS) linear prediction score $L_{\mathrm{OLS}}(\hat{\boldsymbol{Z}})$ obtained by plugging-in the estimates of the parameters of the joint distribution of $(\boldsymbol{X}', Y)'$ in expression (5.4) of $L_{\mathrm{opt}}(\boldsymbol{Z})$:

$$L_{\mathrm{OLS}}(\hat{\boldsymbol{Z}}) \;=\; \bar{Y} + \left(1 + \frac{\hat{\boldsymbol{\gamma}}'\boldsymbol{D}_{\hat{\boldsymbol{\lambda}}}^{-}\hat{\boldsymbol{\gamma}}}{s_y^2}\right)^{-1} \hat{\boldsymbol{Z}}'\boldsymbol{D}_{\hat{\boldsymbol{\lambda}}}^{-}\hat{\boldsymbol{\gamma}}, \tag{5.9}$$

where $\hat{\boldsymbol{\gamma}} = \hat{\boldsymbol{U}}'\boldsymbol{D_s}^{-1}\boldsymbol{s_{xy}}$, $\hat{\boldsymbol{Z}} = \hat{\boldsymbol{U}}'\boldsymbol{D_s}^{-1}(\boldsymbol{X} - \bar{\boldsymbol{X}})$. In expression (5.9), the notation $\boldsymbol{D}_{\hat{\boldsymbol{\lambda}}}^{-}$ denotes the Moore-Penrose generalized inverse of $\boldsymbol{D}_{\hat{\boldsymbol{\lambda}}}$, namely the $p \times p$ diagonal matrix which first $q$ diagonal entries are the inverse of the positive eigenvalues $\hat{\lambda}_j$ and the remaining $p - q$ are 0.

Analogously, the sample counterpart of $L_{\mathrm{N}}(\boldsymbol{Z})$ has the following expression:

$$L_{\mathrm{N}}(\hat{\boldsymbol{Z}}) \;=\; \bar{Y} + \left(1 + \frac{\hat{\boldsymbol{\gamma}}'\hat{\boldsymbol{\gamma}}}{s_y^2}\right)^{-1} \hat{\boldsymbol{Z}}'\hat{\boldsymbol{\gamma}}. \tag{5.10}$$

Note that the previously studied theoretical naive predictor is not used in the remaining of this chapter. Therefore, from now, $L_{\mathrm{N}}$ refers to the empirical naive

predictor. Since $(\bar{\boldsymbol{X}}, \bar{Y}, \hat{\boldsymbol{\lambda}}, \hat{\boldsymbol{\gamma}}, \boldsymbol{s}, s_y^2)$ are consistent estimators of $(\boldsymbol{\mu_x}, \mu_y, \boldsymbol{\lambda}, \boldsymbol{\gamma}, \boldsymbol{\sigma}, \sigma_y^2)$, then the bounds given in Theorem 2.1 hold asymptotically for the comparison of $L_{\mathrm{N}}(\hat{\boldsymbol{Z}})$ and $L_{\mathrm{OLS}}(\hat{\boldsymbol{Z}})$. As illustrated below, these bounds do not hold in small-sample and high-dimensional situations.

## 2.5 Illustrative comparative study in high dimension

The benchmark dataset described in Lu et al. (2004), freely available in the R package `care` (Zuber and Strimmer, 2017) or from https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE1572, is used hereafter to illustrate the comparison of $L_{\mathrm{N}}(\hat{\boldsymbol{Z}})$ and $L_{\mathrm{OLS}}(\hat{\boldsymbol{Z}})$. It contains gene expression profiles $\boldsymbol{X}$ of $p = 403$ genes for 30 human brain samples, used to predict the age $Y$ of each patient. The list of selected 403 genes results from prescreening and preprocessing as described in Zuber and Strimmer (2011). A data-driven simulation study is conducted hereafter, in which the conditional variance matrix between the explanatory variables given the response is the same as estimated by $\boldsymbol{S}$, using the former illustrative gene expression dataset.

We now consider a first pattern of association between $Y$ and $\boldsymbol{X}$ in which the asymptotic prediction performance of $L_{\mathrm{N}}(\hat{\boldsymbol{Z}})$ is the lowest possible with respect to $L_{\mathrm{OLS}}(\hat{\boldsymbol{Z}})$, according to Theorem 2.1. For that, a vector $\boldsymbol{\gamma}$ is obtained by taking the square-roots of the coordinates of the only eigenvector $\boldsymbol{v}(\hat{\boldsymbol{\lambda}})$ of $\hat{\boldsymbol{\lambda}}\hat{\boldsymbol{\lambda}}^{\boldsymbol{-1\prime}} + \hat{\boldsymbol{\lambda}}^{\boldsymbol{-1}}\hat{\boldsymbol{\lambda}}'$ associated to a positive eigenvalue. Since $\hat{\boldsymbol{C}}$ is not full rank, $\hat{\boldsymbol{\lambda}}$ and $\hat{\boldsymbol{\lambda}}^{\boldsymbol{-1}}$ are restricted to the $q = n - 2 = 28$ nonzero eigenvalues of $\hat{\boldsymbol{C}}$. The resulting vector $\boldsymbol{\gamma}$ is multiplied by a scalar, adequately chosen so that $\mathrm{R}^2_{\mathrm{opt}} = 0.8$ (arbitrarily). The covariance vector $\hat{\boldsymbol{\sigma}}_{\boldsymbol{xy}} = \boldsymbol{D_s}\hat{\boldsymbol{U}}\boldsymbol{\gamma}$ is deduced, where the $q$ columns of $\hat{\boldsymbol{U}}$ are the eigenvectors of $\hat{\boldsymbol{C}}$ corresponding to nonzero eigenvalues. Finally, the estimation of $\boldsymbol{\Sigma_x}$ is updated to be consistent with the fixed patterns of conditional variance-covariance of $\boldsymbol{X}$ and association between $\boldsymbol{X}$ and $Y$: $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{x}} = \boldsymbol{S} + (\hat{\boldsymbol{\sigma}}_{\boldsymbol{xy}}\hat{\boldsymbol{\sigma}}'_{\boldsymbol{xy}})/s_y^2$. Using the former set $(\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}, \hat{\boldsymbol{\sigma}}_{\boldsymbol{xy}}, s_y^2)$ of variance parameters for the joint distribution of $(\boldsymbol{X}', Y)'$, it is straightforwardly checked that the asymptotic squared correlation between the naive predictor and $Y$, given by equation 5.6, is 0.2. Note that the lower bound for $\mathrm{R}^2_{\mathrm{N}}$ given by Kantorovitch inequality (see equation 5.8) yields 0.09, smaller than the smallest reachable limit.

A second pattern of association between $Y$ and $\boldsymbol{X}$ is obtained by starting from a vector $\boldsymbol{\gamma}$ with only one nonzero coordinate. The same sequence of operations as described above for the first pattern of association leads here to a set $(\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}, \hat{\boldsymbol{\sigma}}_{\boldsymbol{xy}}, s_y^2)$ of variance parameters for the joint distribution of $(\boldsymbol{X}', Y)'$, for which both $L_{\mathrm{N}}(\hat{\boldsymbol{Z}})$ and $L_{\mathrm{opt}}(\hat{\boldsymbol{Z}})$ have a squared correlation with $Y$ equal to 0.8.

The above two sets of variance parameters are now used to simulate 1,000 training datasets of dimension $n \times (p+1)$, which rows are independent realizations of $(\boldsymbol{X}', Y)$, with expectation $(\boldsymbol{\mu}_{\boldsymbol{x}}', \mu_y) = \mathbf{0}$, in asymptotic ($n = 1000$) and non-asymptotic ($n = 30$, as in the original dataset) conditions. For each training dataset, a test dataset of $1,000$ individuals is also simulated with the same joint distribution as that of $(\boldsymbol{X}', Y)$. Using each training dataset, the response value of the individuals in the corresponding test dataset is predicted using $L_{\mathrm{N}}(\hat{\boldsymbol{Z}})$ and $L_{\mathrm{OLS}}(\hat{\boldsymbol{Z}})$. Moreover, Ridge regression (Hoerl and Kennard, 1970) and Partial Least Squares regression (PLS, Wold et al. (1983, 1984)) are implemented using the R packages `glmnet` (Friedman et al., 2010) and `pls` (Mevik et al., 2019), respectively. For each method, the hyperparameter (the penalty coefficient for Ridge regression and the number of components for PLS regression) is chosen by minimizing the mean squared error of prediction estimated using a 10-fold cross-validation procedure. Table 5.1 reproduces the mean squared correlation between the predicted and observed values of the response in the test dataset over 1,000 simulations for each of the 4 scenarios (two patterns of association parameters, two values for $n$).

As expected, in both scenarios of association between the response and the explanatory variables and in asymptotic conditions ($n = 1,000$), the prediction performances of $L_{\mathrm{N}}(\hat{\boldsymbol{Z}})$ and $L_{\mathrm{OLS}}(\hat{\boldsymbol{Z}})$ are close to the values obtained with the fixed simulation parameters, namely $(\mathrm{R}^2_{\mathrm{N}} = 0.2, \mathrm{R}^2_{\mathrm{opt}} = 0.8)$ in the first scenario and $(\mathrm{R}^2_{\mathrm{N}} = 0.8, \mathrm{R}^2_{\mathrm{opt}} = 0.8)$ in the second scenario. Still in these asymptotic conditions, the prediction performances of Ridge and PLS regression are close to the best performance, which can be explained by the fact that the present asymptotic conditions favor an accurate choice of the optimal hyperparameter in those methods, either a small regularization parameter in Ridge regression or a large number of PLS components, for which the corresponding predictors are close to $L_{\mathrm{OLS}}(\hat{\boldsymbol{Z}})$. The fact that any predictor reaches its optimal theoretical performance demonstrates that the sample size $n = 1,000$ can be considered as asymptotic in the considered scenarios.

Non-asymptotic conditions ($n = 30$) have a strong negative impact on the prediction performance of the four methods in the first scenario of association parameters, especially for $L_{\mathrm{N}}(\hat{\boldsymbol{Z}})$, $L_{\mathrm{OLS}}(\hat{\boldsymbol{Z}})$ and PLS regression. Interestingly, still in non-asymptotic conditions but now in the second scenario of association parameters, the prediction performance of $L_{\mathrm{N}}(\hat{\boldsymbol{Z}})$, Ridge and PLS regression are as good as in the asymptotic conditions, whereas the prediction performance of $L_{\mathrm{OLS}}(\hat{\boldsymbol{Z}})$ is markedly lower.

Clearly, for the same conditional dependence pattern across explanatory variables but in two different scenarios of association between $\boldsymbol{X}$ and $Y$, comparing $L_{\mathrm{N}}(\hat{\boldsymbol{Z}})$ to $L_{\mathrm{OLS}}(\hat{\boldsymbol{Z}})$ leads to very different conclusions, the second scenario being far more favorable to the ignorance of dependence. Ridge and PLS regression offer alternative ways of handling dependence, in-between ignorance as in $L_{\mathrm{N}}(\hat{\boldsymbol{Z}})$ and a complete whitening as in $L_{\mathrm{OLS}}(\hat{\boldsymbol{Z}})$, using a shrinked (Ridge regression) or rank-reduced (PLS regression) estimation of the covariance matrix of the explanatory variables. In small sample conditions, the simulation study above demonstrates that this alternative handling of dependence can show prediction improvement.

Table 5.1: Mean squared correlation between the predicted and observed values of the response in the test dataset over 1000 simulations for each of 4 scenarios (two patterns of association parameters, two values for $n$) and four prediction methods (Naive, OLS, Ridge and PLS). The numbers between brackets are the 2.5% and 97.5% quantiles.

| | | Naive | OLS | Ridge | PLS |
|---|---|---|---|---|---|
| Scenario 1 | $n = 1000$ | 0.23 [0.09, 0.53] | 0.79 [0.77, 0.81] | 0.79 [0.77, 0.81] | 0.79 [0.76, 0.81] |
| | $n = 30$ | 0.13 [0,0.44] | 0.22 [0, 0.53] | 0.55 [0.31, 0.69] | 0.38 [0.02, 0.66] |
| Scenario 2 | $n = 1000$ | 0.80 [0.78, 0.82] | 0.79 [0.77,0.82] | 0.80 [0.78, 0.82] | 0.80 [0.78, 0.82] |
| | $n = 30$ | 0.80 [0.78, 0.82] | 0.23 [0, 0.53] | 0.78 [0.64, 0.82] | 0.80 [0.78, 0.82] |

To complete the above study, the four prediction methods considered above are now implemented to predict the age of a patient from the original illustrative dataset (Lu et al., 2004). Figure 5.2 displays the distributions of the squared correlations

between predicted and observed values with 50 random splittings of the dataset in a 10-fold cross validation set-up. In the present situation, $L_{\mathrm{N}}(\hat{\boldsymbol{Z}})$ turns out to show better prediction performance than $L_{\mathrm{OLS}}(\hat{\boldsymbol{Z}})$, Ridge and PLS regression. Note that, although the OLS predictor is not considered in most comparative studies of high-dimensional prediction methods, its prediction performance is here comparable to that of Ridge and PLS regression.

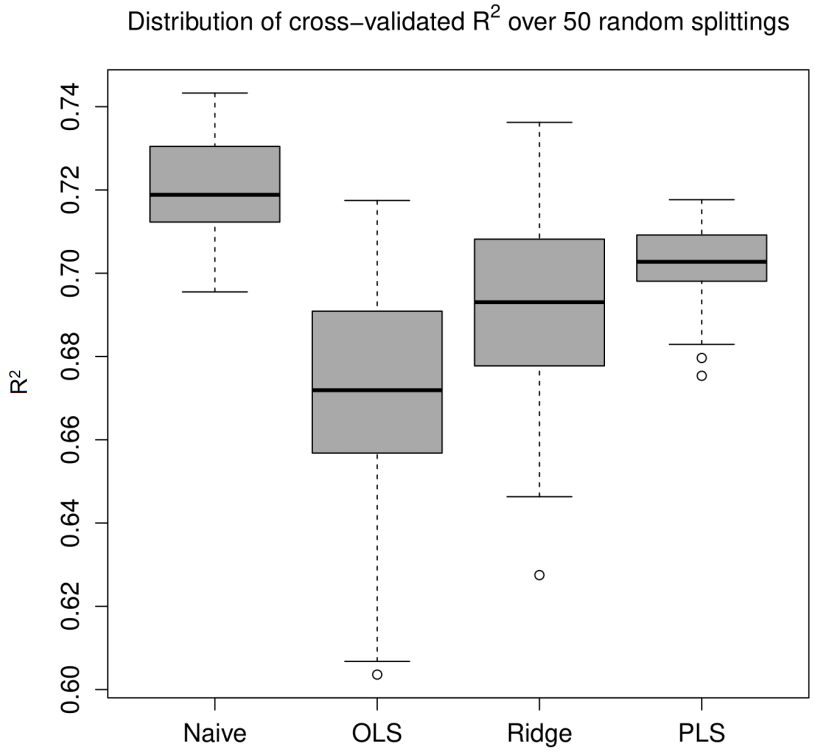


Figure 5.2: Distribution of squared correlations between the predicted and observed values of the response for the prediction of the age of a patient by the genomic profiles in the illustrative dataset, over 50 random splittings of the dataset in a 10-fold cross-validation set-up. Four prediction methods are compared: $L_{\mathrm{N}}(\hat{\boldsymbol{Z}})$, $L_{\mathrm{OLS}}(\hat{\boldsymbol{Z}})$, Ridge and PLS.

In the following, we introduce a new class of linear predictors, including the four regression methods introduced in the comparative study above, and deduce an optimal choice within this class for a flexible handling of dependence.

# 3 A new class of prediction scores

## 3.1 Introduction of a new class of prediction scores

In the following, we will focus on prediction performance in terms of squared correlation between the linear score and the response. Therefore, from now on, two predictors $L_1(\boldsymbol{X})$ and $L_2(\boldsymbol{X})$ will be said to be equivalent, which will be denoted $L_1 \equiv L_2$, if there exists scalars $a$ and $b$, with $b \neq 0$, such that, for all $\boldsymbol{X}$, $L_2(\boldsymbol{X}) = a + bL_1(\boldsymbol{X})$. For example, it is deduced from (5.9) and (5.10) respectively that $L_{\mathrm{OLS}}(\hat{\boldsymbol{Z}}) \equiv \hat{\boldsymbol{Z}}'\boldsymbol{D}_{\hat{\boldsymbol{\lambda}}}^{-}\hat{\boldsymbol{\gamma}}$ and $L_{\mathrm{N}}(\hat{\boldsymbol{Z}}) \equiv \hat{\boldsymbol{Z}}'\hat{\boldsymbol{\gamma}}$.

Interestingly, $L_{\mathrm{OLS}}(\hat{\boldsymbol{Z}})$ and $L_{\mathrm{N}}(\hat{\boldsymbol{Z}})$ are both equivalent to linear combinations of the coordinates of the $p$-vector $\boldsymbol{\xi}(\hat{\boldsymbol{Z}}) = \hat{\boldsymbol{Z}} \odot \hat{\boldsymbol{\gamma}}$, where $\odot$ stands for the term-by-term product of two vectors with same dimension: $L_{\mathrm{OLS}}(\hat{\boldsymbol{Z}}) = \hat{\boldsymbol{\lambda}}^{-\mathbf{1}'}\boldsymbol{\xi}(\hat{\boldsymbol{Z}})$ and $L_{\mathrm{N}}(\hat{\boldsymbol{Z}}) = \mathbf{1}_p'\boldsymbol{\xi}(\hat{\boldsymbol{Z}})$, with $\mathbf{1}_p = (1, \ldots, 1)'$.

Therefore, ignoring dependence in linear prediction or, on the contrary, fully whitening the explanatory variables can both be obtained by an ad-hoc weighting of the coordinates of $\boldsymbol{\xi}(\hat{\boldsymbol{Z}})$. In order to enlarge the scope of dependence handling solutions, the class $\mathcal{L}$ of predictions scores defined as linear combinations of the elements of $\boldsymbol{\xi}(\hat{\boldsymbol{Z}})$ is now introduced:

$$\mathcal{L} = \left\{ L_{\boldsymbol{h}}(\hat{\boldsymbol{Z}}) = \boldsymbol{h}'\boldsymbol{\xi}(\hat{\boldsymbol{Z}}),\ \boldsymbol{h} = (h_1, \ldots, h_p)',\ \text{with } \boldsymbol{h}'\boldsymbol{h} = 1 \right\}. \tag{5.11}$$

The arbitrary restriction $\boldsymbol{h}'\boldsymbol{h} = 1$ just aims at reducing equivalence subclasses of predictors to a unique element in $\mathcal{L}$. Note that, similarly as in expressions (5.9) and (5.10) of $L_{\mathrm{OLS}}(\hat{\boldsymbol{Z}})$ and $L_{\mathrm{N}}(\hat{\boldsymbol{Z}})$ respectively, we can define for each predictor $L_{\boldsymbol{h}}(\hat{\boldsymbol{Z}})$ in $\mathcal{L}$, an equivalent predictor $L_{\boldsymbol{h}}(\hat{\boldsymbol{Z}}) \equiv \bar{Y} + \left\{1 + (\hat{\boldsymbol{\gamma}}'\boldsymbol{D}_{\boldsymbol{h}}\hat{\boldsymbol{\gamma}})/s_y^2\right\}^{-1} \boldsymbol{h}'\boldsymbol{\xi}(\hat{\boldsymbol{Z}})$ that shall be preferred if the prediction performance is measured in terms of mean squared error of prediction.

In the following, we show that Ridge (Hoerl and Kennard, 1970) and Partial Least Squares (Wold et al., 1983, 1984) prediction scores form subclasses of $\mathcal{L}$. Both methods can be viewed as resulting from a least-squares optimization under restriction on the vector of regression coefficients.

## 3.2 $\mathcal{L}$ contains Ridge predictions scores

Provided the explanatory variables have been centered and scaled using their conditional standard deviations $\boldsymbol{s}$, then Ridge regression consists in the minimization over $\boldsymbol{\beta}$ of the penalized least-squares criterion $\sum_{i=1}^{n}(Y_i-\bar{Y}-(\boldsymbol{X_i}-\bar{\boldsymbol{X}})'\boldsymbol{D_s^{-1}}\boldsymbol{\beta})^2+n\kappa\boldsymbol{\beta}'\boldsymbol{\beta}$, where $\kappa$ is a nonnegative penalty parameter, which leads to the following estimate of the regression coefficients $\boldsymbol{\beta}$:

$$\hat{\boldsymbol{\beta}}_\kappa = (\boldsymbol{D_s^{-1}}\boldsymbol{S_x}\boldsymbol{D_s^{-1}}+\kappa\mathbf{I}_p)^{-1}\boldsymbol{D_s^{-1}}\boldsymbol{s_{xy}}, \tag{5.12}$$

A closed form expression of the corresponding Ridge prediction score is deduced:

$$\begin{aligned} L_{\text{Ridge}}(\boldsymbol{X},\kappa) &\equiv (\boldsymbol{X}-\bar{\boldsymbol{X}})'\boldsymbol{D_s^{-1}}\hat{\boldsymbol{\beta}}_\kappa \\ &\equiv (\boldsymbol{X}-\bar{\boldsymbol{X}})'\boldsymbol{D_s^{-1}}(\boldsymbol{D_s^{-1}}\boldsymbol{S_x}\boldsymbol{D_s^{-1}}+\kappa\mathbf{I}_p)^{-1}\boldsymbol{D_s^{-1}}\boldsymbol{s_{xy}}. \end{aligned}$$

Hereafter, $\mathcal{L}_{\text{Ridge}}$ denotes the class of Ridge prediction scores $L_{\text{Ridge}}(\boldsymbol{X},\kappa)$.

**Theorem 3.1** *$\mathcal{L}_{Ridge}$ is a subclass of $\mathcal{L}$. Moreover, if $\boldsymbol{h_\kappa}$ denotes the weighting vector, with $\boldsymbol{h_\kappa'h_\kappa}=1$, such that the Ridge prediction score obtained with penalty parameter $\kappa$ is $\boldsymbol{h_\kappa'\xi}(\hat{\boldsymbol{Z}})$, then $\lim_{\kappa\to+\infty}\boldsymbol{h_\kappa}=(1/\sqrt{p})\mathbf{1}_p$ and $\lim_{\kappa\to 0}\boldsymbol{h_\kappa}=\hat{\boldsymbol{\lambda}}^{\boldsymbol{-1}}/\sqrt{\hat{\boldsymbol{\lambda}}^{\boldsymbol{-1'}}\hat{\boldsymbol{\lambda}}^{\boldsymbol{-1}}}$. As a consequence, $L_N(\hat{\boldsymbol{Z}})$ (resp. $L_{OLS}(\hat{\boldsymbol{Z}})$) can be approximated by a Ridge prediction score as close as desired provided $\kappa$ is chosen sufficiently large (resp. small).*

*Proof*: see Section 8.

Ridge regression is generally presented as a shrinkage estimation procedure aiming at minimizing the mean-squared error of prediction: in this framework, when the shrinkage parameter $\kappa$ gets larger, the Ridge regression coefficients tends to zero. Each estimated Ridge regression model then appears as a compromise between the OLS estimation of the full model and the null model, with no explanatory variables. In the present context where the goal is to maximize the squared correlation between the predictor and the response, due to the normalization restriction on $\boldsymbol{h}$, the limiting Ridge prediction score is the naive predictor.

## 3.3 $\mathcal{L}$ contains Partial Least Squares (PLS) prediction scores

Let us now introduce PLS regression (Wold et al., 1983, 1984) using the sequence of Krylov spaces representation used in Helland (1988); Blazère et al. (2014). As above for Ridge regression, it is supposed that the explanatory variables are centered and scaled using their conditional standard deviations $\boldsymbol{s}$. Then, as recalled by Lingjaerde and Christophersen (2000), the PLS estimate $\hat{\boldsymbol{\beta}}_{\mathrm{PLS},m}$ of the regression coefficient $\boldsymbol{\beta}$, with $m \geq 1$ PLS components, is obtained by minimizing the least-squares criterion $\sum_{i=1}^{n}(Y_i - \bar{Y} - (\boldsymbol{X_i} - \bar{\boldsymbol{X}})'\boldsymbol{D_s}^{-1}\boldsymbol{\beta})^2)$, subject to $\boldsymbol{\beta} \in \mathcal{K}_m(\boldsymbol{D_s}^{-1}\boldsymbol{S_x}\boldsymbol{D_s}^{-1}; \boldsymbol{D_s}^{-1}\boldsymbol{s_{xy}})$, where, for a positive $p \times p$ matrix $\boldsymbol{V}$ and a $p$-vector $\boldsymbol{v}$, the $m$-th Krylov space $\mathcal{K}_m(\boldsymbol{V}; \boldsymbol{v})$ is defined as follows: $\mathcal{K}_m(\boldsymbol{V}; \boldsymbol{v}) = \mathrm{span}\,\{\boldsymbol{v}, \boldsymbol{V}\boldsymbol{v}, \ldots, \boldsymbol{V}^{m-1}\boldsymbol{v}\}$.

Consequently, there exists $\boldsymbol{a} = (a_1, \ldots, a_m)'$, such that $\hat{\boldsymbol{\beta}}_{\mathrm{PLS},m}$ can be expressed as follows:

$$\hat{\boldsymbol{\beta}}_{\mathrm{PLS},m} = \sum_{i=1}^{m} a_i (\boldsymbol{D_s}^{-1}\boldsymbol{S_x}\boldsymbol{D_s}^{-1})^{i-1}\boldsymbol{D_s}^{-1}\boldsymbol{s_{xy}},$$

The corresponding PLS prediction score is deduced:

$$\begin{aligned} L_{\mathrm{PLS}}(\boldsymbol{X}, m) &\equiv (\boldsymbol{X} - \bar{\boldsymbol{X}})'\boldsymbol{D_s}^{-1}\hat{\boldsymbol{\beta}}_{\mathrm{PLS},m}, \\ &\equiv (\boldsymbol{X} - \bar{\boldsymbol{X}})'\boldsymbol{D_s}^{-1}\Big(\sum_{i=1}^{m} a_i (\boldsymbol{D_s}^{-1}\boldsymbol{S_x}\boldsymbol{D_s}^{-1})^{i-1}\Big)\boldsymbol{D_s}^{-1}\boldsymbol{s_{xy}}. \end{aligned}$$

Let $\mathcal{L}_{\mathrm{PLS}}$ denote the class of PLS prediction scores $L_{\mathrm{PLS}}(\boldsymbol{X}, m)$, indexed by the number $m$ of PLS components.

**Theorem 3.2** *$\mathcal{L}_{PLS}$ is a subclass of $\mathcal{L}$, which contains $L_N(\hat{\boldsymbol{Z}}) = L_{PLS}(\boldsymbol{X}, m = 1)$ and $L_{OLS}(\hat{\boldsymbol{Z}}) = L_{PLS}(\boldsymbol{X}, m = q)$.*

*Proof.* First, it is deduced from Lemma 9.1 given in Section 9 that $\mathcal{K}_m(\boldsymbol{D_s}^{-1}\boldsymbol{S_x}\boldsymbol{D_s}^{-1}; \boldsymbol{D_s}^{-1}\boldsymbol{s_{xy}}) = \mathcal{K}_m(\hat{\boldsymbol{C}} = \boldsymbol{D_s}^{-1}\boldsymbol{S}\boldsymbol{D_s}^{-1}; \boldsymbol{D_s}^{-1}\boldsymbol{s_{xy}})$. Therefore, there exists $\boldsymbol{b} = (b_1, \ldots, b_m)'$, such that $\hat{\boldsymbol{\beta}}_{\mathrm{PLS},m}$ can be expressed as follows:

$$\hat{\boldsymbol{\beta}}_{\mathrm{PLS},m} = \sum_{i=1}^{m} b_i \hat{\boldsymbol{C}}^{i-1}\boldsymbol{D_s}^{-1}\boldsymbol{s_{xy}},$$

The corresponding PLS prediction score is deduced:

$$\begin{aligned} L_{\mathrm{PLS}}(\boldsymbol{X}, m) &\equiv (\boldsymbol{X} - \bar{\boldsymbol{X}})' \boldsymbol{D}_{\boldsymbol{s}}^{-1} \hat{\boldsymbol{\beta}}_{\mathrm{PLS},m}, \\ &\equiv (\boldsymbol{X} - \bar{\boldsymbol{X}})' \boldsymbol{D}_{\boldsymbol{s}}^{-1} \Big( \sum_{i=1}^{m} b_i \hat{\boldsymbol{C}}^{i-1} \Big) \boldsymbol{D}_{\boldsymbol{s}}^{-1} \boldsymbol{s}_{\boldsymbol{xy}}. \end{aligned} \tag{5.13}$$

Introducing the eigendecomposition of $\hat{\boldsymbol{C}}$ leads to:

$$\begin{aligned} L_{\mathrm{PLS}}(\boldsymbol{X}, m) &\equiv (\boldsymbol{X} - \bar{\boldsymbol{X}})' \boldsymbol{D}_{\boldsymbol{s}}^{-1} \hat{\boldsymbol{U}} \boldsymbol{D}_{\boldsymbol{b}} \hat{\boldsymbol{U}}' \boldsymbol{D}_{\boldsymbol{s}}^{-1} \boldsymbol{s}_{\boldsymbol{xy}}, \\ &\equiv \hat{\boldsymbol{Z}}' \boldsymbol{D}_{\boldsymbol{b}} \hat{\boldsymbol{\gamma}}, \end{aligned}$$

where $\boldsymbol{D}_{\boldsymbol{b}}$ is the $p \times p$ diagonal matrix whose diagonal entries are the coordinates of the $p$-vector $\boldsymbol{h}_{\boldsymbol{b}} = (\sum_{i=1}^{m} b_i \hat{\lambda}_1^{i-1}, \ldots, \sum_{i=1}^{m} b_i \hat{\lambda}_m^{i-1}, 0, \ldots, 0)'$. Finally, up to a scaling factor, $L_{\mathrm{PLS}}(\boldsymbol{X}, m)$ belongs to $\mathcal{L}$:

$$L_{\mathrm{PLS}}(\boldsymbol{X}, m) \equiv \boldsymbol{h}_{\boldsymbol{b}}' \boldsymbol{\xi}(\hat{\boldsymbol{Z}}).$$

In the special case where $m = 1$, expression (5.13) simplifies:

$$\begin{aligned} L_{\mathrm{PLS}}(\boldsymbol{X}, m) &\equiv (\boldsymbol{X} - \bar{\boldsymbol{X}})' \boldsymbol{D}_{\boldsymbol{s}}^{-1} \boldsymbol{D}_{\boldsymbol{s}}^{-1} \boldsymbol{s}_{\boldsymbol{xy}}, \\ &\equiv L_{\mathrm{N}}(\hat{\boldsymbol{Z}}). \end{aligned}$$

## 3.4 Alternative prediction scores

At least two other existing predictors offer alternative solutions to tune the handling of dependence, although not belonging to the formerly proposed class $\mathcal{L}$: a shrinkage-based predictor (SLM, Opgen-Rhein and Strimmer (2007b)) and the principal component regression predictor (PCR, Jolliffe (1982, 2011)).

### 3.4.1 Shrinkage-based linear model (SLM)

Opgen-Rhein and Strimmer (2007b) (see also Zuber and Strimmer (2011)) proposed a doubly shrunken version of the OLS estimate $\hat{\boldsymbol{\beta}}_{\mathrm{OLS}}$ of the vector $\boldsymbol{\beta}$ of regression coefficients. Indeed, starting from the observation that $\hat{\boldsymbol{\beta}}_{\mathrm{OLS}}$ can be decomposed as follows:

$$\hat{\boldsymbol{\beta}}_{\mathrm{OLS}} = s_y \boldsymbol{D}_{\boldsymbol{s_x}}^{-1/2} \hat{\boldsymbol{C}}_{\boldsymbol{x}}^{-1} \hat{\boldsymbol{r}}_{\boldsymbol{xy}},$$

where $\boldsymbol{D}_{\boldsymbol{s_x}}$ is the $p \times p$ diagonal matrix which diagonal entries are the sample (unconditional) standard deviations of the explanatory variables, $\hat{\boldsymbol{C}}_{\boldsymbol{x}}$ is the sample correlation matrix of the explanatory variables and $\hat{\boldsymbol{r}}_{\boldsymbol{xy}}$ is the $p$-vector of sample correlations between the response and the explanatory variables, Opgen-Rhein and Strimmer (2007b) propose the following estimate $\hat{\boldsymbol{\beta}}_{\mathrm{SLM}}$ of $\boldsymbol{\beta}$:

$$\hat{\boldsymbol{\beta}}_{\mathrm{SLM}}(\tau, \tau') \;=\; s_y^{(\tau)} \boldsymbol{D}^{-1}_{\boldsymbol{s_x^{(\tau)}}} (\hat{\boldsymbol{C}}_{\boldsymbol{x}}^{(\boldsymbol{\tau'})})^{-1} \hat{\boldsymbol{r}}_{\boldsymbol{xy}}^{(\boldsymbol{\tau'})},$$

where $s_y^{(\tau)}$ and $\boldsymbol{s_x^{(\tau)}}$ are James-Stein estimates of the standard deviations of the response and of the explanatory variables respectively, with shrinkage parameter $0 < \tau < 1$, and $\hat{\boldsymbol{C}}_{\boldsymbol{x}}^{(\boldsymbol{\tau'})}$ and $\hat{\boldsymbol{r}}_{\boldsymbol{xy}}^{(\boldsymbol{\tau'})}$ are James-Stein estimates of the correlation matrix of the explanatory variables and of the $p$-vector of correlations between the response and the explanatory variables respectively, with shrinkage parameter $0 < \tau' < 1$.

In particular, $\hat{\boldsymbol{C}}_{\boldsymbol{x}}^{(\boldsymbol{\tau'})} = \tau' \mathbf{I}_p + (1 - \tau') \hat{\boldsymbol{C}}_{\boldsymbol{x}}$. It can thus be remarked that, for values of $\tau'$ close to 1, $\hat{\boldsymbol{C}}_{\boldsymbol{x}}^{(\boldsymbol{\tau'})}$ tends to the identity matrix, which amounts to ignoring the unconditional dependence among explanatory variables. However, since $\hat{\boldsymbol{r}}_{\boldsymbol{xy}}^{(\boldsymbol{\tau'})} = (1 - \tau') \hat{\boldsymbol{r}}_{\boldsymbol{xy}}$, $\lim_{\tau' \to 1} \hat{\boldsymbol{\beta}}_{\mathrm{SLM}}(\tau, \tau') = \mathbf{0}$ and therefore the naive prediction score is not reachable with this shrinkage estimation method. It can be seen that this predictor does not belong to the former class $\mathcal{L}$. First, it can be noted that $\hat{\boldsymbol{C}}_{\boldsymbol{x}}^{(\boldsymbol{\tau'})} = \hat{\boldsymbol{C}}^{(\boldsymbol{\tau'})} + (1-\tau') \hat{\boldsymbol{r}}_{\boldsymbol{xy}} \hat{\boldsymbol{r}}'_{\boldsymbol{xy}}$ where $\hat{\boldsymbol{C}}^{(\boldsymbol{\tau'})} = (1-\tau') \hat{\boldsymbol{C}} + \tau' \mathbf{I}_p$. It can then be shown that $(\hat{\boldsymbol{C}}_{\boldsymbol{x}}^{(\boldsymbol{\tau'})})^{-1} = (\hat{\boldsymbol{C}}^{(\boldsymbol{\tau'})})^{-1} - a (\hat{\boldsymbol{C}}^{(\boldsymbol{\tau'})})^{-1} \hat{\boldsymbol{r}}_{\boldsymbol{xy}} \hat{\boldsymbol{r}}'_{\boldsymbol{xy}} (\hat{\boldsymbol{C}}^{(\boldsymbol{\tau'})})^{-1}$ where $a = \left( \frac{1}{1-\tau'} + \hat{\boldsymbol{r}}'_{\boldsymbol{xy}} (\hat{\boldsymbol{C}}^{(\boldsymbol{\tau'})})^{-1} \hat{\boldsymbol{r}}_{\boldsymbol{xy}} \right)^{-1} \in \mathbb{R}$. Then, some calculus can be done on the expression of the SLM predictor:

$$\begin{aligned}
L_{\mathrm{SLM}}(\boldsymbol{X}) &\equiv (\boldsymbol{X} - \bar{\boldsymbol{X}})' s_y^{(\tau)} \boldsymbol{D}^{-1}_{\boldsymbol{s_x^{(\tau)}}} (\hat{\boldsymbol{C}}_{\boldsymbol{x}}^{(\boldsymbol{\tau'})})^{-1} \hat{\boldsymbol{r}}_{\boldsymbol{xy}}^{(\boldsymbol{\tau'})} \\
&\equiv (\boldsymbol{X} - \bar{\boldsymbol{X}})' \boldsymbol{D}^{-1}_{\boldsymbol{s_x^{(\tau)}}} (\hat{\boldsymbol{C}}_{\boldsymbol{x}}^{(\boldsymbol{\tau'})})^{-1} \hat{\boldsymbol{r}}_{\boldsymbol{xy}} \\
&\equiv [\boldsymbol{D}^{-1}_{\boldsymbol{s_x^{(\tau)}}} (\boldsymbol{X} - \bar{\boldsymbol{X}})]' [(\hat{\boldsymbol{C}}^{(\boldsymbol{\tau'})})^{-1} - a (\hat{\boldsymbol{C}}^{(\boldsymbol{\tau'})})^{-1} \hat{\boldsymbol{r}}_{\boldsymbol{xy}} \hat{\boldsymbol{r}}'_{\boldsymbol{xy}} (\hat{\boldsymbol{C}}^{(\boldsymbol{\tau'})})^{-1}] \hat{\boldsymbol{r}}_{\boldsymbol{xy}} \\
&= [\boldsymbol{D}^{-1}_{\boldsymbol{s_x^{(\tau)}}} (\boldsymbol{X} - \bar{\boldsymbol{X}})]' [(\hat{\boldsymbol{C}}^{(\boldsymbol{\tau'})})^{-1} - a (\hat{\boldsymbol{r}}'_{\boldsymbol{xy}} (\hat{\boldsymbol{C}}^{(\boldsymbol{\tau'})})^{-1} \hat{\boldsymbol{r}}_{\boldsymbol{xy}}) (\hat{\boldsymbol{C}}^{(\boldsymbol{\tau'})})^{-1}] \hat{\boldsymbol{r}}_{\boldsymbol{xy}} \\
&\equiv [\boldsymbol{D}^{-1}_{\boldsymbol{s_x^{(\tau)}}} (\boldsymbol{X} - \bar{\boldsymbol{X}})]' (\hat{\boldsymbol{C}}^{(\boldsymbol{\tau'})})^{-1} \hat{\boldsymbol{r}}_{\boldsymbol{xy}} \\
&\equiv [\boldsymbol{D}^{-1}_{\boldsymbol{s_x^{(\tau)}}} (\boldsymbol{X} - \bar{\boldsymbol{X}})]' (\hat{\boldsymbol{C}}^{(\boldsymbol{\tau'})})^{-1} \boldsymbol{D}^{-1}_{\boldsymbol{s_x}} \boldsymbol{s_{xy}}
\end{aligned}$$

Note that $(\hat{\boldsymbol{C}}^{(\boldsymbol{\tau'})})^{-1} = \hat{\boldsymbol{U}} \boldsymbol{D}^{-1}_{\boldsymbol{(1-\tau')\hat{\lambda}+\tau'}} \hat{\boldsymbol{U}}'$ where $\boldsymbol{D}^{-1}_{\boldsymbol{(1-\tau')\hat{\lambda}+\tau'}}$ is the diagonal matrix whose diagonal elements are the inverse of $(1 - \tau') \hat{\lambda}_i \mathbf{1}_{\{\hat{\lambda}_i > 0\}} + \tau'$. The predictor can then be rewritten in a form very similar to those of predictors belonging to $\mathcal{L}$; however, $\hat{\boldsymbol{U}}' \boldsymbol{D}^{-1}_{\boldsymbol{s_x^{(\tau)}}} (\boldsymbol{X} - \bar{\boldsymbol{X}})$ cannot be rewritten using $\hat{\boldsymbol{Z}}$ because of $\boldsymbol{D}^{-1}_{\boldsymbol{s_x^{(\tau)}}}$ instead of $\boldsymbol{D}_{\boldsymbol{s}}^{-1}$. Similarly, $\hat{\boldsymbol{U}}' \boldsymbol{D}^{-1}_{\boldsymbol{s_x}} \boldsymbol{s_{xy}}$ cannot be rewritten using $\hat{\boldsymbol{\gamma}}$.

Explicit formulas for optimal shrinkage parameters $\tau$ and $\tau'$ are given in Schäfer and Strimmer (2005). Even if this is computationally lighter than a cross-validation procedure, those formulas are not specifically designed to account for the pattern of association between the explanatory variables and the response.

### 3.4.2 Principal component regression predictor

Also very popular, notably for applications in chemometrics, the Principal Component Regression (PCR) prediction score, using $1 \leq m \leq q$ Principal Components (Jolliffe (1982); see also Jolliffe (2011), Chapter 8) is equivalent to (assuming, as for the Ridge and PLS predictors, that the explanatory variables are centered and scaled using their conditional standard deviations):

$$L_{\text{PCR}}(\boldsymbol{X}, m) \equiv (\boldsymbol{X} - \bar{\boldsymbol{X}})' \boldsymbol{D}_{\boldsymbol{s}}^{-1} \left( \sum_{i=1}^{m} \frac{1}{\delta_i} \boldsymbol{v_i} \boldsymbol{v_i}' \right) \boldsymbol{D}_{\boldsymbol{s}}^{-1} \boldsymbol{s_{xy}}$$

where $\boldsymbol{v_i}$ is the $i$-th eigenvector of $\hat{\boldsymbol{C}}_{\boldsymbol{x}} = \boldsymbol{D}_{\boldsymbol{s}}^{-1} \boldsymbol{S}_{\boldsymbol{x}} \boldsymbol{D}_{\boldsymbol{s}}^{-1}$, associated to the eigenvalue $\delta_i$. The term $\sum_{i=1}^{m} \boldsymbol{v_i} \boldsymbol{v_i}' / \delta_i$ is a rank-reduced approximation of the inverse of $\hat{\boldsymbol{C}}_{\boldsymbol{x}}$. Thus, PCR is based on an intermediate handling of dependence through the number $m$ of principal components, *i.e.* the number $m$ of eigenvectors of $\hat{\boldsymbol{C}}_{\boldsymbol{x}}$ retained in the construction of the predictor: as $m$ increases, $L_{\text{PCR}}(\boldsymbol{X}, m)$ gets closer to $L_{\text{OLS}}(\boldsymbol{X})$. Contrary to the PLS prediction score, the case $m = 1$ does not yield the naive prediction score $L_{\text{N}}(\boldsymbol{X})$. Moreover, the PCR predictor cannot be written as a linear combination of the elements of $\boldsymbol{\xi}(\hat{\boldsymbol{Z}})$. Indeed, it can be rewritten as:

$$L_{\text{PCR}}(\boldsymbol{X}, m) \equiv \hat{\boldsymbol{Z}}' \hat{\boldsymbol{U}}' \left( \sum_{i=1}^{m} \frac{1}{\delta_i} \boldsymbol{v_i} \boldsymbol{v_i}' \right) \hat{\boldsymbol{U}} \hat{\boldsymbol{\gamma}}$$

and $\hat{\boldsymbol{U}}' \left( \sum_{i=1}^{m} \frac{1}{\delta_i} \boldsymbol{v_i} \boldsymbol{v_i}' \right) \hat{\boldsymbol{U}}$ is not a diagonal matrix.

# 4 Optimal prediction within $\mathcal{L}$

We now propose to search for the optimal predictor within $\mathcal{L}$, *i.e.* the vector $\boldsymbol{h_{\text{opt}}}$ of weights fulfilling:

$$\boldsymbol{h_{\text{opt}}} = \underset{\boldsymbol{h} \in \mathbb{R}^p}{\text{argmax}} \, \text{cor}^2(L_{\boldsymbol{h}}(\hat{\boldsymbol{Z}}), Y).$$

Assuming the moments of $\boldsymbol{\xi}(\hat{\boldsymbol{Z}})$ are known and provided $\mathrm{Var}(\boldsymbol{\xi}(\hat{\boldsymbol{Z}}))$ is non-singular, $\boldsymbol{h}_{\mathbf{opt}}$ has the following expression:

$$\boldsymbol{h}_{\mathbf{opt}} = \left\{\mathrm{Var}(\boldsymbol{\xi}(\hat{\boldsymbol{Z}})\right\}^{-1} \mathrm{Cov}(\boldsymbol{\xi}(\hat{\boldsymbol{Z}}), Y). \tag{5.14}$$

## 4.1 Closed-form expression of the best predictor

The following Theorem gives the non-asymptotic moments of $\boldsymbol{\xi}(\hat{\boldsymbol{Z}})$.

**Theorem 4.1** *Let $\boldsymbol{S}$ be a random $p \times p$ matrix distributed as $\mathcal{W}_p(\boldsymbol{\Sigma}; n-2)/n$. Let $\boldsymbol{D_s}$ denote the $p \times p$ diagonal matrix whose diagonal entries are the square roots of the diagonal entries of $\boldsymbol{S}$. Let $\boldsymbol{U}$ denote the $p \times p$ matrix of eigenvectors of $\boldsymbol{D_s}^{-1}\boldsymbol{S}\boldsymbol{D_s}^{-1}$.*

*Under the normality assumption introduced in Section 2 for the joint distribution of the explanatory variables and the response, the expectation and variance of $\boldsymbol{\xi}(\hat{\boldsymbol{Z}})$ are as follows:*

$$\begin{aligned}
\mathbb{E}\left[\boldsymbol{\xi}(\hat{\boldsymbol{Z}})\right] &= \mathbf{0}, \\
\mathrm{Var}(\boldsymbol{\xi}(\hat{\boldsymbol{Z}})) &= \frac{n+1}{n-1}\frac{\sigma_y^2}{n}\,\mathbb{E}\left\{\left(\boldsymbol{U}'\boldsymbol{D_s}^{-1}\boldsymbol{\Sigma_x}\boldsymbol{D_s}^{-1}\boldsymbol{U}\right)^{\odot 2}\right\} + \\
&\quad \frac{n+1}{n-1}\,\mathbb{E}\left\{\left(\boldsymbol{U}'\boldsymbol{D_s}^{-1}\boldsymbol{\Sigma_x}\boldsymbol{D_s}^{-1}\boldsymbol{U}\right) \odot \left(\boldsymbol{U}'\boldsymbol{D_s}^{-1}\boldsymbol{\sigma_{xy}}\boldsymbol{\sigma_{xy}}'\boldsymbol{D_s}^{-1}\boldsymbol{U}\right)\right\},
\end{aligned}$$

*Moreover, the covariance between the response variable $Y$ and $\boldsymbol{\xi}(\hat{\boldsymbol{Z}})$ has the following expression:*

$$\mathrm{Cov}\left\{\boldsymbol{\xi}(\hat{\boldsymbol{Z}}), Y\right\} = \mathbb{E}\left\{\left(\boldsymbol{U}'\boldsymbol{D_s}^{-1}\boldsymbol{\sigma_{xy}}\right)^{\odot 2}\right\}.$$

*Proof*: see Section 10.

Due to its complexity, the above expression of $\mathrm{Var}(\boldsymbol{\xi}(\hat{\boldsymbol{Z}}))$ does not give much insight into the conditions under which it is nonsingular, which is required for the calculation of $\boldsymbol{h}_{\mathbf{opt}}$ (see equation (5.14)). Hereafter, in cases where $\mathrm{Var}(\boldsymbol{\xi}(\hat{\boldsymbol{Z}}))$ turns out to be singular, its inverse is replaced by its Moore-Penrose generalized inverse in expression (5.14) of $\boldsymbol{h}_{\mathbf{opt}}$.

The prediction performance of $L_{\boldsymbol{h}_{\text{opt}}}(\hat{\boldsymbol{Z}})$, measured by its squared correlation $\mathrm{R}^2_{\boldsymbol{h}_{\text{opt}}}$ with the response, can straightforwardly be deduced from Theorem 4.1. It turns out to depend both on the vector $\boldsymbol{\sigma_{xy}}$ of one-to-one association parameters between the explanatory variables and the response and on the conditional dependence across explanatory variables in a non-trivial manner through expectations of quadratic forms in the coordinates of $\boldsymbol{U'D_s^{-1}\sigma_{xy}}$. The expectation of the former vector $\boldsymbol{U'D_s^{-1}\sigma_{xy}}$ approximates the vector $\boldsymbol{\gamma}$ of one-to-one association parameters between the whitened explanatory variables and the response.

To illustrate how optimal prediction within $\mathcal{L}$ adapts to the specific combination of $\boldsymbol{\gamma}$ and $\boldsymbol{C}$, let us consider the two situations introduced in Section 2, where the conditional covariance matrix of the explanatory variables is the same as estimated using data from Lu et al. (2004) and two association signals are considered: in scenario 1, the naive predictor is clearly outperformed by the OLS predictor whereas in scenario 2, the naive predictor shows equivalent asymptotic performance and turns out to outperform the OLS predictor in non-asymptotic conditions (see Table 5.1).

For both scenarios, Theorem 4.1 is used to derive the squared correlation between the response and the optimal predictor $L_{\boldsymbol{h}_{\text{opt}}}(\hat{\boldsymbol{Z}})$, the moments of $\boldsymbol{\xi}(\hat{\boldsymbol{Z}})$ being estimated using 1,000 simulated Monte-Carlo samples. A sequence of values for the sample size $n$ going from $n = 50$ ($n \ll p$) to $n = 1,000$ ($n \gg p$) is considered. Figure 5.3 shows how the prediction performance varies along $n$ in both scenarios for the optimal, OLS and naive predictors. It appears that the prediction performance of the optimal predictor is equivalent to the best one, either the OLS predictor in scenario 1 or the naive predictor in scenario 2. Therefore, it can be used to adapt to the specific combination of a conditional dependence structure and a pattern of association between the response and the predictors.

## 4.2 Estimation of the optimal predictor within $\mathcal{L}$

The most straightforward idea to take advantage of the expressions of $\mathrm{Var}(\boldsymbol{\xi}(\hat{\boldsymbol{Z}}))$ and $\mathrm{Cov}(\boldsymbol{\xi}(\hat{\boldsymbol{Z}}), Y)$ given in Theorem 4.1 is to plug-in the sample estimates of $\boldsymbol{\Sigma_x}, \boldsymbol{\sigma_{xy}}$ and $\sigma_y^2$. The vector of weights can then be estimated by using a Monte-Carlo procedure generating random matrices according to the Wishart distribution. Figure

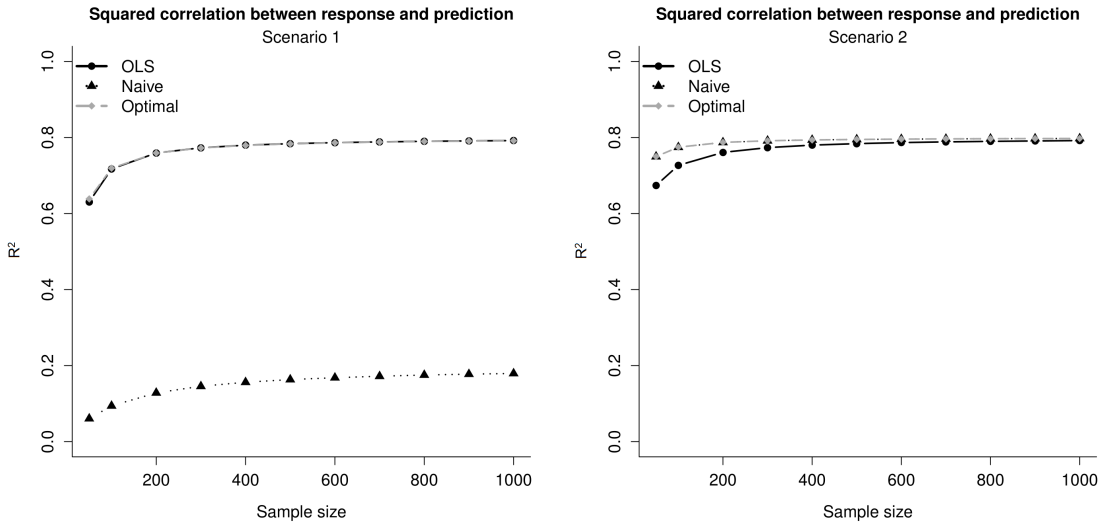


Figure 5.3: Squared correlation between the response variable and $L_{\text{OLS}}(\hat{\boldsymbol{Z}})$ (OLS), $L_{\text{N}}(\hat{\boldsymbol{Z}})$ (Naive) and the optimal predictor (Optimal) within $\mathcal{L}$ in two illustrative scenarios introduced in Section 2 based on Lu et al. (2004) data.

5.4 displays the optimal weights estimated using this approach for the prediction of the age of a patient using the 403 gene expression levels with the data introduced in Lu et al. (2004). Figure 5.4 also shows the weights corresponding to the OLS and the naive predictors. It is interesting to note that the optimal weights differ quite markedly from the weights used for the OLS predictor. Those optimal weights are generally not simple functions of the eigenvalues $\boldsymbol{\lambda}$, as for OLS, Ridge and PLS predictors.

Since the computation of the optimal weights involves time-consuming Monte-Carlo experiments, we propose an approximation based on the convergence in distribution of the unit-length eigenvectors of a sample correlation matrix of normal profiles to the unit-length eigenvectors of the corresponding population correlation matrix (see Kollo and Neudecker (1993)). Indeed, it can be deduced that:

$$\begin{aligned}
\mathbb{E}\left\{(\boldsymbol{U}'\boldsymbol{D}_{\boldsymbol{s}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{D}_{\boldsymbol{s}}^{-1}\boldsymbol{U})^{\odot 2}\right\} &= \left(\boldsymbol{D}_{\boldsymbol{\lambda}} + \frac{\boldsymbol{\gamma}\boldsymbol{\gamma}'}{\sigma_y^2}\right)^{\odot 2} + o(n) \\
\mathbb{E}\left\{\left(\boldsymbol{U}'\boldsymbol{D}_{\boldsymbol{s}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{D}_{\boldsymbol{s}}^{-1}\boldsymbol{U}\right) \odot \left(\boldsymbol{U}'\boldsymbol{D}_{\boldsymbol{s}}^{-1}\boldsymbol{\sigma}_{\boldsymbol{xy}}\boldsymbol{\sigma}_{\boldsymbol{xy}}'\boldsymbol{D}_{\boldsymbol{s}}^{-1}\boldsymbol{U}\right)\right\} &= \left(\boldsymbol{D}_{\boldsymbol{\lambda}} + \frac{\boldsymbol{\gamma}\boldsymbol{\gamma}'}{\sigma_y^2}\right) \odot \left(\boldsymbol{\gamma}\boldsymbol{\gamma}'\right) + o(n), \\
\mathbb{E}\left\{\left(\boldsymbol{U}'\boldsymbol{D}_{\boldsymbol{s}}^{-1}\boldsymbol{\sigma}_{\boldsymbol{xy}}\right)^{\odot 2}\right\} &= \boldsymbol{\gamma}^{\odot 2} + o(n).
\end{aligned}$$

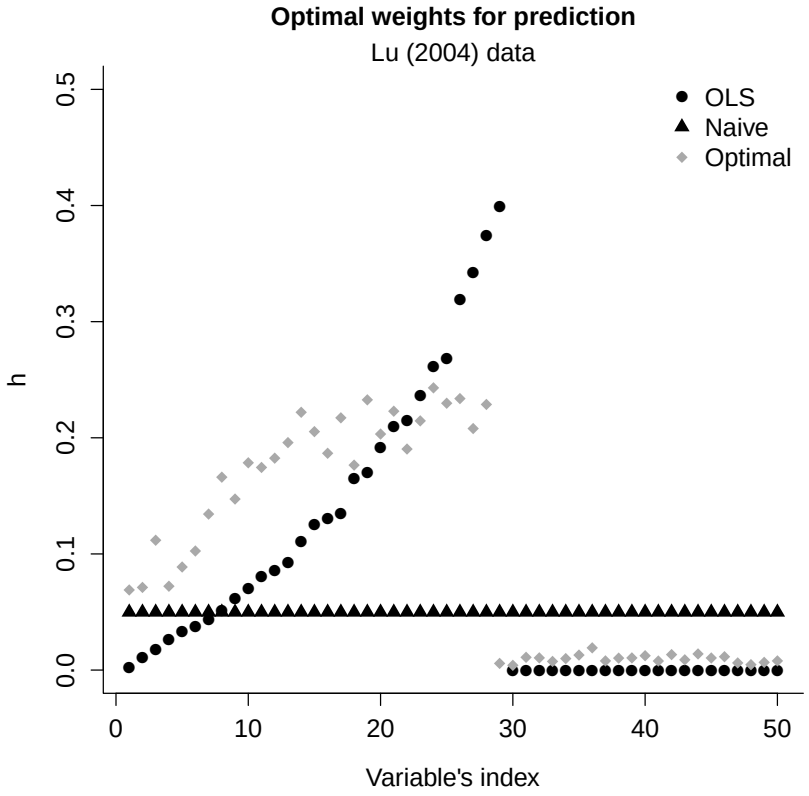


Figure 5.4: Weights for the prediction of the age of a patient using the 403 gene expression levels with the data introduced in Lu et al. (2004) by the optimal predictor in $\mathcal{L}$ and the OLS and naive predictors. The weights after the 50th explanatory variable are not shown since they remain constant until the 403rd value.

Therefore:

$$\begin{aligned} \mathrm{Var}(\boldsymbol{\xi}(\hat{\boldsymbol{Z}})) &= \left(\boldsymbol{D}_{\boldsymbol{\lambda}} + \frac{\boldsymbol{\gamma}\boldsymbol{\gamma}'}{\sigma_y^2}\right) \odot \left(\boldsymbol{\gamma}\boldsymbol{\gamma}'\right) + o(n), \\ \mathrm{Cov}\left\{\boldsymbol{\xi}(\hat{\boldsymbol{Z}}), Y\right\} &= \boldsymbol{\gamma}^{\odot 2} + o(n). \end{aligned}$$

Consistently, the following alternative estimator of the vector of optimal weights is proposed:

$$\begin{aligned} \hat{\boldsymbol{h}}_{\mathbf{opt}} &= \left[\left(\boldsymbol{D}_{\hat{\boldsymbol{\lambda}}} + \frac{\hat{\boldsymbol{\gamma}}\hat{\boldsymbol{\gamma}}'}{s_y^2}\right) \odot \left(\hat{\boldsymbol{\gamma}}\hat{\boldsymbol{\gamma}}'\right)\right]^{-} \hat{\boldsymbol{\gamma}}^{\odot 2}, \\ &= \left(\boldsymbol{D}_{\hat{\boldsymbol{\lambda}}\hat{\boldsymbol{\gamma}}^{\odot \mathbf{2}}} + \frac{\hat{\boldsymbol{\gamma}}^{\odot 2}\hat{\boldsymbol{\gamma}}^{\odot 2'}}{s_y^2}\right)^{-} \hat{\boldsymbol{\gamma}}^{\odot 2}, \end{aligned}$$

where $\boldsymbol{D}_{\hat{\boldsymbol{\lambda}}\hat{\boldsymbol{\gamma}}^{\odot 2}}$ is the $p \times p$ diagonal matrix which diagonal entries are the coordinates of the vector $\hat{\boldsymbol{\lambda}} \odot \hat{\boldsymbol{\gamma}}^{\odot 2}$. However, in practice, the use of a Moore-Penrose generalized inverse involves the choice of a threshold under which the eigenvalues are set to zero. In some cases, disentangling zero and non-zero eigenvalues being not obvious, we introduce the number of positive eigenvalues as an hyperparameter of our method. In the comparative studies conducted in the next Section, this hyperparameter is tuned by cross-validation. In the following, the predictor associated to this procedure and using the vector of weights $\hat{\boldsymbol{h}}_{\mathbf{opt}}$ is denoted $L_{\hat{\boldsymbol{h}}_{\mathbf{opt}}}$.

# 5 Comparative study

The proposed adaptive method is now compared to the OLS, Naive, Lasso, Ridge, PLS, PCR and SLM methods. When a hyperparameter is required (Lasso, Ridge, PLS, $L_{\hat{\boldsymbol{h}}_{\mathbf{opt}}}$), it is optimized by a 10-fold cross validation procedure. The PLS and PCR methods are implemented using the R package `pls` (Mevik et al., 2019), the SLM method is implemented using the R package `care` (Zuber and Strimmer, 2017) and the Ridge and Lasso methods are implemented using the R package `glmnet` (Friedman et al., 2010).

## 5.1 Simulation study

### 5.1.1 Data-driven simulation setup

First, 100,000 random $p$-vectors of explanatory variables are generated according to the normal distribution, with mean $\mathbf{0}_p$ and variance-covariance $\boldsymbol{\Sigma}_{\boldsymbol{x}}$, where $\boldsymbol{\Sigma}_{\boldsymbol{x}}$ is the correlation matrix estimated on a dataset consisting of 124 near-infrared spectra ($p = 256$ wavelengths) of wine samples. The data table is freely available in the R package `cggd` (Zhang and Melnik, 2012). Figure 5.5 displays a heatmap of $\boldsymbol{\Sigma}_{\boldsymbol{x}}$, showing a two-block structure, with strong within-block positive correlation. The two blocks are themselves correlated, with both positive and negative correlation coefficients depending on the positions of the variables.

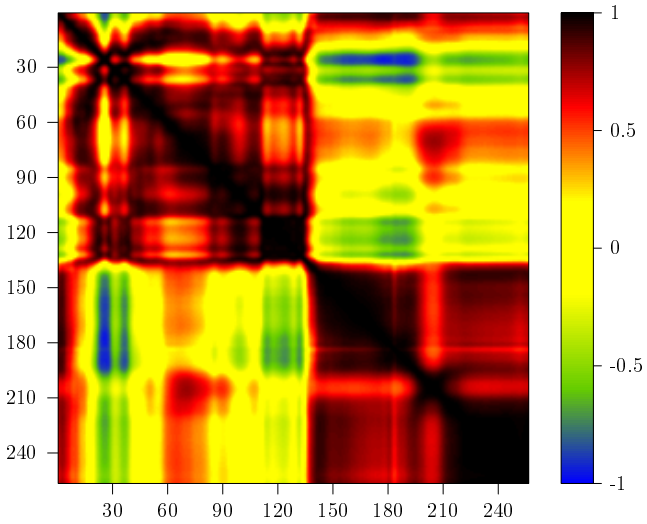


Figure 5.5: Correlation matrix of the explanatory variables in the simulation study (estimated on the `wine` dataset from the R package `cggd` (Zhang and Melnik, 2012))

For each profile $\boldsymbol{X}$ of explanatory variables, the corresponding value of the response variable $Y$ is generated by the standard linear regression model:

$$Y = \boldsymbol{X}'\boldsymbol{\beta} + \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, \sigma^2)$$

where the values of $\boldsymbol{\beta}$ and $\sigma^2$ are discussed below.

Several scenarios are considered for the association between the explanatory variables and the response, either by setting the value of $\boldsymbol{\beta}$ or alternatively $\boldsymbol{\sigma_{xy}}$:

- scenario 1: $\boldsymbol{\beta} = (1, 0, \ldots, 0)'$
- scenario 2: $\boldsymbol{\beta}$ contains 50 coefficients equal to 1 and 50 coefficients equal to $-1$; the remaining coefficients are equal to 0
- scenario 3: $\boldsymbol{\sigma_{xy}} = (1, 0, \ldots, 0)'$
- scenario 4: $\boldsymbol{\sigma_{xy}}$ contains 50 coefficients equal to 1 and 50 coefficients equal to $-1$; the remaining coefficients are equal to 0.

For the last two scenarios, $\boldsymbol{\beta}$ is deduced from $\boldsymbol{\sigma_{xy}}$ using the following expression: $\boldsymbol{\beta} = \boldsymbol{\Sigma_x^- \sigma_{xy}}$. Because of the strong correlation across the explanatory variables, defining the association signal by choosing $\boldsymbol{\beta}$ or $\boldsymbol{\sigma_{xy}}$ yields very different patterns after converting the signal vectors on the same scale, either $\boldsymbol{\beta}$ or $\boldsymbol{\sigma_{xy}}$. In particular, a sparse (resp. non-sparse) $\boldsymbol{\beta}$ is associated to a non-sparse (resp. sparse) $\boldsymbol{\sigma_{xy}}$.

The signal-to-noise ratio (SNR) is defined as $\text{var}(\boldsymbol{X'\beta})/\text{var}(\varepsilon) = \boldsymbol{\beta' \Sigma_x \beta}/\sigma^2$. A sequence of increasing values is chosen for the SNR, obtained by *ad-hoc* values of $\sigma^2$. Training samples are obtained by randomly sampling 100 individuals. For each training sample, a validation sample is obtained by randomly sampling $10,000$ individuals (with no overlap with the training sample). For each combination of an association signal and a SNR, 100 pairs of training/validation samples are generated. For each pair of a training and a validation sample, the root mean squared error of prediction (RMSEP) between the predicted values using the explanatory variables in the validation sample and the corresponding response values is calculated for each of the eight methods introduced above.

### 5.1.2 Results

Figure 5.6 shows the mean RMSEP along the SNR over 100 replications for each prediction method. The first interesting result is that, depending on the scenario, the ranking of the methods is completely different. Indeed,

- In scenarios 1 and 2: the OLS predictor has by far the lowest prediction performance, while the Lasso, SLM, PCR and PLS predictors clearly outperform the other methods. The naive predictor turns out to be better than the OLS predictor, but does not perform as well as the best methods. The performance of the Ridge predictor is similar to that of the naive one.

- In scenarios 3 and 4: the OLS and PLS predictors are now by far better than the other prediction methods. In particular, the naive, SLM and Ridge predictors are totally unable to predict the response. The Lasso predictor clearly does not perform as well as under scenarios 1 and 2, which can be explained by the fact that $\boldsymbol{\beta}$ is not sparse here. PCR performs much better than these approaches, although not as well as the PLS and OLS predictors.

This illustrates the fact that, for a given dependence pattern, the best method can be different according to the pattern of association between the explanatory variables and the response. However, in any scenario, the proposed adaptive method remains among the best prediction methods.

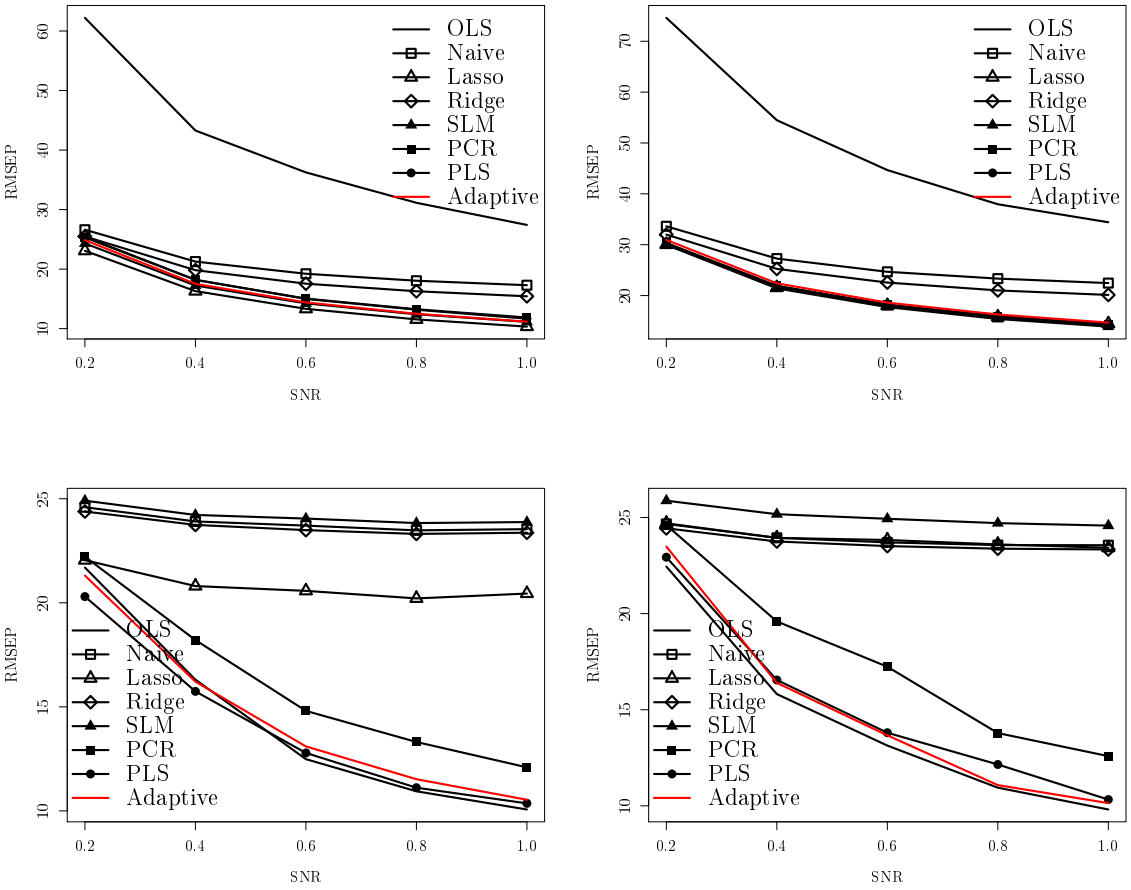


Figure 5.6: Simulation study results: mean RMSEP for scenarios 1 to 4 (the proposed method is represented in red)

## 5.2 Performance comparisons on datasets

The same prediction methods as in the above simulation study are now compared on public datasets by their RMSEP estimated in a 10-fold cross validation setup. The cross-validation procedure is repeated 50 times on random splittings to give an insight of the variability of the performances of each method.

### 5.2.1 Orange juice near infrared spectra

The first dataset used for the comparisons is available in the `cggd` R package (Zhang and Melnik, 2012) and was initially described and analyzed in Li et al. (1996). This dataset contains 218 near infrared spectra of samples of orange juice measured between 1100 and 2500 nm at 2 nm intervals. Three spectra are outliers and are removed from the dataset. Finally, the dataset is composed of 215 individuals and 700 explanatory variables. Based on these spectra, we aim at predicting the level of saccharose of the juice.

The results are displayed on Figure 5.7. For each method, a boxplot of the 50 RMSEP values estimated by 10-fold cross validation is given. Quite surprisingly as $n \ll p$, the OLS predictor performs well on this dataset, compared to other approaches. Its performance is indeed similar to those of the Lasso and Ridge predictors, though these two methods are dedicated to high dimensional datasets. The naive predictor is by far the weakest one. $L_{\hat{\boldsymbol{h}}_{\mathbf{opt}}}$ is the best one with PLS and PCR, these three methods showing similar performance.

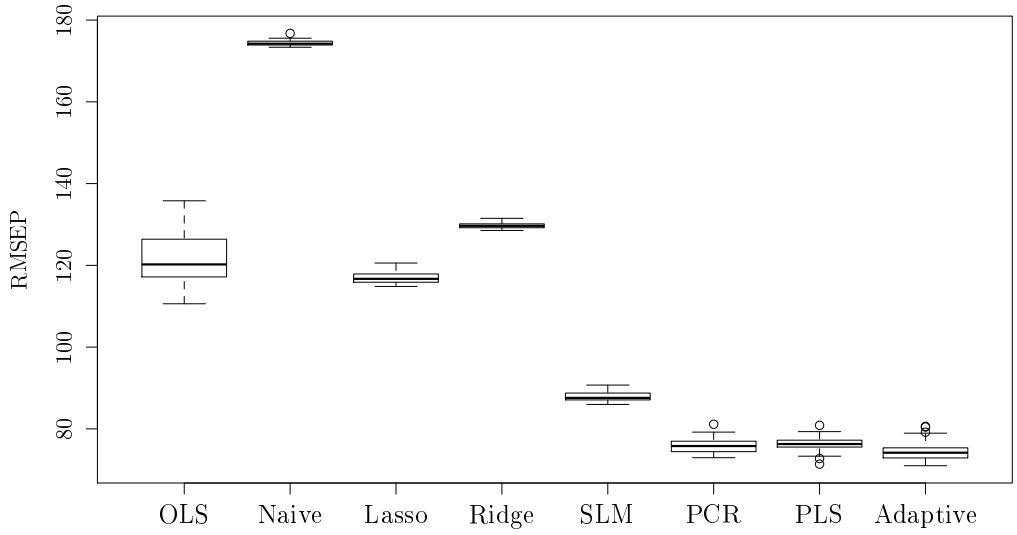


Figure 5.7: Boxplots of 10-fold cross-validation RMSEP for the orange juice dataset

### 5.2.2 Soil near infrared spectra

The second dataset used for the comparisons is available in the `prospectr` R package (Stevens and Ramirez-Lopez, 2013) and was namely used for the "Chimiometrie 2006" challenge (Pierna and Dardenne, 2008; Minasny and McBratney, 2008). This dataset contains 645 absorbance spectra of samples of soil measured between 1100 and 2500 nm at 2 nm intervals. Six spectra are outliers and are removed from the dataset. The response variable is the level of total nitrogen in g/Kg of dry soil.

The results are displayed on Figure 5.8. Contrary to the previous dataset, the OLS predictor is by far the weakest one. The naive predictor is clearly better than the OLS. By looking only at these two predictors, one could conclude that ignoring dependence yields better performance on this dataset. Nevertheless, adaptive approaches, namely PLS, PCR and $L_{\hat{\boldsymbol{h}}_{\mathbf{opt}}}$ outperform the naive one. This demonstrates that even in a context where ignoring dependence might yield better performances than completely taking it into account, determining the amount of dependence to take into account in an adaptive way can outperform the two extremes. Quite surprisingly, the results for the OLS and naive predictors are very different from the previous dataset, whereas the ranking of the other predictors is very similar.

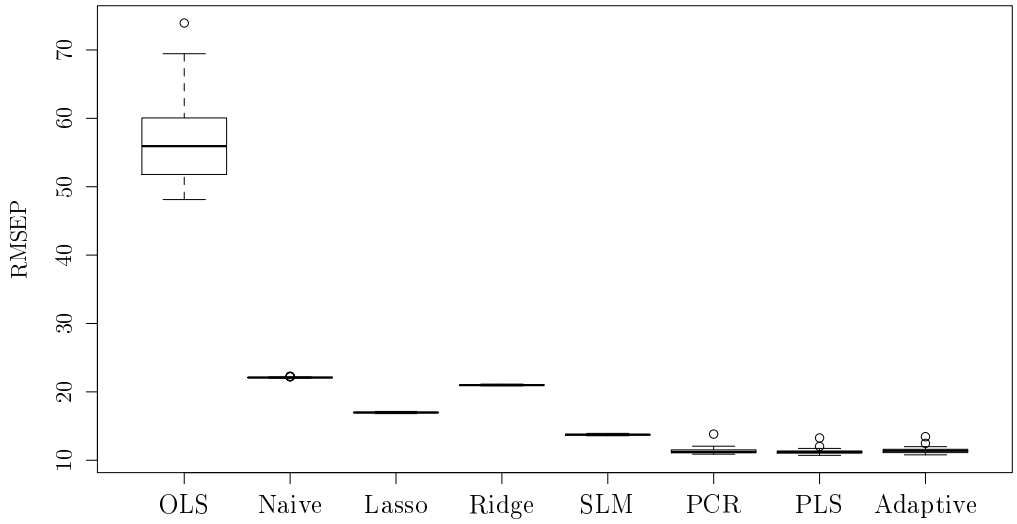


Figure 5.8: Boxplots of 10-fold cross-validation RMSEP for the soil dataset

# 6 Discussion

How to handle dependence in large and high dimensional prediction issues should not be restricted to a choice between a naive approach in which dependence across explanatory variables is ignored and an opposite approach in which explanatory variables are fully whitened as in the standard Ordinary Least Squares method. Unfortunately, for any given dependence structure, there is no uniformly best dependence handling strategy over all patterns of association signals and similarly, for any given pattern of association, there is no uniformly best strategy over all dependence structures. Indeed, the present chapter aims at demonstrating that the choice of a strategy to handle dependence has to account for the interplay between the conditional dependence of the explanatory variables and the pattern of association between the response and the explanatory variables. To illustrate this point, for an arbitrary conditional dependence pattern, a closed-form expression of an association signal for which ignoring dependence is optimal is even given in Section 2.

Some well-known prediction methods, such as Ridge and PLS, designed to address estimation of linear regression models in high-dimension, can be viewed as intermediate strategies in which dependence is just partly reduced. Although both of these methods show good prediction performance in many situations, the conditions under which one has to be preferred with respect to the other are very unclear. The class $\mathcal{L}$ of predictors introduced in Section 3 defines a general framework for the comparison of the naive, OLS, Ridge and PLS prediction methods and more generally offers a wide scope of dependence handling strategies. Each predictor in $\mathcal{L}$ is a linear combination of the coordinates of a random vector $\boldsymbol{\xi}(\hat{\boldsymbol{Z}})$. Whereas OLS, Ridge with regularization parameter $\kappa > 0$ and PLS with $m$ latent components correspond to linear coefficients defined as simple functions of the eigenvalues $\hat{\lambda}_j$ of the conditional correlation matrix of the explanatory variables solely, respectively proportional to $1/\hat{\lambda}_j$, $1/(\hat{\lambda}_j + \kappa)$ and a $m$-th order polynomial in $\hat{\lambda}_j$, the explicit expression of the optimal weights turns out to depend in a more complex way both on $\hat{\boldsymbol{\lambda}}$ and on the association signal through a vector $\hat{\boldsymbol{\gamma}}$.

The optimal predictor within $\mathcal{L}$ straightforwardly deduced from this explicit expression of the optimal weights adapts to various combinations of a dependence structure and a pattern of association signal, in the comparative studies conducted

in Section 5. Clearly, although these first results are very promising, much remains to be done to improve the estimation of the optimal weights, our proposition of a moment estimator raising numerical issues that may generate a strong variability in some high-dimensional situations. One idea could be to propose a non-parametric model for the relationship between the optimal weights and $\hat{\boldsymbol{\lambda}}$, that would extend the parametric approaches by Ridge and PLS.

Furthermore, the theoretical comparison with Ridge and PLS can also be investigated more deeply. Indeed, conditions under which the prediction performance of those methods are close to optimal can be stated explicitly. The explicit expression of the squared correlation between the response and the predictors can also be used to define optimization procedures for the regularization parameter or the number of latent components, without requiring cross-validation.

Finally, the present chapter focuses on prediction of a real-valued response but similar results can straightforwardly be deduced for the case of a two-class response variable, extending the results of Dudoit et al. (2002); Bickel and Levina (2004).

# 7 Appendix: proof of Theorem 2.1

$g(\tilde{\boldsymbol{\sigma}}_{\boldsymbol{xy}}, \boldsymbol{C})$ can alternatively be expressed as follows:

$$\begin{aligned} g(\tilde{\boldsymbol{\sigma}}_{\boldsymbol{xy}}, \boldsymbol{C}) &= \frac{\tilde{\boldsymbol{\sigma}}'_{\boldsymbol{xy}} \boldsymbol{C} \tilde{\boldsymbol{\sigma}}_{\boldsymbol{xy}} \tilde{\boldsymbol{\sigma}}'_{\boldsymbol{xy}} \boldsymbol{C}^{-1} \tilde{\boldsymbol{\sigma}}_{\boldsymbol{xy}}}{\tilde{\boldsymbol{\sigma}}'_{\boldsymbol{xy}} \tilde{\boldsymbol{\sigma}}_{\boldsymbol{xy}})^2} \\ &= \frac{\boldsymbol{\gamma}' \boldsymbol{D}_{\boldsymbol{\lambda}} \boldsymbol{\gamma} \boldsymbol{\gamma}' \boldsymbol{D}_{\boldsymbol{\lambda}}^{-1} \boldsymbol{\gamma}}{(\boldsymbol{\gamma}' \boldsymbol{\gamma})^2}, \\ &= \frac{\sum_{j=1}^{p} \lambda_j \gamma_j^2 \sum_{j=1}^{p} \frac{\gamma_j^2}{\lambda_j}}{\left( \sum_{j=1}^{p} \gamma_j^2 \right)^2} \\ &:= g(\boldsymbol{\gamma}, \boldsymbol{\lambda}). \end{aligned} \tag{5.15}$$

A sharp upper bound for $g(\boldsymbol{\gamma}, \boldsymbol{\lambda})$ over all possible $\boldsymbol{\gamma}$ is then deduced from expression (5.15). Let us denote $w_j = \gamma_j^2$ and rewrite $g(\boldsymbol{\gamma}, \boldsymbol{\lambda})$ as a function of $\boldsymbol{w}$ as:

$$h(\boldsymbol{w}, \boldsymbol{\lambda}) = \frac{\sum_{j=1}^{p} \lambda_j w_j \sum_{j=1}^{p} \frac{w_j}{\lambda_j}}{\left(\sum_{j=1}^{p} w_j\right)^2}.$$

The partial derivative of $h(\boldsymbol{w}, \boldsymbol{\lambda})$ with respect to $w_k$ has the following expression:

$$\frac{\partial h(\boldsymbol{w}, \boldsymbol{\lambda})}{\partial w_k} = \frac{\left(\lambda_k \sum_{j=1}^{p} \frac{w_j}{\lambda_j} + \frac{1}{\lambda_k} \sum_{j=1}^{p} \lambda_j w_j\right)\left(\sum_{j=1}^{p} w_j\right) - 2\left(\sum_{j=1}^{p} \lambda_j w_j \sum_{j=1}^{p} \frac{w_j}{\lambda_j}\right)}{\left(\sum_{j=1}^{p} w_j\right)^3}.$$

Thus, the gradient vector has the following expression:

$$\nabla h(\boldsymbol{w}, \boldsymbol{\lambda}) = \frac{(\boldsymbol{\lambda}\boldsymbol{\lambda}^{-\mathbf{1}\prime} + \boldsymbol{\lambda}^{-\mathbf{1}}\boldsymbol{\lambda}')\boldsymbol{w}\boldsymbol{w}'\mathbf{1}_p - 2(\boldsymbol{\lambda}'\boldsymbol{w}\boldsymbol{\lambda}^{-\mathbf{1}\prime}\boldsymbol{w})\mathbf{1}_p}{\left(\sum_{j=1}^{p} w_j\right)^3}.$$

Equating the gradient to zero then yields

$$\begin{aligned} & (\boldsymbol{\lambda}\boldsymbol{\lambda}^{-\mathbf{1}\prime} + \boldsymbol{\lambda}^{-\mathbf{1}}\boldsymbol{\lambda}')\boldsymbol{w}\boldsymbol{w}'\mathbf{1}_p - 2(\boldsymbol{\lambda}'\boldsymbol{w}\boldsymbol{\lambda}^{-\mathbf{1}\prime}\boldsymbol{w})\mathbf{1}_p &&= \mathbf{0} \\ \Leftrightarrow \quad & \left[(\boldsymbol{\lambda}\boldsymbol{\lambda}^{-\mathbf{1}\prime} + \boldsymbol{\lambda}^{-\mathbf{1}}\boldsymbol{\lambda}')\boldsymbol{w}\boldsymbol{w}' - 2(\boldsymbol{\lambda}'\boldsymbol{w}\boldsymbol{\lambda}^{-\mathbf{1}\prime}\boldsymbol{w})\mathbf{I}_p\right]\mathbf{1}_p &&= \mathbf{0} \\ \Leftrightarrow \quad & (\boldsymbol{\lambda}\boldsymbol{\lambda}^{-\mathbf{1}\prime} + \boldsymbol{\lambda}^{-\mathbf{1}}\boldsymbol{\lambda}')\boldsymbol{w} &&= \frac{2\boldsymbol{\lambda}'\boldsymbol{w}\boldsymbol{\lambda}^{-\mathbf{1}\prime}\boldsymbol{w}}{\boldsymbol{w}'\boldsymbol{w}}\boldsymbol{w} \end{aligned}$$

and consequently the vector $\boldsymbol{w}$ maximizing $h(\boldsymbol{w}, \boldsymbol{\lambda})$ is an eigenvector of $\boldsymbol{\lambda}\boldsymbol{\lambda}^{-\mathbf{1}\prime} + \boldsymbol{\lambda}^{-\mathbf{1}}\boldsymbol{\lambda}'$. Since $w_j = \gamma_j^2$, the vector $\boldsymbol{\gamma}$ for which $g(\boldsymbol{\gamma}, \boldsymbol{\lambda})$ reaches its upper limit is the vector whose coordinates are the square roots of the coordinates of $\boldsymbol{w}$.

It can be seen that

$$\boldsymbol{\lambda}\boldsymbol{\lambda}^{-\mathbf{1}\prime} + \boldsymbol{\lambda}^{-\mathbf{1}}\boldsymbol{\lambda}' = \frac{1}{2}[\boldsymbol{\lambda} + \boldsymbol{\lambda}^{-\mathbf{1}}, \boldsymbol{\lambda} - \boldsymbol{\lambda}^{-\mathbf{1}}] \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} [\boldsymbol{\lambda} + \boldsymbol{\lambda}^{-\mathbf{1}}, \boldsymbol{\lambda} - \boldsymbol{\lambda}^{-\mathbf{1}}]'$$

which shows that $\boldsymbol{\lambda}\boldsymbol{\lambda}^{-\mathbf{1}\prime} + \boldsymbol{\lambda}^{-\mathbf{1}}\boldsymbol{\lambda}'$ has only one positive eigenvalue and that the coordinates of the corresponding eigenvector $\boldsymbol{v}(\boldsymbol{\lambda})$ are all positive, thus allowing to take their square roots. Consequently, the vector $\boldsymbol{\gamma}$ for which $g(\boldsymbol{\gamma}, \boldsymbol{\lambda})$ reaches its

upper limit is the vector whose coordinates are the square roots of the coordinates of $\boldsymbol{v}(\boldsymbol{\lambda})$.

Moreover, for any vector $\boldsymbol{\gamma}$ with only one nonzero coordinate, the corresponding value of $g(\boldsymbol{\gamma}, \boldsymbol{\lambda})$ equals one and $\mathrm{R}^2_{\mathrm{N}}$ reaches its upper limit.

# 8 Appendix: proof of Theorem 3.1

First, let us reformulate expression (5.12) of $\hat{\boldsymbol{\beta}}_\kappa$ by introducing the conditional covariance matrix $\boldsymbol{S}$:

$$\begin{aligned} \boldsymbol{D_s}^{-1}\boldsymbol{S_x}\boldsymbol{D_s}^{-1} + \kappa \mathbf{I}_p &= \boldsymbol{D_s}^{-1}\boldsymbol{S}\boldsymbol{D_s}^{-1} + \frac{\boldsymbol{D_s}^{-1}\boldsymbol{s_{xy}}\boldsymbol{s'_{xy}}\boldsymbol{D_s}^{-1}}{s_y^2} + \kappa \mathbf{I}_p, \\ &= \hat{\boldsymbol{C}} + \frac{\boldsymbol{D_s}^{-1}\boldsymbol{s_{xy}}\boldsymbol{s'_{xy}}\boldsymbol{D_s}^{-1}}{s_y^2} + \kappa \mathbf{I}_p. \end{aligned}$$

Therefore, using the Sherman-Morrison identity (see Hager (1989), equation (2)) for the inversion of the former matrix,

$$(\boldsymbol{D_s}^{-1}\boldsymbol{S_x}\boldsymbol{D_s}^{-1} + \kappa \mathbf{I}_p)^{-1} = (\hat{\boldsymbol{C}} + \kappa \mathbf{I}_p)^{-1} - \frac{(\hat{\boldsymbol{C}} + \kappa \mathbf{I}_p)^{-1}\boldsymbol{D_s}^{-1}\boldsymbol{s_{xy}}\boldsymbol{s'_{xy}}\boldsymbol{D_s}^{-1}(\hat{\boldsymbol{C}} + \kappa \mathbf{I}_p)^{-1}}{s_y^2 + \boldsymbol{s'_{xy}}\boldsymbol{D_s}^{-1}(\hat{\boldsymbol{C}} + \kappa \mathbf{I}_p)^{-1}\boldsymbol{D_s}^{-1}\boldsymbol{s_{xy}}}.$$

Expression (5.12) of $\hat{\boldsymbol{\beta}}_\kappa$ can therefore be reformulated as follows:

$$\hat{\boldsymbol{\beta}}_\kappa = \frac{s_y^2}{s_y^2 + \boldsymbol{s'_{xy}}\boldsymbol{D_s}^{-1}(\hat{\boldsymbol{C}} + \kappa \mathbf{I}_p)^{-1}\boldsymbol{D_s}^{-1}\boldsymbol{s_{xy}}}(\hat{\boldsymbol{C}} + \kappa \mathbf{I}_p)^{-1}\boldsymbol{D_s}^{-1}\boldsymbol{s_{xy}}.$$

It is deduced that $\hat{\boldsymbol{\beta}}_\kappa$ and $(\hat{\boldsymbol{C}} + \kappa \mathbf{I}_p)^{-1}\boldsymbol{D_s}^{-1}\boldsymbol{s_{xy}}$ are collinear, which also implies that:

$$L_{\mathrm{Ridge}}(\boldsymbol{X}, \kappa) \equiv (\boldsymbol{X} - \bar{\boldsymbol{X}})'\boldsymbol{D_s}^{-1}(\hat{\boldsymbol{C}} + \kappa \mathbf{I}_p)^{-1}\boldsymbol{D_s}^{-1}\boldsymbol{s_{xy}}.$$

Introducing the Singular Value Decomposition of $\hat{\boldsymbol{C}}$ leads to:

$$L_{\mathrm{Ridge}}(\boldsymbol{X}, \kappa) \equiv (\boldsymbol{X} - \bar{\boldsymbol{X}})'\boldsymbol{D_s}^{-1}\hat{\boldsymbol{U}}\boldsymbol{D_\kappa}^{-1}\hat{\boldsymbol{U}}'\boldsymbol{D_s}^{-1}\boldsymbol{s_{xy}},$$

where $\boldsymbol{D_\kappa}$ is the $p \times p$ diagonal matrix which vector of diagonal entries is $\hat{\boldsymbol{\lambda}} + \kappa \mathbf{1}_p$. Finally,

$$L_{\mathrm{Ridge}}(\hat{\boldsymbol{Z}}, \kappa) \equiv \hat{\boldsymbol{Z}}'\boldsymbol{D_\kappa}^{-1}\hat{\boldsymbol{\gamma}}.$$

Therefore, up to a scaling factor, $L_{\text{Ridge}}(\hat{\boldsymbol{Z}}, \kappa)$ belongs to $\mathcal{L}$:

$$L_{\text{Ridge}}(\hat{\boldsymbol{Z}}, \kappa) \equiv \boldsymbol{h}'_{\kappa} \boldsymbol{\xi}(\hat{\boldsymbol{Z}}),$$

where the first $q$ coordinates of the weighting vector $\boldsymbol{h}_{\boldsymbol{\kappa}}$ are

$$\boldsymbol{h}_{i,\kappa} = \frac{\frac{1}{\hat{\lambda}_i + \kappa}}{\sqrt{\sum_{j=1}^{q} \frac{1}{(\hat{\lambda}_j + \kappa)^2} + \frac{p-q}{\kappa^2}}}, \ i = 1, \ldots, q,$$

and the last $p - q$ coordinates are all equal to:

$$\boldsymbol{h}_{i,\kappa} = \frac{\frac{1}{\kappa}}{\sqrt{\sum_{j=1}^{q} \frac{1}{(\hat{\lambda}_j + \kappa)^2} + \frac{p-q}{\kappa^2}}}, \ i = q+1, \ldots, p.$$

As a consequence, $\mathcal{L}_{\text{Ridge}} \subset \mathcal{L}$. Moreover, it is straightforwardly checked that $\lim_{\kappa \to +\infty} \boldsymbol{h}_{\boldsymbol{\kappa}} = (1/\sqrt{p})\mathbf{1}_p$ and $\lim_{\kappa \to 0} \boldsymbol{h}_{\boldsymbol{\kappa}} = \hat{\boldsymbol{\lambda}}^{-\mathbf{1}} / \sqrt{\hat{\boldsymbol{\lambda}}^{-1\prime} \hat{\boldsymbol{\lambda}}^{-1}}$.

# 9 Appendix: Lemma 9.1

**Lemma 9.1** *For all $p \geq 1$, let $\boldsymbol{v}$ be a $p$-vector. Let $\boldsymbol{V}$ be a $p \times p$ positive definite matrix. For all $m \geq 1$, $\mathcal{K}_m(\boldsymbol{V}; \boldsymbol{v}) = \mathcal{K}_m(\boldsymbol{V} - \boldsymbol{v}\boldsymbol{v}'; \boldsymbol{v})$.*

*Proof.* We first show using induction that $\mathcal{K}_m(\boldsymbol{V}; \boldsymbol{v}) \subseteq \mathcal{K}_m(\boldsymbol{V} - \boldsymbol{v}\boldsymbol{v}'; \boldsymbol{v})$. First, note that:

$$\begin{aligned} \boldsymbol{V}\boldsymbol{v} &= \boldsymbol{v}\boldsymbol{v}'\boldsymbol{v} + (\boldsymbol{V} - \boldsymbol{v}\boldsymbol{v}')\boldsymbol{v} \\ &= (\boldsymbol{v}'\boldsymbol{v})\boldsymbol{v} + (\boldsymbol{V} - \boldsymbol{v}\boldsymbol{v}')\boldsymbol{v}. \end{aligned}$$

Therefore $\boldsymbol{V}\boldsymbol{v} \in \text{span}\{\boldsymbol{v}, (\boldsymbol{V} - \boldsymbol{v}\boldsymbol{v}')\boldsymbol{v}\}$. Let us assume that the claim holds at rank $m$:

$$\boldsymbol{V}^m \boldsymbol{v} = \sum_{i=0}^{m} a_i (\boldsymbol{V} - \boldsymbol{v}\boldsymbol{v}')^i \boldsymbol{v}.$$

Then,

$$\begin{aligned} \boldsymbol{V}^{m+1}\boldsymbol{v} &= \boldsymbol{V}\boldsymbol{V}^m\boldsymbol{v} \\ &= \boldsymbol{V} \sum_{i=0}^{m} a_i (\boldsymbol{V} - \boldsymbol{v}\boldsymbol{v}')^i \boldsymbol{v} \\ &= ((\boldsymbol{V} - \boldsymbol{v}\boldsymbol{v}') + \boldsymbol{v}\boldsymbol{v}') \sum_{i=0}^{m} a_i (\boldsymbol{V} - \boldsymbol{v}\boldsymbol{v}')^i \boldsymbol{v} \end{aligned}$$

$$
\begin{aligned}
&= (\boldsymbol{V}-\boldsymbol{v}\boldsymbol{v}')\sum_{i=0}^{m} a_i(\boldsymbol{V}-\boldsymbol{v}\boldsymbol{v}')^i\boldsymbol{v} + \boldsymbol{v}\boldsymbol{v}'\sum_{i=0}^{m} a_i(\boldsymbol{V}-\boldsymbol{v}\boldsymbol{v}')^i\boldsymbol{v} \\
&= \sum_{i=0}^{m} a_i(\boldsymbol{V}-\boldsymbol{v}\boldsymbol{v}')^{i+1}\boldsymbol{v} + \boldsymbol{v}\sum_{i=0}^{m} a_i\boldsymbol{v}'(\boldsymbol{V}-\boldsymbol{v}\boldsymbol{v}')^i\boldsymbol{v} \\
&= \sum_{i=0}^{m} a_i(\boldsymbol{V}-\boldsymbol{v}\boldsymbol{v}')^{i+1}\boldsymbol{v} + \boldsymbol{v}\sum_{i=0}^{m} b_i \qquad \text{since} \quad \boldsymbol{v}'(\boldsymbol{V}-\boldsymbol{v}\boldsymbol{v}')^i\boldsymbol{v} \in \mathbb{R} \\
&= \sum_{i=0}^{m} a_i(\boldsymbol{V}-\boldsymbol{v}\boldsymbol{v}')^{i+1}\boldsymbol{v} + \sum_{i=0}^{m} b_i(\boldsymbol{V}-\boldsymbol{v}\boldsymbol{v}')^0\boldsymbol{v} \\
&= \sum_{i=0}^{m+1} \alpha_i(\boldsymbol{V}-\boldsymbol{v}\boldsymbol{v}')^i\boldsymbol{v}.
\end{aligned}
$$

Consequently, the claim still holds at rank $m+1$.

We now show by induction that $\mathcal{K}_m(\boldsymbol{V};\boldsymbol{v}) \supseteq \mathcal{K}_m(\boldsymbol{V}-\boldsymbol{v}\boldsymbol{v}';\boldsymbol{v})$. First,

$$
\begin{aligned}
(\boldsymbol{V}-\boldsymbol{v}\boldsymbol{v}')\boldsymbol{v} &= \boldsymbol{V}\boldsymbol{v} - \boldsymbol{v}\boldsymbol{v}'\boldsymbol{v} \\
&= (\boldsymbol{v}'\boldsymbol{v})\boldsymbol{v} + \boldsymbol{V}\boldsymbol{v}
\end{aligned}
$$

and $(\boldsymbol{V}-\boldsymbol{v}\boldsymbol{v}')\boldsymbol{v} \in \operatorname{span}\{\boldsymbol{v}, \boldsymbol{V}\boldsymbol{v}\}$. Let us assume that the claim holds at rank $m$:

$$
(\boldsymbol{V}-\boldsymbol{v}\boldsymbol{v}')^m\boldsymbol{v} = \sum_{i=0}^{m} a_i\boldsymbol{V}^i\boldsymbol{v}.
$$

Then,

$$
\begin{aligned}
(\boldsymbol{V}-\boldsymbol{v}\boldsymbol{v}')^{m+1}\boldsymbol{v} &= (\boldsymbol{V}-\boldsymbol{v}\boldsymbol{v}')(\boldsymbol{V}-\boldsymbol{v}\boldsymbol{v}')^m\boldsymbol{v} \\
&= (\boldsymbol{V}-\boldsymbol{v}\boldsymbol{v}')\sum_{i=0}^{m} a_i\boldsymbol{V}^i\boldsymbol{v} \\
&= \boldsymbol{V}\sum_{i=0}^{m} a_i\boldsymbol{V}^i\boldsymbol{v} - \boldsymbol{v}\boldsymbol{v}'\sum_{i=0}^{m} a_i\boldsymbol{V}^i\boldsymbol{v} \\
&= \sum_{i=0}^{m} a_i\boldsymbol{V}^{i+1}\boldsymbol{v} - \boldsymbol{v}\sum_{i=0}^{m} a_i\boldsymbol{v}'\boldsymbol{V}^i\boldsymbol{v} \\
&= \sum_{i=0}^{m} a_i\boldsymbol{V}^{i+1}\boldsymbol{v} - \boldsymbol{v}\sum_{i=0}^{m} b_i \qquad \text{since } \boldsymbol{v}'\boldsymbol{V}^i\boldsymbol{v} \in \mathbb{R} \\
&= \sum_{i=0}^{m} a_i\boldsymbol{V}^{i+1}\boldsymbol{v} - \sum_{i=0}^{m} b_i\boldsymbol{V}^0\boldsymbol{v} \\
&= \sum_{i=0}^{m+1} \alpha_i\boldsymbol{V}^i\boldsymbol{v}.
\end{aligned}
$$

Consequently, the claim still holds at rank $m+1$.

Finally, $\mathcal{K}_m(\boldsymbol{V};\boldsymbol{v}) = \mathcal{K}_m(\boldsymbol{V} - \boldsymbol{v}\boldsymbol{v}';\boldsymbol{v})$.

# 10 Appendix: proof of Theorem 4.1

First, let us recall that

$$\begin{pmatrix} \boldsymbol{X}^* \\ Y^* \end{pmatrix} \sim \mathcal{N}_{p+1}\left(\begin{pmatrix} \boldsymbol{\mu_x} \\ \mu_y \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma_x} & \boldsymbol{\sigma_{xy}} \\ \boldsymbol{\sigma}'_{\boldsymbol{xy}} & \sigma_y^2 \end{pmatrix}\right)$$

and that

$$\boldsymbol{X}^*|Y^* = y^* \sim \mathcal{N}_p\left(\boldsymbol{\mu_x} + \frac{1}{\sigma_y^2}\boldsymbol{\sigma_{xy}}(y^* - \mu_y), \boldsymbol{\Sigma}\right)$$

where $\boldsymbol{\Sigma} = \boldsymbol{\Sigma_x} - \frac{1}{\sigma_y^2}\boldsymbol{\sigma_{xy}}\boldsymbol{\sigma}'_{\boldsymbol{xy}}$.

For all $j = 1, \ldots, p$, the $j$th coordinate $\hat{Z}_j\hat{\gamma}_j$ of the vector $\boldsymbol{\xi}(\hat{\boldsymbol{Z}})$, where $\odot$ stands for the term-by-term product of two vectors with equal dimension, has the following conditional expectation:

$$\begin{aligned} \mathbb{E}[\hat{Z}_j\hat{\gamma}_j \mid Y^*, \boldsymbol{S}, \boldsymbol{s_{xy}}, \bar{\boldsymbol{X}}] &= \hat{\gamma}_j\,\mathbb{E}[\hat{Z}_j \mid Y^*, \boldsymbol{S}, \boldsymbol{s_{xy}}, \bar{\boldsymbol{X}}], \\ &= \hat{\gamma}_j\,\mathbb{E}[\hat{\boldsymbol{U}}'_{\boldsymbol{j}}\boldsymbol{D}_{\boldsymbol{s}}^{-1}(\boldsymbol{X^*} - \bar{\boldsymbol{X}})|Y^*, \boldsymbol{S}, \boldsymbol{s_{xy}}, \bar{\boldsymbol{X}}] \\ &= \hat{\gamma}_j\hat{\boldsymbol{U}}'_{\boldsymbol{j}}\boldsymbol{D}_{\boldsymbol{s}}^{-1}(\boldsymbol{\mu_x} - \bar{\boldsymbol{X}}) - \frac{Y^* - \mu_y}{\sigma_y^2}\hat{\gamma}_j\hat{\boldsymbol{U}}'_{\boldsymbol{j}}\boldsymbol{D}_{\boldsymbol{s}}^{-1}\boldsymbol{\sigma_{xy}}. \end{aligned}$$

Hence,

$$\mathbb{E}[\boldsymbol{\xi}(\hat{\boldsymbol{Z}}) \mid Y^*, \boldsymbol{S}, \boldsymbol{s_{xy}}, \bar{\boldsymbol{X}}] = \boldsymbol{D}_{\hat{\boldsymbol{\gamma}}}\hat{\boldsymbol{U}}'\boldsymbol{D}_{\boldsymbol{s}}^{-1}(\boldsymbol{\mu_x} - \bar{\boldsymbol{X}}) - \frac{Y^* - \mu_y}{\sigma_y^2}\boldsymbol{D}_{\hat{\boldsymbol{\gamma}}}\hat{\boldsymbol{U}}'\boldsymbol{D}_{\boldsymbol{s}}^{-1}\boldsymbol{\sigma_{xy}},$$

where $\boldsymbol{D}_{\hat{\boldsymbol{\gamma}}}$ is the $p \times p$ diagonal matrix which diagonal entries are the coordinates of $\hat{\boldsymbol{\gamma}}$. Then,

$$\mathbb{E}[\boldsymbol{\xi}(\hat{\boldsymbol{Z}}) \mid \boldsymbol{S}, \boldsymbol{s_{xy}}, \bar{\boldsymbol{X}}] = \boldsymbol{D}_{\hat{\boldsymbol{\gamma}}}\hat{\boldsymbol{U}}'\boldsymbol{D}_{\boldsymbol{s}}^{-1}(\boldsymbol{\mu_x} - \bar{\boldsymbol{X}}).$$

Since $\bar{\boldsymbol{X}}$ is independent from $\boldsymbol{S}$ and $\boldsymbol{s_{xy}}$, the conditioning can be reduced to $\boldsymbol{S}$ and $\boldsymbol{s_{xy}}$:

$$\mathbb{E}[\boldsymbol{\xi}(\hat{\boldsymbol{Z}}) \mid \boldsymbol{S}, \boldsymbol{s_{xy}}] = \boldsymbol{0}.$$

Conditionally to $\boldsymbol{S}$, $\hat{\boldsymbol{\gamma}}$ and $\hat{\boldsymbol{Z}}$ are independent. Consequently,

$$\begin{aligned}\mathbb{E}[\boldsymbol{\xi}(\hat{\boldsymbol{Z}}) \mid \boldsymbol{S}] &= \mathbb{E}[\hat{\boldsymbol{\gamma}} \mid \boldsymbol{S}] \odot \mathbb{E}[\hat{\boldsymbol{Z}} \mid \boldsymbol{S}] \\ &= \mathbf{0}.\end{aligned}$$

We finally get $\mathbb{E}[\boldsymbol{\xi}(\hat{\boldsymbol{Z}})] = \mathbf{0}$.

Similarly, the vector $\boldsymbol{\xi}(\hat{\boldsymbol{Z}})$ has the following conditional variance:

$$\begin{aligned}\operatorname{Var}(\boldsymbol{\xi}(\hat{\boldsymbol{Z}}) \mid Y^*, \boldsymbol{S}, \boldsymbol{s_{xy}}, \bar{\boldsymbol{X}}) &= \boldsymbol{D}_{\hat{\boldsymbol{\gamma}}}\hat{\boldsymbol{U}}'\boldsymbol{D}_{\boldsymbol{s}}^{-1}\operatorname{Var}(\boldsymbol{X^*} \mid Y^*, \boldsymbol{S}, \boldsymbol{s_{xy}}, \bar{\boldsymbol{X}})\boldsymbol{D}_{\boldsymbol{s}}^{-1}\hat{\boldsymbol{U}}\boldsymbol{D}_{\hat{\boldsymbol{\gamma}}} \\ &= \boldsymbol{D}_{\hat{\boldsymbol{\gamma}}}\hat{\boldsymbol{U}}'\boldsymbol{D}_{\boldsymbol{s}}^{-1}\boldsymbol{\Sigma}\boldsymbol{D}_{\boldsymbol{s}}^{-1}\hat{\boldsymbol{U}}\boldsymbol{D}_{\hat{\boldsymbol{\gamma}}}.\end{aligned}$$

Then,

$$\begin{aligned}\operatorname{Var}(\boldsymbol{\xi}(\hat{\boldsymbol{Z}}) \mid \boldsymbol{S}, \boldsymbol{s_{xy}}, \bar{\boldsymbol{X}}) &= \boldsymbol{D}_{\hat{\boldsymbol{\gamma}}}\hat{\boldsymbol{U}}'\boldsymbol{D}_{\boldsymbol{s}}^{-1}\operatorname{Var}(\boldsymbol{X^*} \mid \boldsymbol{S}, \boldsymbol{s_{xy}}, \bar{\boldsymbol{X}})\hat{\boldsymbol{U}}\boldsymbol{D}_{\hat{\boldsymbol{\gamma}}} \\ &= \boldsymbol{D}_{\hat{\boldsymbol{\gamma}}}\hat{\boldsymbol{U}}'\boldsymbol{D}_{\boldsymbol{s}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{D}_{\boldsymbol{s}}^{-1}\hat{\boldsymbol{U}}\boldsymbol{D}_{\hat{\boldsymbol{\gamma}}}\end{aligned}$$

and

$$\begin{aligned}\operatorname{Var}(\boldsymbol{\xi}(\hat{\boldsymbol{Z}}) \mid \boldsymbol{S}, \boldsymbol{s_{xy}}) &= \boldsymbol{D}_{\hat{\boldsymbol{\gamma}}}\hat{\boldsymbol{U}}'\boldsymbol{D}_{\boldsymbol{s}}^{-1}\operatorname{Var}(\boldsymbol{X^*} - \bar{\boldsymbol{X}} \mid \boldsymbol{S}, \boldsymbol{s_{xy}})\boldsymbol{D}_{\boldsymbol{s}}^{-1}\hat{\boldsymbol{U}}\boldsymbol{D}_{\hat{\boldsymbol{\gamma}}} \\ &= \boldsymbol{D}_{\hat{\boldsymbol{\gamma}}}\hat{\boldsymbol{U}}'\boldsymbol{D}_{\boldsymbol{s}}^{-1}[\operatorname{Var}(\boldsymbol{X^*} \mid \boldsymbol{S}, \boldsymbol{s_{xy}}) + \operatorname{Var}(\bar{\boldsymbol{X}} \mid \boldsymbol{S}, \boldsymbol{s_{xy}})]\boldsymbol{D}_{\boldsymbol{s}}^{-1}\hat{\boldsymbol{U}}\boldsymbol{D}_{\hat{\boldsymbol{\gamma}}} \\ &= \frac{n+1}{n}\boldsymbol{D}_{\hat{\boldsymbol{\gamma}}}\hat{\boldsymbol{U}}'\boldsymbol{D}_{\boldsymbol{s}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{D}_{\boldsymbol{s}}^{-1}\hat{\boldsymbol{U}}\boldsymbol{D}_{\hat{\boldsymbol{\gamma}}} \\ &= \frac{n+1}{n}\hat{\boldsymbol{U}}'\boldsymbol{D}_{\boldsymbol{s}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{D}_{\boldsymbol{s}}^{-1}\hat{\boldsymbol{U}} \odot (\hat{\boldsymbol{\gamma}}\hat{\boldsymbol{\gamma}}').\end{aligned}$$

It is deduced from Christensen (2015) that $\boldsymbol{s_{xy}}$ has the following unconditional expectation and variance:

$$\begin{aligned}\mathbb{E}[\boldsymbol{s_{xy}}] &= \boldsymbol{\sigma_{xy}}, \\ \operatorname{Var}(\boldsymbol{s_{xy}}) &= \frac{1}{n-1}(\sigma_y^2\boldsymbol{\Sigma}_{\boldsymbol{x}} + \boldsymbol{\sigma_{xy}}\boldsymbol{\sigma}'_{\boldsymbol{xy}}).\end{aligned}$$

Thus, by the law of total variance,

$$\operatorname{Var}(\boldsymbol{\xi}(\hat{\boldsymbol{Z}}) \mid \boldsymbol{S}) = \mathbb{E}[\operatorname{Var}(\boldsymbol{\xi}(\hat{\boldsymbol{Z}}) \mid \boldsymbol{S}, \boldsymbol{s_{xy}}) \mid \boldsymbol{S}] + \operatorname{Var}(\mathbb{E}[\boldsymbol{\xi}(\hat{\boldsymbol{Z}}) \mid \boldsymbol{S}, \boldsymbol{s_{xy}}] \mid \boldsymbol{S}).$$

As shown previously, $\mathbb{E}[\boldsymbol{\xi}(\hat{\boldsymbol{Z}}) \mid \boldsymbol{S}, \boldsymbol{s_{xy}}] = \mathbf{0}$. Consequently,

$$\begin{aligned}\operatorname{Var}(\boldsymbol{\xi}(\hat{\boldsymbol{Z}}) \mid \boldsymbol{S}) &= \mathbb{E}[\operatorname{Var}(\boldsymbol{\xi}(\hat{\boldsymbol{Z}}) \mid \boldsymbol{S}, \boldsymbol{s_{xy}}) \mid \boldsymbol{S}] \\ &= \frac{n+1}{n}\mathbb{E}[\hat{\boldsymbol{U}}'\boldsymbol{D}_{\boldsymbol{s}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{x}}\hat{\boldsymbol{U}} \odot (\hat{\boldsymbol{\gamma}}\hat{\boldsymbol{\gamma}}') \mid \boldsymbol{S}] \\ &= \frac{n+1}{n}\hat{\boldsymbol{U}}'\boldsymbol{D}_{\boldsymbol{s}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{D}_{\boldsymbol{s}}^{-1}\hat{\boldsymbol{U}} \odot \mathbb{E}[\hat{\boldsymbol{\gamma}}\hat{\boldsymbol{\gamma}}' \mid \boldsymbol{S}].\end{aligned}$$

Then,

$$\begin{aligned}
\mathbb{E}[\hat{\boldsymbol{\gamma}}\hat{\boldsymbol{\gamma}}' \mid \boldsymbol{S}] &= \mathbb{E}[\hat{\boldsymbol{U}}'\boldsymbol{D}_{\boldsymbol{s}}^{-1}\boldsymbol{s}_{\boldsymbol{xy}}\boldsymbol{s}_{\boldsymbol{xy}}'\boldsymbol{D}_{\boldsymbol{s}}^{-1}\hat{\boldsymbol{U}} \mid \boldsymbol{S}] \\
&= \hat{\boldsymbol{U}}'\boldsymbol{D}_{\boldsymbol{s}}^{-1}\,\mathbb{E}[\boldsymbol{s}_{\boldsymbol{xy}}\boldsymbol{s}_{\boldsymbol{xy}}' \mid \boldsymbol{S}]\boldsymbol{D}_{\boldsymbol{s}}^{-1}\hat{\boldsymbol{U}} \\
&= \hat{\boldsymbol{U}}'\boldsymbol{D}_{\boldsymbol{s}}^{-1}\,\mathbb{E}[\boldsymbol{s}_{\boldsymbol{xy}}\boldsymbol{s}_{\boldsymbol{xy}}']\boldsymbol{D}_{\boldsymbol{s}}^{-1}\hat{\boldsymbol{U}} \\
&= \hat{\boldsymbol{U}}'\boldsymbol{D}_{\boldsymbol{s}}^{-1}(\mathrm{Var}(\boldsymbol{s}_{\boldsymbol{xy}}) + \boldsymbol{\sigma}_{\boldsymbol{xy}}\boldsymbol{\sigma}_{\boldsymbol{xy}}')\boldsymbol{D}_{\boldsymbol{s}}^{-1}\hat{\boldsymbol{U}} \\
&= \hat{\boldsymbol{U}}'\boldsymbol{D}_{\boldsymbol{s}}^{-1}\left(\frac{1}{n-1}(\sigma_y^2\boldsymbol{\Sigma}_{\boldsymbol{x}} + \boldsymbol{\sigma}_{\boldsymbol{xy}}\boldsymbol{\sigma}_{\boldsymbol{xy}}') + \boldsymbol{\sigma}_{\boldsymbol{xy}}\boldsymbol{\sigma}_{\boldsymbol{xy}}'\right)\boldsymbol{D}_{\boldsymbol{s}}^{-1}\hat{\boldsymbol{U}} \\
&= \frac{1}{n-1}\sigma_y^2\hat{\boldsymbol{U}}'\boldsymbol{D}_{\boldsymbol{s}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{D}_{\boldsymbol{s}}^{-1}\hat{\boldsymbol{U}} + \frac{n}{n-1}\hat{\boldsymbol{U}}'\boldsymbol{D}_{\boldsymbol{s}}^{-1}\boldsymbol{\sigma}_{\boldsymbol{xy}}\boldsymbol{\sigma}_{\boldsymbol{xy}}'\boldsymbol{D}_{\boldsymbol{s}}^{-1}\hat{\boldsymbol{U}}.
\end{aligned}$$

We finally get that:

$$\begin{aligned}
\mathrm{Var}(\boldsymbol{\xi}(\hat{\boldsymbol{Z}}) \mid \boldsymbol{S}) &= \frac{n+1}{n-1}\left[\frac{\sigma_y^2}{n}\left(\hat{\boldsymbol{U}}'\boldsymbol{D}_{\boldsymbol{s}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{D}_{\boldsymbol{s}}^{-1}\hat{\boldsymbol{U}}\right)^{\odot 2}\right. \\
&\quad \left. +(\hat{\boldsymbol{U}}'\boldsymbol{D}_{\boldsymbol{s}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{D}_{\boldsymbol{s}}^{-1}\hat{\boldsymbol{U}}) \odot (\hat{\boldsymbol{U}}'\boldsymbol{D}_{\boldsymbol{s}}^{-1}\boldsymbol{\sigma}_{\boldsymbol{xy}}\boldsymbol{\sigma}_{\boldsymbol{xy}}'\boldsymbol{D}_{\boldsymbol{s}}^{-1}\hat{\boldsymbol{U}})\right].
\end{aligned}$$

The final expression for $\mathrm{Var}(\boldsymbol{\xi}(\hat{\boldsymbol{Z}}))$ is obtained by the law of total variance:

$$\begin{aligned}
\mathrm{Var}(\boldsymbol{\xi}(\hat{\boldsymbol{Z}})) &= \mathbb{E}[\mathrm{Var}(\boldsymbol{\xi}(\hat{\boldsymbol{Z}})) \mid \boldsymbol{S}] + \mathrm{Var}(\mathbb{E}[\boldsymbol{\xi}(\hat{\boldsymbol{Z}}) \mid \boldsymbol{S}]) \\
&= \mathbb{E}[\mathrm{Var}(\boldsymbol{\xi}(\hat{\boldsymbol{Z}})) \mid \boldsymbol{S}] \qquad \text{since } \mathbb{E}[\boldsymbol{\xi}(\hat{\boldsymbol{Z}}) \mid \boldsymbol{S}] = \boldsymbol{0}
\end{aligned}$$

and the result is obtained.

Finally, $\boldsymbol{\xi}(\hat{\boldsymbol{Z}})$ has the following conditional covariance with $Y^*$:

$$\mathrm{Cov}(\boldsymbol{\xi}(\hat{\boldsymbol{Z}}), Y^* \mid Y^*, \boldsymbol{S}, \boldsymbol{s}_{\boldsymbol{xy}}, \bar{\boldsymbol{X}}) = \boldsymbol{0}.$$

Then,

$$\mathrm{Cov}(\boldsymbol{\xi}(\hat{\boldsymbol{Z}}), Y^* \mid \boldsymbol{S}, \boldsymbol{s}_{\boldsymbol{xy}}, \bar{\boldsymbol{X}}) = \boldsymbol{D}_{\hat{\boldsymbol{\gamma}}}\hat{\boldsymbol{U}}'\boldsymbol{D}_{\boldsymbol{s}}^{-1}\boldsymbol{\sigma}_{\boldsymbol{xy}}$$

and

$$\mathrm{Cov}(\boldsymbol{\xi}(\hat{\boldsymbol{Z}}), Y^* \mid \boldsymbol{S}, \boldsymbol{s}_{\boldsymbol{xy}}) = \boldsymbol{D}_{\hat{\boldsymbol{\gamma}}}\hat{\boldsymbol{U}}'\boldsymbol{D}_{\boldsymbol{s}}^{-1}\boldsymbol{\sigma}_{\boldsymbol{xy}}.$$

Finally, by the law of total covariance:

$$\begin{aligned}
\mathrm{Cov}(\boldsymbol{\xi}(\hat{\boldsymbol{Z}}), Y^* \mid \boldsymbol{S}) &= \mathbb{E}[\mathrm{Cov}(\boldsymbol{\xi}(\hat{\boldsymbol{Z}}), Y^* \mid \boldsymbol{S}, \boldsymbol{s}_{\boldsymbol{xy}}) \mid \boldsymbol{S}] \\
&\quad + \mathrm{Cov}(\mathbb{E}[Y^* \mid \boldsymbol{S}, \boldsymbol{s}_{\boldsymbol{xy}}], \mathbb{E}[\boldsymbol{\xi}(\hat{\boldsymbol{Z}}) \mid \boldsymbol{S}, \boldsymbol{s}_{\boldsymbol{xy}}] \mid \boldsymbol{S}).
\end{aligned}$$

$\mathbb{E}[Y^*|\boldsymbol{S}, \boldsymbol{s}_{\boldsymbol{xy}}] = \mu_y$ is constant; consequently, its covariance with any random variable is zero. Thus,

$$\begin{aligned}
\mathrm{Cov}(\boldsymbol{\xi}(\hat{\boldsymbol{Z}}), Y^* \mid \boldsymbol{S}) &= \mathbb{E}[\boldsymbol{D}_{\hat{\boldsymbol{\gamma}}}\left(\hat{\boldsymbol{U}}'\boldsymbol{D}_{\boldsymbol{s}}^{-1}\boldsymbol{\sigma}_{\boldsymbol{xy}}\right) \mid \boldsymbol{S}] \\
&= \left(\hat{\boldsymbol{U}}'\boldsymbol{D}_{\boldsymbol{s}}^{-1}\boldsymbol{\sigma}_{\boldsymbol{xy}}\right)^{\odot 2}.
\end{aligned}$$

The final expression for $\mathrm{Cov}(\boldsymbol{\xi}(\hat{\boldsymbol{Z}}), Y^*)$ is obtained using the law of total covariance again:

$$\begin{aligned}\mathrm{Cov}(\boldsymbol{\xi}(\hat{\boldsymbol{Z}}), Y^*) &= \mathbb{E}[\mathrm{Cov}(\boldsymbol{\xi}(\hat{\boldsymbol{Z}}), Y^* \mid \boldsymbol{S})] + \mathrm{Cov}(\mathbb{E}[\boldsymbol{\xi}(\hat{\boldsymbol{Z}}) \mid \boldsymbol{S}], \mathbb{E}[Y^* \mid \boldsymbol{S}]) \\ &= \mathbb{E}[\mathrm{Cov}(\boldsymbol{\xi}(\hat{\boldsymbol{Z}}), Y^* \mid \boldsymbol{S})] \qquad \text{since } \mathbb{E}[\boldsymbol{\xi}(\hat{\boldsymbol{Z}}) \mid \boldsymbol{S}] = \mathbf{0}\end{aligned}$$

and the result is obtained.

# Chapter 6

# Conclusion

The general topic of the present thesis is the impact of dependence between explanatory variables on the performance of high-dimensional prediction and testing procedures under assumption of a standard linear regression model. Indeed, for various types of data, including functional data generated by high-throughput spectroscopy for example or more unstructured data encountered in genomic studies, a strong dependence is often observed, which raises questions about the proper way to handle it. Choosing how to take this dependence into account remains an open issue, most often addressed in comparative studies by simulations. Throughout this thesis, the effect of ignoring or not dependence is investigated and discussed, especially for the global testing and prediction issues. The former studies leading to the general conclusion that the question cannot be reduced to a two-choice issue, adaptive procedures are proposed, aiming at a flexible way to handle dependence.

*Outline*

In Chapter 1, the general context of the thesis is introduced. Existing methods and viewpoints on dependence handling are also reviewed and discussed. In particular, we point out that, although some authors advocate for handling dependence by full or partial whitening procedures, others argue that, paradoxically, completely ignoring it can improve the performance of statistical methods. Interestingly, this dichotomy appears similarly both for prediction and global testing issues. As an illustration in prediction problems, the performance of two famous

methods are compared, one ignoring dependence whereas the other is based on a preliminary complete whitening of the pointwise test statistics, using two datasets similar in terms of dependence across explanatory variables but with different patterns of association signals. It is observed that ignoring dependence improves the prediction accuracy for the first dataset, whereas better results are obtained by properly handling dependence on the other one. Therefore, for global testing issues, it is demonstrated in a very common situation that, for a given correlation structure, the preference for ignoring or not dependence depends on the pattern of the association signal.

Chapter 2 is dedicated to the impact of dependence on the detection ability of global testing procedures. Analogously as in many Genome Wide Association Studies (GWAS) in genetic epidemiology (Conneely and Boehnke, 2007), global testing is defined here as the test of the null hypothesis that the response is not related to any of the explanatory variables by the aggregation of one-to-one association test statistics between each explanatory variable and the response, under assumption of a generalized linear model. In so-called SNP-set approaches of GWAS, the same global testing method is indeed applied on segments of the genome showing different within-block dependence structures and possibly different patterns of within-block association signal. Consistently with the preliminary results displayed in Chapter 1, a comparative study based on many scenarios with different association signals first confirmed that for a given dependence structure, detection ability highly depends on the pattern of the association signal. Based on this observation, a general class of aggregation methods is proposed, covering a wide scope of dependence handling strategies. An adaptive testing method within the former class is deduced, leading to a flexible handling of dependence through a weighted decorrelation procedure. The proposed approach turns out to be always close to the most powerful methods in a comparative study using simulations, which guarantee an overall good power in a complete GWAS over all the blocks forming the genome.

Chapters 3 and 4 are devoted to the detection of interactions in GWAS. Extensions of the generalized linear model introduced in Chapter 2 are proposed to address the challenge of gene - environment interaction in Chapter 3 and gene - gene interaction in Chapter 4. In Chapter 3, a set of factors are introduced to model the marginal effect of the environment as well as the interaction effect on the response.

In this context, a proper use of global testing to test for gene - environment interaction requires a control for both marginal gene and environment effects. A parametric bootstrap approach has been introduced to correctly control the significance of the global test. A simulation study has first confirmed the efficiency of the parametric bootstrap control under the absence of gene - environment interaction effect. Similarly as in Chapter 2, the results obtained through a large comparative study showed that a flexible handling of dependence allows to make the global testing approach more robust to various dependence structures and signal patterns. The adaptation of the proposed method in Chapter 2 turns out to be always among the most powerful methods. In Chapter 4, the original model was modified by introducing gene - gene interaction components. The challenge of detecting gene - gene interaction using a global testing approach lies in correctly accounting for marginal effects and in accurately estimating pointwise statistics (*i.e.* statistic of the interaction between a pair of single elements in both sets of explanatory variables) together with their correlation matrix. Similarly as in Chapter 3, marginal effects are accounted for by using an efficient parametric bootstrap approach. Furthermore, a formal expression of the vector of pointwise statistics is derived based on the assumption of an underlying generalized linear model. The correlation turns out to show a very specific structure, which offers computationally reasonable solutions for its estimation, despite the potentially large dimension of the vector of test statistics. Results obtained through a large simulation study confirmed that adapting to the interplay between dependence structure and signal pattern is essential for a method to be powerful in every situation. Despite the good performance of our proposal to adapt to the signal/dependence pattern in many scenarios, it turns out that it lacks power in situations where the number of pointwise statistics is large. In such situations, methods like Higher Criticism show a potentially high gain in power. Furthermore, since the class of gene - gene interaction models is much more complex than the class of single SNP-set association models, global testing methods are likely to be underpowered when pointwise statistics are misspecified. To account for the different modeling of SNPs, an omnibus strategy was proposed and simulation studies demonstrate the robustness of this approach in a wide range of scenarios.

Finally, Chapter 5 addresses the handling of dependence for large or high-dimensional prediction issues. Indeed, for a two-class response variable, under the

usual assumption in Linear Discriminant Analysis (LDA) of a two-component mixture of normal distribution for the explanatory variables with equal within-class variance matrices, a naive decision rule called Diagonal Discriminant Analysis (DDA) assuming the explanatory variables are independent given the response is known to perform well in some situations (Dudoit et al., 2002; Bickel and Levina, 2004). Surprisingly, although DDA has been introduced more than fifteen years ago, its counterpart for the prediction of a real-valued response has not yet been proposed, to the best of our knowledge. A theoretical comparison of the former naive regression rule ignoring dependence and the Best Linear Unbiased Predictor based on the Ordinary Least Squares (OLS) estimation of the regression model leads to the conclusion that their prediction performance can be similar for some specific patterns of association signal. A new class of predictors is introduced, containing the naive and the OLS predictors but also Partial Least Squares (Wold et al., 1983, 1984) and Ridge (Hoerl and Kennard, 1970) predictors. This general framework offers flexibility for the handling of dependence throughout the choice of a vector of linear coefficients. The explicit expression of the optimal vector of linear coefficients is given and the corresponding predictor turns out to outperform leading prediction methods in many high-dimensional situations. Interestingly, comparative studies based on public benchmark datasets reveal that the naive and OLS predictors can show very good prediction performance and outperform more sophisticated methods especially designed for high dimension.

*Main messages and perspectives*

The first main result of this thesis is the following: how to handle dependence in testing and prediction issues does not solely depend on the pattern of dependence across explanatory variables. Indeed, we show that methods ignoring dependence can perform much better than approaches based on an explicit handling of dependence for some association signals, but also that ignoring dependence can impair the performance of statistical procedures for some other association signals. The first striking example of this fact is given at the very beginning of this manuscript, with the Leukemia and Colon datasets, sharing several important characteristics since they both are high dimensional gene expression datasets generated with similar technology. On the Leukemia dataset, the naive DDA rule performs much better than Fisher's LDA rule, whereas it is clearly the opposite on the Colon dataset (see

figure 1.1).

The second main result is that both the detection ability in global testing and prediction performance can generally be improved by an adaptive handling of dependence. Indeed, for a wide range of situations, the alternative between complete whitening and ignoring dependence is not fully satisfactory. We propose alternative approaches where the handling of dependence adapts both to the dependence structure itself and to the association signal. As a matter of fact, these approaches show more stable performance than the previous opposite approaches over different combinations for the patterns of dependence and association signal.

Several elements remain to be discussed. First, our class of global test statistics introduced in Chapter 2 is intentionally defined as linear combinations of the squared decorrelated pointwise test statistics, which includes the $L^2$-norm statistic and the Hotelling-type statistic (Derkach et al., 2014). Nevertheless, our general principles of adaptive decorrelation could also be applied on other global test statistics, such as $L^\infty$-norm based statistics or Higher Criticism statistic (Donoho and Jin, 2004). Moreover, a hyperparameter $t$ is introduced to control decorrelation on the pointwise test statistics. More advanced theoretical results are needed here to help for the search of an optimal value for this hyperparameter.

Furthermore, the performance of methods explicitly taking into account dependence, both in global testing or in prediction, is affected by the quality of the estimation of the correlation matrix of the explanatory variables. Plus, for global testing problems in large or high dimension, the proposed test might be underpowered or lack stability. This could be (at least partially) corrected by a potentially more stable rank-reduced or regularized (Hoerl and Kennard, 1970; Schäfer and Strimmer, 2005; Zuber and Strimmer, 2009; Opgen-Rhein and Strimmer, 2007b) estimate of the correlation matrix. Nevertheless, this involves new hyperparameters that can be tuned, at least for prediction issues, using cross-validation. For the detection of gene - gene interaction effects, we decided to consider eigenvectors associated to an eigenvalue greater than 1 to ensure the stability of our test. This basic rule could probably be improved and adapted to the present testing issue.

Finally, it would be interesting to provide an extension of the adaptive regression rule for classification problems in a linear discriminant analysis context. In

particular, we expose several similarities between the proposed method and the partial least squares regression method (Wold et al., 1983, 1984). Therefore, it would be interesting to compare the extension of the proposed rule to the partial least squares discriminant analysis rule (Boulesteix, 2004; Gottfries et al., 1995; Barker and Rayens, 2003), which is known to perform generally well.

# List of Works

## 1 Articles

Chapter 2 led to the submission of the following article:
F. Hébert, M. Emily and D. Causeur. An adaptive decorrelation procedure for signal detection. Submitted to *Computational Statistics and Data Analysis*.

Chapter 5 led to the submission of the following article:
F. Hébert, M. Emily and D. Causeur. Adaptive handling of dependence for regression modeling. Submitted to *Biometrika*.

## 2 Conferences

### 2.1 Peer-reviewed conferences

F. Hébert, M. Emily and D. Causeur. Adaptive handling of dependence in regression modeling. *51èmes Journées de Statistique*. Nancy, France, 3 - 7 juin 2019.

F. Hébert, M. Emily and D. Causeur. Signal detection and dependence in genome-wide association studies. *50èmes Journées de Statistique*. Saclay, France, 28 mai - 1er juin 2018.

F. Hébert, M. Emily and D. Causeur. Signal detection and dependence in genome-wide association studies. *Journée des Jeunes Chercheurs de la Société Française de Biométrie*. Paris, France, 29 mai 2018.

F. Hébert, M. Emily and D. Causeur. Signal detection and dependence in genome-wide association studies. *Jeunes Probabilistes et Statisticiens*. Saint Pierre d'Oléron, France, May 14 - 18, 2018.

F. Hébert, M. Emily and D. Causeur. Combination of dependent tests in genome-wide association studies. *49èmes Journées De Statistique*. Avignon, France, 29 mai - 2 juin 2017.

F. Hébert, M. Emily and D. Causeur. Block testing approach in genome-wide association studies using a multilevel modeling of the dependence structure. *Statistical Methods for Post Genomic Data (SMPGD) 2017*. Londres, Royaume-Uni, 12-13 janvier 2017.

### 2.2 Invited conference

F. Hébert, M. Emily and D. Causeur. Signal detection and dependence in genome-wide association studies. *7ème Journée Young Statisticians and Probabilists*. Paris, France, 25 janvier 2019.

## 3 R packages

The following R packages were developed during the thesis to implement and distribute the proposed methods, or to perform simulation studies. All of these packages can be installed from `https://github.com/fhebert/`, using the `install_github` function of the `devtools` package. The functionalities of each of these packages are detailed in appendices present at the end of this manuscript.

### 3.1 Package SNPSetSimulations

The `SNPSetSimulations` package was developed to generate SNP data with specific dependence structures. In particular, it provides a function which can be used to generate genotypic profiles with the same correlation matrix and marginal distribu-

tions as those observed on data. This package was used for the simulation study in Chapter 2.

F. Hébert, M. Emily and D. Causeur (2019). SNPSetSimulations: Simulation of genotypic profiles and binary phenotypes for GWAS. R package version 1.0.

## 3.2 Package MGFRTest

The `MGFRTest` package provides functions which can be used to compute the $p$-value of the MGF-R test introduced in Chapter 2.

F. Hébert, M. Emily and D. Causeur (2019). MGFRTest: Moment Generating Function Ratio Test. R package version 1.0.

## 3.3 Package GeneEnvInteractions

The `GeneEnvInteractions` package provides functions which can be used to generate samples under a gene-environment interaction model and compute the vector of test statistics for the detection of gene - environment interaction effects introduced in Chapter 4.

F. Hébert, M. Emily and D. Causeur (2019). GeneEnvInteractions: Global tests for the detection of gene-environment interaction effects on binary phenotypes. R package version 1.0.

## 3.4 Package GeneGeneInteractions

The `GeneGeneInteractions` package provides functions which can be used to generate samples under a gene-gene interaction model and compute the vector of test statistics for the detection of gene - gene interaction effects introduced in Chapter 4.

F. Hébert, M. Emily and D. Causeur (2019). GeneGeneInteractions: Global tests for the detection of gene-gene interaction effects on binary phenotypes. R package

version 1.0.

## 3.5 Package AdaptivePrediction

The functions of the `AdaptivePrediction` package enable to construct the adaptive regression predictor introduced in Chapter 5, and to predict the value of the response variable for new observations.

F. Hébert, M. Emily and D. Causeur (2019). AdaptivePrediction: Prediction rules based on an adaptive handling of dependence. R package version 1.0.

# Appendix A

# Simulation of SNP Data: R Package SNPSetSimulations

## 1 Introduction

To carry out the GWAS simulation studies presented in this manuscript, an R package was developed to generate correlated SNP data. This appendix provides examples of usage of the functions of this R package, named `SNPSetSimulations`. This package depends on packages `lattice` (Sarkar, 2008), `Matrix` (Bates and Maechler, 2019) and, most importantly, `GenOrd` (Barbiero and Ferrari, 2015b).

In GWAS simulation studies, SNP-sets (or genes for simplicity) are often considered. A gene can be represented as a data matrix of genotypes denoted $\boldsymbol{X}$, its generic term being denoted $x_{ij}$. $x_{ij} \in \{0, 1, 2\}$ is the number of copies of the minor allele for the $i$-th individual and $j$-th SNP of the set. Strong and heterogeneous dependence structures are often observed among adjacent SNPs. Simulation studies must replicate such dependence structures to obtain realistic and reliable results.

The `SNPSetSimulations` package aims at giving tools for generating simulated data related to case-control genome-wide association studies. In particular, it includes tools for simulating genotype and phenotype data under realistic scenarios.

# 2 Generating correlated genotype data

The `PopulationSNPSet` function provides a way to efficiently generate a matrix of genotypes for a population of `n` individuals and `p` SNPs according to given marginal distributions and within-gene dependence structure. This function can be used through several ways illustrated by the following examples. In each example, the obtained matrix is stored in the variable `G`. The following inputs are taken by the function:

- `n`: the desired size of the simulated population
- `Sigma`: the correlation matrix of the simulated SNP-set
- `maf`: a vector of minor allele frequencies (one value for each SNP); if this argument is given, genotypes are simulated according to Hardy-Weinberg equilibrium (HWE)
- `marginal`: a list of marginal distributions. This argument can be used to generate genotypes under a chosen distribution, which can be different from HWE. Each element of the list is a vector of cumulative marginal probabilities. If the $k$-th SNP takes values 0, 1 and 2 with probabilities `p0`, `p1` and `p2`, the $k$-th vector of the list is `c(p0,p0+p1)`. The third cumulative probability is supposed to be equal to one and is not given
- `X`: an observed SNP-set; this argument can be used to generate a population of `n` individuals with the same dependence structure and marginal distributions as those observed in the given SNP-set.

As a first example, we generate `n = 100000` individuals for `p = 10` autocorrelated SNPs (correlation coefficient equal to 0.8) with a minor allele frequency `p = 0.4`, under HWE:

```
G = PopulationSNPSet(n=100000,Sigma=0.8^abs(outer(1:10,1:10,"-")),
    p=rep(0.4,10))
```

As a second example, we generate SNPs with marginal distributions different from HWE. The complete marginal distribution of each SNP is thus given. Each SNP is supposed to be uniformly distributed (*i.e.* the probability for an individual to get 0, 1 or 2 copies of the minor allele equals 1/3). These marginal distributions are given in the argument `marginal`, as a list of `p` elements. Each element of the list is a vector of cumulative probabilities.

```
M = cbind(rep(1/3,10),rep(1/3+1/3,10))
M = lapply(1:nrow(M),function(i){M[i,]})
G = PopulationSNPSet(n=100000,Sigma=0.8^abs(outer(1:10,1:10,"-")),
    marginal=M)
```

Finally, we generate SNPs according to the dependence structure and marginal distributions observed on the PDZRN4 gene from the WTCCC data (Wellcome Trust Case Control Consortium, 2007).

```
data("PDZRN4")
G = PopulationSNPSet(n=100000,X=PDZRN4)
```

In the last example, the dependence structure, dimension (number of SNPs) and marginal distributions are automatically estimated on the object given in the `X` input. One can check that the simulated data has the requested dependence structure and marginal distributions by comparing them to the observed SNP-set.

The `PopulationSNPSet` function is based on modified versions of functions from the `GenOrd` package (Barbiero and Ferrari, 2015b), which gives tools for generating discrete data with given dependence structure and marginal distributions.

# 3 Generating phenotypes

Once genetic profiles for a population are generated, a disease status associated to these genotypes can be generated using a specified model. The `PopulationPhenotype`

function generates disease status conditionally to a matrix of genotypes `X` using a logistic model. Let $\boldsymbol{x_i}$ and $\boldsymbol{u_i}$ be the vector of genotypes and the optional vector of covariates of the $i$-th individual, respectively. Then the $i$-th individual is diseased with probability $\pi_i$ with

$$\pi_i = \frac{\exp(\beta_0 + \boldsymbol{u_i'}\boldsymbol{\alpha} + \boldsymbol{x_i'}\boldsymbol{\beta})}{1 + \exp(\beta_0 + \boldsymbol{u_i'}\boldsymbol{\alpha} + \boldsymbol{x_i'}\boldsymbol{\beta})}.$$

The disease status is then sampled as a Bernoulli variable with probability of success $\pi_i$. Cases are coded 1 and controls 0. Several inputs must be given:

- `X`: a matrix of genotype data, such as obtained with function `PopulationSNPSet`
- `beta0`: the intercept of the logistic model
- `beta`: the values of the non-zero coordinates of the vector $\boldsymbol{\beta}$ of the logistic model
- `I`: the vector of positions of the non-zero coordinates of $\boldsymbol{\beta}$
- `U`: an optional matrix of covariates
- `alpha`: the vector of effect parameters corresponding to the covariates
- `mod`: a vector of characters giving the association model of each involved SNP - `"A"` for additive association (default), `"R"` for recessive association and `"D"` for dominant association.

In the following example, $\beta_0 = -3$ and the 2nd and 7th coordinates of $\boldsymbol{\beta}$ are non-zero (both equal 0.2). The association between the disease and the involved SNPs corresponds to a recessive model. No covariates are used.

```
Y = PopulationPhenotype(X=G,beta0=-3,beta=c(0.2,0.2),I=c(2,7),
    mod=c("R","R"))
```

The association models can be mixed (*e.g.* a SNP can be associated to the disease through an additive model and the other one through a recessive model).

One can also desire to generate a phenotype that is independent from the SNPs. This corresponds to $\boldsymbol{\beta} = \mathbf{0}$; therefore, it can be done by setting `beta = 0` and `I = 1`.

# 4 Generating samples of genotype and phenotype data

One of the most useful functions of this package is the `SampleSNPPhenotype` function, which enables to generate samples of genotype data and corresponding disease status, *i.e.* samples similar to that used in case-control GWAS. Using a matrix of genotype data corresponding to a population (and an optional matrix of covariates), it generates the corresponding disease status using the `PopulationPhenotype` function. Then, it constructs a sample by randomly sampling a specified number of cases and controls. The result is given as a list of three elements: the first one is the genotype matrix of the sample, the second one is the vector of disease status corresponding to each row of the genotype matrix and the third one is the matrix of covariates corresponding to the sample. The function takes the following inputs:

- `X`: a matrix of genotype data
- `beta0`: the intercept of the logistic model
- `beta`: the values of the non-zero coordinates of the vector $\boldsymbol{\beta}$ of the logistic model
- `I`: the vector of positions of the non-zero coordinates of $\boldsymbol{\beta}$
- `n0`: number of controls
- `n1`: number of cases
- `U`: an optional matrix of covariates
- `alpha`: the vector of effect parameters corresponding to the covariates

- `mod`: a vector of characters giving the association model of each involved SNP - `"A"` for additive association (default), `"R"` for recessive association and `"D"` for dominant association.

In the following example, we generate a sample of 1,000 cases and 1,000 controls with the same genotype data and the same association model as in the example for the `PopulationPhenotype` function :

```
tmp = SampleSNPPhenotype(X=G,beta0=-3,beta=c(0.2,0.2),I=c(2,7),n0=1000,
      n1=1000,mod=c("R","R"))
```

The elements of the list are named `SNP`, `Phenotype` and `Covariates`. One can then compute association tests between each SNP and the phenotype. For example, using the $\chi^2$ test:

```
Z = apply(tmp$SNP,2,function(x){chisq.test(x,tmp$Phenotype,
    correct=FALSE)$p.value})
```

# Appendix B

# Implementation of the MGF-R Test: R Package MGFRTest

## 1 Introduction

An implementation of the MGF-R test introduced in this thesis is available as an R package named `MGFRTest`. This appendix provides details and examples of usage of the functions of the package.

GWAS aim at identifying genetic markers (such as Single Nucleotide Polymorphisms or SNPs) associated to a disease. Due to the great size of the genome, SNP-sets or genes, which correspond to groups of neighboring SNPs, are often considered. First, for a given gene, a test statistic for the association between each SNP and the disease is computed. Then, the individual test statistics are aggregated to construct a global test at the gene scale. This package provides functions to compute the $p$-value of the MGF-R (Moment Generating Function - Ratio) test. The $p$-value is computed using permutations of the phenotype.

In the following, examples of usage of the functions of this package are given in the GWAS context. However, it can be used more generally for testing the nullity of the mean vector of a multivariate normal distribution.

# 2 Computing the vector of association test statistics between a gene and a phenotype

Here, `X` is assumed to be a matrix of $n$ rows and $p$ columns, each row corresponding to an individual and each column to a SNP. Typically, `X` is a matrix of genotypes corresponding to a gene. It takes the values 0, 1 or 2, corresponding to the number of copies of the minor allele. `Y` is a vector giving the disease status, 0 or 1 for a control or a case, respectively. An optional covariates matrix `U` of size $n \times q$ can be given. Each row of `U` corresponds to an individual and each column to a covariate.

First, to compute the vector of test statistics between each SNP of the gene and the phenotype, the function `ScoreTest` can be used. It takes as inputs the SNP matrix `X`, the vector of disease status `Y`, the optional covariates matrix `U`, an optional matrix `Y0` containing permuted versions of `Y` and a number of permutations `N` (default to 1,000). If `Y0` is given, each column must be a permutation or resampled version of `Y`. It can be used for example to use specific resampling methods, such as parametric bootstrap. If `Y0` is not given, `N` permutations of `Y` are computed. The output is a list containing three elements named `Z`, `Z0` and `Sigma`. `Z` is the vector of test statistics, `Z0` is the matrix containing on its $i$-th row the vector of test statistics corresponding to the $i$-th permutation of `Y` and `Sigma` is the estimated correlation matrix of `Z`. The test statistics vectors and the estimated correlation matrix are computed using the formulas given in chapter 2.

As an example, we generate a matrix of genotype values for 10 SNPs and 2000 individuals (for simplicity, no dependence is introduced between the SNPs), and the vector of disease status:

```
X = matrix(sample(0:2,2000*10,TRUE),ncol=10)
Y = sample(c(rep(0,1000),rep(1,1000)))
```

The two following commands are equivalent (and give exactly the same result if the random seed is fixed):

```
res1 = ScoreTest(X=X,Y=Y,U=NULL,Y0=NULL,N=1000)
```

```
res2 = ScoreTest(X=X,Y=Y,U=NULL,Y0=sapply(1:1000,function(i){sample(Y)}))
```

For the first command, the number of permutations of the phenotype is set to 1,000. The permutations are computed automatically by the function. For the second command, the permutations are given as an input; these permutations are used to compute the $p$-value. Both `res1` and `res2` are lists containing the test statistics vector `Z` computed on the true phenotype `Y`, the matrix `Z0` of test statistics vectors computed on the permuted phenotypes (the $i$-throw of `Z0` is the test statistics vector obtained on the $i$-th permutation of the phenotype) and the estimated correlation matrix `Sigma`.

# 3 Computing the $p$-value of the MGF-R test

The $p$-value of the MGF-R test can be computed by using the `MGFR` function. It takes as inputs the vector `Z` of test statistics, the matrix `Z0` of test statistics corresponding to the permutations of `Y` and the correlation matrix `Sigma`.The following commands compute the $p$-value with 1,000 permutations:

```
res = ScoreTest(X=X,Y=Y,U=NULL,Y0=NULL,N=1000)
MGFR(res$Z,res$Z0,res$Sigma)
```

Instead of `Sigma`, the eigenvalue decomposition of `Sigma` can be given using the `eigSigma` argument. This can be useful for instance for using only the first $K$ eigenvalues.

```
ev = eigen(Z$Sigma)
MGFR(res$Z,res$Z0,eigSigma=ev)
```

# 4 Power study

The following example aims at reproducing the subfigure ($r_1$) of Figure 2.4 (see the power study in Chapter 2). For simplicity purposes, only the minP, HC, Hotelling

and $L^2$-norm tests are included in the comparison.

We first give R functions implementing each of the above mentioned tests:

```
minP = function(Z,Z0){
       stat = max(Z^2)
       stat0 = apply(Z0^2,1,max)
       p = mean(stat0>=stat)
       return(p)
}

L2norm = function(Z,Z0){
         stat = sum(Z^2)
         stat0 = (Z0^2)%*%matrix(1,ncol=1,nrow=ncol(Z0))
         p = mean(stat0>=stat)
         return(p)
}

HC = function(Z,Z0){
     m = length(Z)
     i = 1:floor(m/2)
     pZ = sort(2*(1-pnorm(abs(Z))))[i]
     stat = max(sqrt(m)*(i/m-pZ)/sqrt(pZ*(1-pZ)))
     pZ0 = apply(2*(1-pnorm(abs(Z0))),1,sort)[i,,drop=FALSE]
     stat0 = apply(sqrt(m)*(matrix(rep(i/m,ncol(pZ0)),
             ncol=ncol(pZ0))-pZ0)/sqrt(pZ0*(1-pZ0)),2,max)
     p = mean(stat0>=stat)
     return(p)
}

Hotelling = function(Z,Z0,Sigma){
            ev = eigen(Sigma)
            A = ev$vectors%*%diag(1/sqrt(ev$values))%*%t(ev$vectors)
            stat = sum((Z%*%A)^2)
```

```
            stat0 = rowSums((Z0%*%A)^2)
            p = mean(stat<=stat0)
            return(p)
}
```

In the following, the `SNPSetSimulations` package also developed during this thesis is used to generate genotype and phenotype data used in simulation studies. In particular, the `PopulationSNPSet` function if used to generate a population of genetic profiles reproducing the dependence structure observed in a gene (in the following, the PDZRN4 gene), and the `SampleSNPPhenotype` function is used to generated a sample of genetic profiles and associated phenotype values. The tests are then applied to these samples.

The following commands generate the population of genetic profiles:

```
library(SNPSetSimulations)
data("PDZRN4")
X = PopulationSNPSet(n=100000,X=PDZRN4)
```

The power of each test is now estimated as a function of the signal strength. For each value of the signal strength, 1,000 samples are generated to estimate the power:

```
res = matrix(0,ncol=5,nrow=11)
for(k in 0:10){
    res.tmp = sapply(1:1000,function(i){
              tmp = SampleSNPPhenotype(X,beta0=-3,beta=c(0.4,0.4,0.8)*k/10,
                    I=c(18,25,33),n0=1000,n1=1000)
              tmp = ScoreTest(tmp$SNP,tmp$Phenotype)
              tmp = c(minP(tmp$Z,tmp$Z0),
                    L2norm(tmp$Z,tmp$Z0), HC(tmp$Z,tmp$Z0),
                    Hotelling(tmp$Z,tmp$Z0,tmp$Sigma),
                    MGFR(tmp$Z,tmp$Z0,tmp$Sigma))
              return(tmp)
```

```
    })
    res.tmp = rowMeans(res.tmp<0.05)
    res[k+1,] = res.tmp
}
```

The results can finally be displayed:

```
matplot(seq(0,1,by=0.1),res,type="o",lty=1,lwd=2,ylim=c(0,1),
        col=c(rep("black",4),"red"),ylab="Power",cex=1.5,
        xlab="Signal strength",pch=c(NA,15,16,17,NA))
legend("topleft",bty="n",lwd=2,pch=c(NA,15,16,17,NA),cex=1.5,
       col=c(rep("black",4),"red"),
       legend=c("minP","L2-norm","HC","Hotelling","MGF-R"))
```

# Appendix C

# Gene - Environment Interaction Effects Detection: R Package GeneEnvInteractions

## 1 Introduction

To compute the simulation study of Chapter 3, an R package named `GenEnvInteractions` was developed. It provides functions to generate samples under a gene - environment interaction model and to compute the vector of test statistics on which global testing methods can be applied to test the presence of a gene - environment interaction effect.

## 2 Generating samples under a gene - environment interaction model

The `SampleGEInter` function can be used to generate a sample under a gene - environment interaction model for a binary phenotype. Given a matrix `X` of size $N \times p$ corresponding to a population of genotypic profiles, a vector `E` of length $N$

containing the values of the environmental factor corresponding to the rows of `X` and an optional matrix `U` of size $N \times q$ corresponding to additional covariates, the function generates a binary phenotype $Y$ under the following model:

$$\text{logit}(\mathbb{P}[Y = 1|\boldsymbol{X} = \boldsymbol{x}, E = e, \boldsymbol{U} = \boldsymbol{u}]) = \beta_0 + \boldsymbol{x}'\boldsymbol{\beta_X} + e\beta_E + \boldsymbol{s}'\boldsymbol{\gamma} + \boldsymbol{u}'\boldsymbol{\beta_U}$$

where $\boldsymbol{s} = e \cdot \boldsymbol{x}$ is the interaction profile. The function takes the following inputs:

- `X`: a $N \times p$ matrix of genotypes (coded as $0/1/2$) for a given gene for a population
- `E`: a vector of length $N$ containing the values of the environmental variable for the same population
- `mu`: the intercept of the logistic model
- `betaX`: the vector of non-zero parameters for the main genetic effects
- `I.betaX`: the vector of positions corresponding to the coefficients in `betaX`
- `betaE`: the parameter for the main effect corresponding to the environmental variable
- `gamma`: the vector of non-zero parameters for the interaction effects
- `I.gamma`: the vector of positions corresponding to the coefficients in `gamma`
- `n0`: the desired number of controls
- `n1`: the desired number of cases
- `U`: an optional $N \times q$ matrix of covariates
- `betaU`: the vector of parameters associated to the covariates.

The output is a list containing the following elements:

- `SNP`: the sample matrix of genotypic profiles
- `Env`: the sample vector containing the values of the environmental variable
- `Phenotype`: the sample vector of disease status
- `Covariates`: the sample matrix of covariates.

# 3 Computing the test statistics vector corresponding to a gene - environment interaction model

The `GEInterScoreTest` function provides an implementation of the test statistics vector $\mathcal{Z}$ introduced in Chapter 3, on which global testing methods can be applied. The function requires the following inputs:

- `X`: a matrix of size $n \times p$ corresponding to a sample of genotypic profiles (coded as $0/1/2$)
- `E`: a vector of length $n$ containing the values of the environmental variable
- `Y`: a vector of length $n$ containing the values of the binary phenotype $(0/1)$
- `perm.method`: a character giving the chosen resampling method for the phenotype: paramatric bootstrap (`perm.method="parametric"`, default value) or simple permutations (any other character)
- `N`: the number of permutations of the phenotype (default = 1000)
- `U`: an optional matrix of additional covariates.

As a result, the function gives a list containing several elements, namely the test statistics vector `Z`, the matrix `Z0` of permuted vectors of test statistics (displayed as rows) and the estimated correlation matrix `Sigma`.

# Appendix D

# Gene - Gene Interaction Effects Detection: R Package GeneGeneInteractions

## 1 Introduction

To compute the simulation study of Chapter 4, an R package named `GeneGeneInteractions` was developed. It provides functions to generate samples under a gene - gene interaction model and to compute the vector of test statistics on which global testing methods can be applied to test the presence of a gene - gene interaction effect.

## 2 Generating samples under a gene - gene interaction model

The `SampleGGInter` function can be used to generate a sample under a gene - gene interaction model for a binary phenotype. Given a matrix `X1` of size $N \times p_1$ corresponding to a population of genotypic profiles for a given gene, a matrix `X2` of size $N \times p_2$ for another gene and an optional matrix `U` of size $N \times q$ corresponding to

additional covariates, the function generates a binary phenotype $Y$ under a logistic model. The user can choose to use a continuous or a dummy coding for the SNPs for the interaction effect (the continuous coding is used in both cases for the marginal genetic effects). Therefore, the simulation model for the phenotype is the following:

$$\text{logit}(\mathbb{P}[Y = 1|\boldsymbol{X^{(1)}} = \boldsymbol{x^{(1)}}, \boldsymbol{X^{(2)}} = \boldsymbol{x^{(2)}}]) = \beta_0 + \boldsymbol{x^{(1)\prime}\beta_{(1)}} + \boldsymbol{x^{(2)\prime}\beta_{(2)}} + \boldsymbol{s'\gamma} + \boldsymbol{u'\beta_U}$$

where $\boldsymbol{s} = \boldsymbol{x^{(1)}} \otimes \boldsymbol{x^{(2)}}$ or $\boldsymbol{s} = (\boldsymbol{a^{(1)\prime}}, \boldsymbol{b^{(1)\prime}})' \otimes (\boldsymbol{a^{(2)\prime}}, \boldsymbol{b^{(2)\prime}})'$ where $\boldsymbol{a^{(\ell)}} = (a_1^{(\ell)}, \ldots, a_{p_\ell}^{(\ell)})'$ with $a_i^{(\ell)} = \mathbf{1}_{\{X_i^{(\ell)}=1\}}$, and $\boldsymbol{b^{(\ell)}} = (b_1^{(\ell)}, \ldots, b_{p_\ell}^{(\ell)})'$ with $b_i^{(\ell)} = \mathbf{1}_{\{X_i^{(\ell)}=2\}}$.

The function takes the following inputs:

- `X1`: a $N \times p_1$ matrix of genotypes (coded as $0/1/2$) for a given gene for a population
- `X2`: a $N \times p_2$ matrix of genotypes (coded as $0/1/2$) for a given gene for a population
- `beta0`: the intercept of the logistic model
- `beta1`: the vector of non-zero parameters for the main genetic effects of the first gene
- `I.beta1`: the vector of positions corresponding to the coefficients in `beta1`
- `beta2`: the vector of non-zero parameters for the main genetic effects of the second gene
- `I.beta2`: the vector of positions corresponding to the coefficients in `beta2`
- `gamma`: a $K \times 4$ matrix or a vector of length $K$ of gene-gene interaction effects parameters. If it is a $K \times 4$ matrix, each row is a vector of parameters for a pair of SNPs having an interaction effect and a dummy coding of the SNPs is used. The four parameters in a row correspond to the heterozygous/heterozygous, heterozygous/homozygous, homozygous/heterozygous and homozygous/homozygous configurations, respectively. If it is a vector, each coordinate is the interaction parameter for a pair of SNPs having an interaction effect. In this case, a continuous coding is used and the interaction effect is multiplicative

- `I.gamma`: a $K \times 2$ matrix of coordinates for the interacting SNPs. Each row is a pair of SNPs. For a given row $(i, j)$, the $i$-th SNP of the first gene and the $j$-th SNP of the second gene have an interaction effect on the phenotype. The interaction parameters for the pair of SNPs in the $k$-th row of `I.gamma` are given by the $k$-th row (or coordinate) of `gamma`
- `n0`: the desired number of controls
- `n1`: the desired number of cases
- `U`: an optional $N \times q$ matrix of covariates
- `betaU`: the vector of parameters associated to the covariates.

The output is a list containing the following elements:

- `SNP1`: the sample matrix of genotypic profiles for the first gene
- `SNP2`: the sample matrix of genotypic profiles for the second gene
- `Phenotype`: the sample vector of disease status
- `Covariates`: the sample matrix of covariates.

# 3 Computing the test statistics vector corresponding to a gene - gene interaction model

The `GGInterScoreTest` function provides an implementation of the test statistics vector $\mathcal{Z}$ introduced in chapter 4, on which global testing methods can be applied. The function requires the following inputs:

- `X1`: a matrix of size $n \times p_1$ corresponding to a sample of genotypic profiles (coded as 0/1/2) for the first gene
- `X2`: a matrix of size $n \times p_2$ corresponding to a sample of genotypic profiles (coded as 0/1/2) for the second gene

- `Y`: a vector of length $n$ containing the values of the binary phenotype (0/1)
- `perm.method`: a character giving the chosen resampling method for the phenotype: paramatric bootstrap (`perm.method="parametric"`, default value) or simple permutations (any other character)
- `N`: the number of permutations of the phenotype (default = 1000)
- `U`: an optional matrix of additional covariates.

As a result, the functions gives a list containing several elements, namely the test statistics vector `Z`, the matrix `Z0` of permuted vectors of test statistics (displayed as rows), the estimated correlation matrix `Sigma` (computed using the direct formula) and the estimated eigendecomposition `ev` of the correlation matrix, using the Kronecker decomposition.

# Appendix E

# Implementation of the Adaptive Regression Predictor: R Package AdaptivePrediction

An implementation of the regression rule introduced in Chapter 5 is available as an R package named `AdaptivePrediction`. Usage of the functions of this package, namely for constructing the prediction rule and for prediction the value of the response for a new observation, is detailed in this appendix. The soil near infrared spectra dataset used in chapter 5 will be used as an example for the usage of the functions.

The dataset is first loaded and the individuals for which the response is unobserved are removed:

```
library(prospectr)
data(NIRsoil)
indNA = which(is.na(NIRsoil$Nt))
NIRNt = list(x=NIRsoil$spc[-indNA,],y=NIRsoil$Nt[-indNA])
```

The regression rule will be constructed on a training sample; it will then be used to predict the values of the response for a validation sample. The validation sample contains 50 individuals selected at random. In the following, the positions in

the whole sample of the individuals of the validation sample are stored in a vector named `ind.test`.

The `AdaptiveReg` function is first used to construct the regression rule. It takes as necessary inputs the matrix `x` of explanatory variables and the vector `y` containing the corresponding values of the response. Two optional inputs can be given: a scalar `nvmax` giving a maximal numer of latent factors to be used in the construction of the regression rule and a scalar `nfolds` giving the number of folds to use when performing the cross-validation for the determination of the optimal number of latent factors. By default, the function uses as many factors as possible to determine the optimal number, and 10-fold cross-validation is performed. The regression rule is obtained by the following function:

```
fit = AdaptiveReg(NIRNt$x[-ind.test,],NIRNt$y[-ind.test])
```

The obtained `fit` object is a list containing a list named `ZgMoments`, a scalar `nv` giving the optimal number of factors and a vector `R2` giving the value of the cross-validation $R^2$ for each possible number of factors. `ZgMoments` contains several objects related to the moments of the latent factors, as detailed in the development of the method in Chapter 5. A plot representing the $R^2$ as a function of the number of factors can be obtained by using the `PlotAdaptiveReg` function:

```
PlotAdaptiveReg(fit)
```

The plot also gives the optimal number of factors.

Finally, once the regression rule is constructed, the values of the response for new observations can be obtained with the `PredictAdaptiveReg` function. This function takes as inputs a fitted object obtained by the `AdaptiveReg` function, and a matrix `xnew` containing the values of the explanatory variables for new observations. The predicted values for the validation sample can be obtained and compared to the true response by the following commands:

```
pred = PredictAdaptiveReg(fit,NIRNt$x[ind.test,])
plot(NIRNt$y[ind.test],pred,pch=16,xlab="Y",ylab=expression(hat(Y)))
```

# Bibliography


Ahdesmäki, M. and Strimmer, K. (2010). Feature selection in omics prediction problems using cat scores and false nondiscovery rate control. *The Annals of Applied Statistics*, 4(1):503–519.

Ahmad, M. H., Adnan, R., and Adnan, N. (2006). A comparative study on some methods for handling multicollinearity problems. *MATEMATIKA: Malaysian Journal of Industrial and Applied Mathematics*, 22:109–119.

Anderson, T. (2003). *An Introduction to Multivariate Statistical Analysis*. Wiley Series in Probability and Statistics. Wiley.

Arias-Castro, E., Candès, E. J., and Plan, Y. (2011). Global testing under sparse alternatives: Anova, multiple comparisons and the higher criticism. *The Annals of Statistics*, pages 2533–2556.

Barbiero, A. and Ferrari, P. A. (2015a). *GenOrd: Simulation of Discrete Random Variables with Given Correlation Matrix and Marginal Distributions*. R package version 1.4.0.

Barbiero, A. and Ferrari, P. A. (2015b). *GenOrd: Simulation of Discrete Random Variables with Given Correlation Matrix and Marginal Distributions*. R package version 1.4.0.

Barker, M. and Rayens, W. (2003). Partial least squares for discrimination. *Journal of Chemometrics: A Journal of the Chemometrics Society*, 17(3):166–173.

Barnett, I., Mukherjee, R., and Lin, X. (2017). The generalized higher criticism for testing snp-set effects in genetic association studies. *Journal of the American Statistical Association*, 112(517):64–76.

Basu, S. and Pan, W. (2011). Comparison of statistical tests for disease association with rare variants. *Genetic epidemiology*, 35(7):606–619.

Bates, D. and Maechler, M. (2018). *Matrix: Sparse and Dense Matrix Classes and Methods*. R package version 1.2-14.

Bates, D. and Maechler, M. (2019). *Matrix: Sparse and Dense Matrix Classes and Methods*. R package version 1.2-17.

Behzadi, P., Behzadi, E., and Ranjbar, R. (2015). The incidence and prevalence of crohn's disease in global scale. *SOJ Immunol*, 3(2):1–6.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the royal statistical society. Series B (Methodological)*, pages 289–300.

Bickel, P. J. and Levina, E. (2004). 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.

Bůžková, P., Lumley, T., and Rice, K. (2011). Permutation and parametric bootstrap tests for gene–gene and gene–environment interactions. *Annals of human genetics*, 75(1):36–45.

Blazère, M., Gamboa, F., and Loubes, J.-M. (2014). A unified framework to study the properties of the pls vector of regression coefficients. In *International Conference on Partial Least Squares and Related Methods*, pages 227–237. Springer.

Bodnar, T. and Okhrin, Y. (2011). On the product of inverse wishart and normal distributions with applications to discriminant analysis and portfolio theory. *Scandinavian Journal of Statistics*, 38(2):311–331.

Boonstra, P. S., Mukherjee, B., Gruber, S. B., Ahn, J., Schmit, S. L., and Chatterjee, N. (2016). Tests for gene-environment interactions and joint effects with exposure misclassification. *American journal of epidemiology*, 183(3):237–247.

Boulesteix, A.-L. (2004). Pls dimension reduction for classification with microarray data. *Statistical applications in genetics and molecular biology*, 3(1):1–30.

Boulesteix, A.-L., Durif, G., Lambert-Lacroix, S., Peyre, J., and Strimmer., K. (2018). *plsgenomics: PLS Analyses for Genomics*. R package version 1.5-2.

Broeckx, B. J., Derrien, T., Mottier, S., Wucher, V., Cadieu, E., Hédan, B., Le Béguec, C., Botherel, N., Lindblad-Toh, K., Saunders, J. H., et al. (2017). An exome sequencing based approach for genome-wide association studies in the dog. *Scientific reports*, 7(15680).

Cai, T., Liu, W., and Xia, Y. (2014). Two-sample test of high dimensional means under dependence. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(2):349–372.

Chong, I.-G. and Jun, C.-H. (2005). Performance of some variable selection methods when multicollinearity is present. *Chemometrics and intelligent laboratory systems*, 78(1-2):103–112.

Christensen, R. (2015). Covariance of the Wishart distribution with applications to regression.

Clavel, J. (2007). Progress in the epidemiological understanding of gene–environment interactions in major diseases: cancer. *Comptes rendus biologies*, 330(4):306–317.

Conneely, K. N. and Boehnke, M. (2007). So many correlated tests, so little time! rapid adjustment of p values for multiple correlated tests. *The American Journal of Human Genetics*, 81(6):1158–1168.

Coombes, B. J. and Biernacka, J. M. (2018). Application of the parametric bootstrap for gene-set analysis of gene–environment interactions. *European Journal of Human Genetics*, 26(11):1679.

Cordell, H. J. (2002). Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. *Human molecular genetics*, 11(20):2463–2468.

Cordell, H. J. (2009a). Detecting gene–gene interactions that underlie human diseases. *Nature Reviews Genetics*, 10(6):392.

Cordell, H. J. (2009b). Estimation and testing of gene–environment interactions in family-based association studies. *Genomics*, 93(1):5–9.

Cordell, H. J., Todd, J. A., Hill, N. J., Lord, C. J., Lyons, P. A., Peterson, L. B., Wicker, L. S., and Clayton, D. G. (2001). Statistical modeling of interlocus interactions in a complex disease: rejection of the multiplicative model of epistasis in type 1 diabetes. *Genetics*, 158(1):357–367.

Cuevas, A., Febrero, M., and Fraiman, R. (2004). An anova test for functional data. *Computational statistics & data analysis*, 47(1):111–122.

De, R., Verma, S. S., Drenos, F., Holzinger, E. R., Holmes, M. V., Hall, M. A., Crosslin, D. R., Carrell, D. S., Hakonarson, H., Jarvik, G., et al. (2015). Identifying gene-gene interactions that are highly associated with body mass index using quantitative multifactor dimensionality reduction (qmdr). *BioData mining*, 8(1):41.

De Jong, S. (1993). PLS fits closer than PCR. *Journal of chemometrics*, 7(6):551–557.

de Lange, K. M., Moutsianas, L., Lee, J. C., Lamb, C. A., Luo, Y., Kennedy, N. A., Jostins, L., Rice, D. L., Gutierrez-Achury, J., Ji, S.-G., et al. (2017). Genome-wide association study implicates immune activation of multiple integrin genes in inflammatory bowel disease. *Nature genetics*, 49(2):256.

Dehman, A., Ambroise, C., and Neuvial, P. (2015). Performance of a blockwise approach in variable selection using linkage disequilibrium information. *BMC bioinformatics*, 16(1):148.

Derkach, A., Lawless, J. F., and Sun, L. (2013). Robust and powerful tests for rare variants using fisher's method to combine evidence of association from two or more complementary tests. *Genetic epidemiology*, 37(1):110–121.

Derkach, A., Lawless, J. F., and Sun, L. (2014). Pooled association tests for rare genetic variants: a review and some new results. *Statistical Science*, pages 302–321.

Dong, C., Chu, X., Wang, Y., Wang, Y., Jin, L., Shi, T., Huang, W., and Li, Y. (2008). Exploration of gene–gene interaction effects using entropy-based methods. *European Journal of Human Genetics*, 16(2):229.

Donoho, D. and Jin, J. (2004). Higher criticism for detecting sparse heterogeneous mixtures. *Annals of Statistics*, pages 962–994.

Donoho, D. and Jin, J. (2008). Higher criticism thresholding: Optimal feature selection when useful features are rare and weak. *Proceedings of the National Academy of Sciences*, 105(39):14790–14795.

Dudbridge, F. and Fletcher, O. (2014). Gene-environment dependence creates spurious gene-environment interaction. *The American Journal of Human Genetics*, 95(3):301–307.

Dudoit, S., Fridlyand, J., and Speed, T. P. (2002). Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American Statistical Association*, 97(457):77–87.

Eastment, H. and Krzanowski, W. (1982). Cross-validatory choice of the number of components from a principal component analysis. *Technometrics*, 24(1):73–77.

Efron, B. (2009). Empirical bayes estimates for large-scale prediction problems. *Journal of the American Statistical Association*, 104(487):1015–1028.

Emily, M. (2012). IndOR: a new statistical procedure to test for snp–snp epistasis in genome-wide association studies. *Statistics in medicine*, 31(21):2359–2373.

Emily, M. (2016). Aggregator: A gene-based gene-gene interacttion test for case-control association studies. *Statistical Application in Genetics and Molecular Biology*, 15(2):151–171.

Emily, M. (2018). A survey of statistical methods for gene-gene interaction in case-control genome-wide association studies. *Journal de la Société Française de Statistique*, 159(1):27–67.

Epstein, M. P., Duncan, R., Jiang, Y., Conneely, K. N., Allen, A. S., and Satten, G. A. (2012). A permutation procedure to correct for confounders in case-control studies, including tests of rare variation. *The American Journal of Human Genetics*, 91(2):215–223.

Fisher, R. A. (1925). *Statistical methods for research workers*. Oliver and Boyd (Edinburgh).

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

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

Friedman, J. H. (1989). Regularized discriminant analysis. *Journal of the American statistical association*, 84(405):165–175.

Friguet, C., Kloareg, M., and Causeur, D. (2009). A factor model approach to multiple testing under dependence. *Journal of the American Statistical Association*, 104(488):1406–1415.

Gottfries, J., Blennow, K., Wallin, A., and Gottfries, C. (1995). Diagnosis of dementias using partial least squares discriminant analysis. *Dementia and Geriatric Cognitive Disorders*, 6(2):83–88.

Guo, Y., Hastie, T., and Tibshirani, R. (2005). Regularized discriminant analysis and its application in microarrays. *Biostatistics*, 1(1):1–18.

Guo, Y., Wu, C., Guo, M., Liu, X., and Keinan, A. (2018). Gene-based nonparametric testing of interactions using distance correlation coefficient in case-control association studies. *Genes*, 9(12).

Hager, W. W. (1989). Updating the inverse of a matrix. *SIAM review*, 31(2):221–239.

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

Hall, P. and Jin, J. (2010). Innovated higher criticism for detecting sparse signals in correlated noise. *The Annals of Statistics*, 38(3):1686–1732.

Han, F. and Pan, W. (2010). A data-adaptive sum test for disease association with multiple common or rare variants. *Human Heredity*, 70(1):42–54.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference and Prediction, Second Edition*. Springer Series in Statistics New York.

Helland, I. S. (1988). On the structure of partial least squares regression. *Communications in statistics-Simulation and Computation*, 17(2):581–607.

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

Horn, R. A. and Johnson, C. R. (1994). *Topics in matrix analysis*. Cambridge university press Cambridge.

Hotelling, H. (1931). The generalization of student's ratio. *The Annals of Mathematical Statistics*, 2(3):360–378.

Hu, J. K., Wang, X., and Wang, P. (2014). Testing gene–gene interactions in genome wide association studies. *Genetic epidemiology*, 38(2):123–134.

Hunter, D. J. (2005). Gene–environment interactions in human diseases. *Nature Reviews Genetics*, 6(4):287.

Ingster, Y. I. (1997). Some problems of hypothesis testing leading to infinitely divisible distributions. *Mathematical Methods of Statistics*, 6(1):47–69.

Jolliffe, I. (2011). *Principal component analysis*. Springer.

Jolliffe, I. T. (1982). A note on the use of principal components in regression. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 31(3):300–303.

Kaiser, H. F. (1960). The application of electronic computers to factor analysis. *Educational and psychological measurement*, 20(1):141–151.

Kessy, A., Lewin, A., and Strimmer, K. (2018). Optimal whitening and decorrelation. *The American Statistician*, pages 1–6.

Klengel, T. and Binder, E. B. (2013). Gene× environment interactions in the prediction of response to antidepressant treatment. *International Journal of Neuropsychopharmacology*, 16(3):701–711.

Kollo, T. and Neudecker, H. (1993). Asymptotics of eigenvalues and unit-length eigenvectors of sample variance and correlation matrices. *Journal of Multivariate Analysis*, 47(2):283–300.

Krstajic, D., Buturovic, L. J., Leahy, D. E., and Thomas, S. (2014). Cross-validation pitfalls when selecting and assessing regression and classification models. *Journal of cheminformatics*, 6(1):10.

Kwee, L. C., Liu, D., Lin, X., Ghosh, D., and Epstein, M. P. (2008). A powerful and flexible multilocus association test for quantitative traits. *The American Journal of Human Genetics*, 82(2):386–397.

Larson, N. B., Jenkins, G. D., Larson, M. C., Vierkant, R. A., Sellers, T. A., Phelan, C. M., Schildkraut, J. M., Sutphen, R., Pharoah, P. P., Gayther, S. A., et al. (2014). Kernel canonical correlation analysis for assessing gene–gene interactions and application to ovarian cancer. *European Journal of Human Genetics*, 22(1):126.

Lesch, K. P. (2004). Gene–environment interaction and the genetics of depression. *Journal of Psychiatry and Neuroscience*, 29(3):174.

Li, J., Huang, D., Guo, M., Liu, X., Wang, C., Teng, Z., Zhang, R., Jiang, Y., Lv, H., and Wang, L. (2015). A gene-based information gain method for detecting gene–gene interactions in case–control studies. *European Journal of Human Genetics*, 23(11):1566.

Li, J., Tang, R., Biernacka, J. M., and De Andrade, M. (2009a). Identification of gene-gene interaction using principal components. *BMC proceedings*, 3(Suppl 7):S78.

Li, Q., Zheng, G., Liang, X., and Yu, K. (2009b). Robust tests for single-marker analysis in case-control genetic association studies. *Annals of human genetics*, 73(2):245–252.

Li, W., Goovaerts, P., and Meurens, M. (1996). Quantitative analysis of individual sugars and acids in orange juices by near-infrared spectroscopy of dry extract. *Journal of agricultural and food chemistry*, 44(8):2252–2259.

Li, W. and Reich, J. (2000). A complete enumeration and classification of two-locus disease models. *Human heredity*, 50(6):334–349.

Lin, X., Lee, S., Christiani, D. C., and Lin, X. (2013). Test for interactions between a genetic marker set and environment in generalized linear models. *Biostatistics*, 14(4):667–681.

Lin, X., Lee, S., Wu, M. C., Wang, C., Chen, H., Li, Z., and Lin, X. (2016). Test for rare variants by environment interactions in sequencing association studies. *Biometrics*, 72(1):156–164.

Lingjaerde, O. C. and Christophersen, N. (2000). Shrinkage structure of partial least squares. *Scandinavian Journal of Statistics*, 27(3):459–473.

Liu, J. Z., Mcrae, A. F., Nyholt, D. R., Medland, S. E., Wray, N. R., Brown, K. M., Hayward, N. K., Montgomery, G. W., Visscher, P. M., Martin, N. G., et al. (2010). A versatile gene-based test for genome-wide association studies. *The American Journal of Human Genetics*, 87(1):139–145.

Liu, Y. and Xie, J. (2018). Powerful test based on conditional effects for genome-wide screening. *The Annals of Applied Statistics*, 12(1):567.

Loftus Jr, E. V. (2004). Clinical epidemiology of inflammatory bowel disease: incidence, prevalence, and environmental influences. *Gastroenterology*, 126(6):1504–1517.

Lu, T., Pan, Y., Kao, S.-Y., Li, C., Kohane, I., Chan, J., and Yankner, B. A. (2004). Gene regulation and dna damage in the ageing human brain. *Nature*, 429(6994):883.

Luo, L., Peng, G., Zhu, Y., Dong, H., Amos, C. I., and Xiong, M. (2010). Genome-wide gene and pathway analysis. *European Journal of Human Genetics*, 18(9):1045.

Ma, L., Clark, A. G., and Keinan, A. (2013). Gene-based testing of interactions in association studies of quantitative traits. *PLoS Genetics*, 9(2).

Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., McCarthy, M. I., Ramos, E. M., Cardon, L. R., Chakravarti, A., et al. (2009). Finding the missing heritability of complex diseases. *Nature*, 461(7265):747.

Mevik, B.-H., Wehrens, R., and Liland, K. H. (2019). *pls: Partial Least Squares and Principal Component Regression*. R package version 2.7-1.

Minasny, B. and McBratney, A. B. (2008). Regression rules as a tool for predicting soil properties from infrared reflectance spectroscopy. *Chemometrics and intelligent laboratory systems*, 94(1):72–79.

Moskvina, V. and Schmidt, K. M. (2008). On multiple-testing correction in genome-wide association studies. *Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society*, 32(6):567–573.

Musameh, M. D., Wang, W. Y., Nelson, C. P., Lluís-Ganella, C., Debiec, R., Subirana, I., Elosua, R., Balmforth, A. J., Ball, S. G., Hall, A. S., et al. (2015). Analysis of gene-gene interactions among common variants in candidate cardiovascular genes in coronary artery disease. *PloS one*, 10(2):e0117684.

Nakamoto, K., Wang, S., Jenison, R. D., Guo, G. L., Klaassen, C. D., Wan, Y.-J. Y., and Zhong, X.-b. (2006). Linkage disequilibrium blocks, haplotype structure, and htsnps of human cyp7a1 gene. *BMC genetics*, 7(1):29.

Neale, B. and Sham, P. (2004). The future of association studies: gene-based analysis and replication. *American Journal of Human Genetics*, 75:353–362.

Opgen-Rhein, R. and Strimmer, K. (2007a). Accurate ranking of differentially expressed genes by a distribution-free shrinkage approach. *Statistical applications in genetics and molecular biology*, 6(1).

Opgen-Rhein, R. and Strimmer, K. (2007b). From correlation to causation networks: a simple approximate learning algorithm and its application to high-dimensional plant gene expression data. *BMC systems biology*, 1(1):37.

Pan, W. (2009). Asymptotic tests of association with multiple snps in linkage disequilibrium. *Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society*, 33(6):497–507.

Pan, W., Kim, J., Zhang, Y., Shen, X., and Wei, P. (2014). A powerful and adaptive association test for rare variants. *Genetics*, 197(4):1081–1095.

Pattaro, C., Ruczinski, I., Fallin, D. M., and Parmigiani, G. (2008). Haplotype block partitioning as a tool for dimensionality reduction in SNP association studies. *BMC genomics*, 9(1):405.

Peng, Q., Zhao, J., and Xue, F. (2010). A gene-based method for detecting gene–gene co-association in a case–control association study. *European Journal of Human Genetics*, 18(5):582.

Perthame, E., Friguet, C., and Causeur, D. (2016a). Stability of feature selection in classification issues for high-dimensional correlated data. *Statistics and Computing*, 26(4):783–796.

Perthame, E., Friguet, C., and Causeur, D. (2016b). Stability of feature selection in classification issues for high-dimensional correlated data. *Statistics and Computing*, 26(4):783–796.

Pierna, J. A. F. and Dardenne, P. (2008). Soil parameter quantification by nirs as a chemometric challenge at 'chimiométrie 2006'. *Chemometrics and intelligent laboratory systems*, 91(1):94–98.

Preacher, K. J. and MacCallum, R. C. (2003). Repairing tom swift's electric factor analysis machine. *Understanding statistics: Statistical issues in psychology, education, and the social sciences*, 2(1):13–43.

R Development Core Team (2008). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Rajapakse, I., Perlman, M. D., Martin, P. J., Hansen, J. A., and Kooperberg, C. (2012). Multivariate detection of gene-gene interactions. *Genetic epidemiology*, 36(6):622–630.

Ramsay, J., Hooker, G., and Graves, S. (2009). *Functional Data Analysis with R and MATLAB*. Use R! Springer New York.

Ramsay, J. and Silverman, B. (2005). *Functional Data Analysis*. Springer Series in Statistics. Springer.

Reddon, H., Gueant, J.-L., and Meyre, D. (2016). The importance of gene-environment interactions in human obesity. *Clinical science*, 130(18):1571–1597.

Sarkar, D. (2008). *Lattice: Multivariate Data Visualization with R*. Springer, New York. ISBN 978-0-387-75968-5.

Sasieni, P. D. (1997). From genotypes to genes: doubling the sample size. *Biometrics*, 53(4):1253–1261.

Schäfer, J. and Strimmer, K. (2005). A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical applications in genetics and molecular biology*, 4(1).

Shen, Q. and Faraway, J. (2004). An F test for linear models with functional responses. *Statistica Sinica*, pages 1239–1257.

Sheu, C.-F., Perthame, É., Lee, Y.-S., Causeur, D., et al. (2016). Accounting for time dependence in large-scale multiple testing of event-related potential data. *The Annals of Applied Statistics*, 10(1):219–245.

Simes, R. J. (1986). An improved bonferroni procedure for multiple tests of significance. *Biometrika*, 73(3):751–754.

Stevens, A. and Ramirez-Lopez, L. (2013). *An introduction to the prospectr package*. R package version 0.1.3.

Su, Y.-R., Di, C.-Z., and Hsu, L. (2017). A unified powerful set-based test for sequencing data analysis of gxe interactions. *Biostatistics*, 18(1):119–131.

Thomas, D. (2010). Gene–environment-wide association studies: emerging approaches. *Nature Reviews Genetics*, 11(4):259.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288.

Tibshirani, R., Hastie, T., Narasimhan, B., Chu, G., et al. (2003). Class prediction by nearest shrunken centroids, with applications to dna microarrays. *Statistical Science*, 18(1):104–117.

Tukey, J. W. (1976). T13 n: The higher criticism. *Course notes, Stat*, 411.

Twells, R. C., Mein, C. A., Phillips, M. S., Hess, J. F., Veijola, R., Gilbey, M., Bright, M., Metzker, M., Lie, B. A., Kingsnorth, A., et al. (2003). Haplotype

structure, ld blocks, and uneven recombination within the lrp5 gene. *Genome research*, 13(5):845–855.

Ueki, M. (2014). On the choice of degrees of freedom for testing gene–gene interactions. *Statistics in medicine*, 33(28):4934–4948.

Ueki, M. and Cordell, H. J. (2012). Improved statistics for genome-wide interaction analysis. *PLoS Genetics*, 8(4):e1002625.

VanderWeele, T. J. and Laird, N. M. (2011). Tests for compositional epistasis under single interaction-parameter models. *Annals of human genetics*, 75(1):146–156.

Vukcevic, D., Hechter, E., Spencer, C., and Donnelly, P. (2011). Disease model distortion in association studies. *Genetic epidemiology*, 35(4):278–290.

Wald, A. (1943). Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Transactions of the American Mathematical society*, 54(3):426–482.

Wall, J. D. and Pritchard, J. K. (2003). Haplotype blocks and linkage disequilibrium in the human genome. *Nature Reviews Genetics*, 4(8):587.

Wellcome Trust Case Control Consortium (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447(7145):661.

Wellek, S. and Ziegler, A. (2012). Cochran-armitage test versus logistic regression in the analysis of genetic association studies. *Human heredity*, 73(1):14–17.

Westfall, P. and Young, S. (1993). *Resampling-Based Multiple Testing*. New York: Wiley.

Wold, S., Martens, H., and Wold, H. (1983). The multivariate calibration problem in chemistry solved by the pls method. pages 286–293.

Wold, S., Ruhe, A., Wold, H., and Dunn, III, W. (1984). The collinearity problem in linear regression. the partial least squares (pls) approach to generalized inverses. *SIAM Journal on Scientific and Statistical Computing*, 5(3):735–743.

Won, S., Kwon, M.-S., Mattheisen, M., Park, S., Park, C., Kihara, D., Cichon, S., Ophoff, R., Nöthen, M. M., Rietschel, M., et al. (2014). Efficient strategy for detecting gene× gene joint action and its application in schizophrenia. *Genetic epidemiology*, 38(1):60–71.

Woo, H. J., Yu, C., Kumar, K., and Reifman, J. (2017). Large-scale interaction effects reveal missing heritability in schizophrenia, bipolar disorder and posttraumatic stress disorder. *Translational psychiatry*, 7(4):e1089.

Wu, M. C., Lee, S., Cai, T., Li, Y., Boehnke, M., and Lin, X. (2011). Rare-variant association testing for sequencing data with the sequence kernel association test. *The American Journal of Human Genetics*, 89(1):82–93.

Wu, Z., Sun, Y., He, S., Cho, J., Zhao, H., Jin, J., et al. (2014). Detection boundary and higher criticism approach for rare and weak genetic effects. *The Annals of Applied Statistics*, 8(2):824–851.

Yu, Z., Demetriou, M., and Gillen, D. L. (2015). Genome-wide analysis of gene-gene and gene-environment interactions using closed-form wald tests. *Genetic epidemiology*, 39(6):446–455.

Zhang, C.-H. and Melnik, O. (2012). *cggd: Continuous Generalized Gradient Descent*. R package version 0.8.

Zhang, J.-T. (2013). *Analysis of variance for functional data*. CRC Press.

Zhang, J.-T., Cheng, M.-Y., Wu, H.-T., and Zhou, B. (2019). A new test for functional one-way anova with applications to ischemic heart screening. *Computational Statistics & Data Analysis*, 132:3–17.

Zhang, J.-T. and Liang, X. (2014). One-way anova for functional data via globalizing the pointwise F-test. *Scandinavian Journal of Statistics*, 41(1):51–71.

Zhang, X., Yang, X., Yuan, Z., Liu, Y., Li, F., Peng, B., Zhu, D., Zhao, J., and Xue, F. (2013). A plspm-based test statistic for detecting gene-gene co-association in genome-wide association study with case-control design. *PloS one*, 8(4):e62129.

Zhao, J., Jin, L., and Xiong, M. (2006). Test for interaction between two unlinked loci. *The American Journal of Human Genetics*, 79(5):831–845.

Zhao, S. D., Cai, T. T., Cappola, T. P., Margulies, K. B., and Li, H. (2017). Sparse simultaneous signal detection for identifying genetically controlled disease genes. *Journal of the American Statistical Association*, 112(519):1032–1046.

Zheng, G., Joo, J., and Yang, Y. (2009). Pearson's test, trend test, and max are all trend tests with different types of scores. *Annals of human genetics*, 73(2):133–140.

Zhong, P.-S., Chen, S. X., Xu, M., et al. (2013). Tests alternative to higher criticism for high-dimensional means under sparsity and column-wise dependence. *The Annals of Statistics*, 41(6):2820–2851.

Zuber, V. and Strimmer, K. (2009). Gene ranking and biomarker discovery under correlation. *Bioinformatics*, 25(20):2700–2707.

Zuber, V. and Strimmer, K. (2011). High-dimensional regression and variable selection using car scores. *Statistical Applications in Genetics and Molecular Biology*, 10(1).

Zuber, V. and Strimmer, K. (2017). *care: High-Dimensional Regression and CAR Score Variable Selection*. R package version 1.1.10.

Zuk, O., Hechter, E., Sunyaev, S. R., and Lander, E. S. (2012). The mystery of missing heritability: Genetic interactions create phantom heritability. *Proceedings of the National Academy of Sciences*, 109(4):1193–1198.

**Titre :** Prise en compte de la dépendance pour des problèmes de test global et de prédiction

**Mots clés :** dépendance, test global, prédiction, régression

**Résumé :** Dans de nombreux types de données, tels que les données génomiques ou fonctionnelles, les variables explicatives sont caractérisées par une forte structure de dépendance. Pour des problèmes variés, comme la construction de tests globaux ou de règles de prédiction, la prise en compte de cette structure de dépendance reste un problème ouvert. Plusieurs auteurs recommandent de prendre explicitement en compte cette dépendance, tandis que d'autres proposent de l'ignorer complètement. En réalité, il apparaît que le meilleur choix entre ces deux possibilités dépend à la fois de la structure de dépendance elle-même, mais aussi de la forme du signal d'association entre les variables explicatives et la variable réponse.

Dans cette thèse, des approches adaptatives sont proposées, visant à déterminer la meilleure façon de prendre en compte la dépendance. Une méthode de test global est construite, notamment pour les problèmes de tests d'association gène-phénotype en études d'association pangénomiques. De façon similaire, une règle de régression adaptative est développée. Dans les deux cas, la prise en compte adaptative de la dépendance est permise par l'introduction de poids. Une expression explicite des poids optimaux est ensuite obtenue. Celle-ci dépend à la fois de la structure de dépendance des variables explicatives et du signal d'association. Les résultats obtenus sur des simulations et des jeux de données non simulés démontrent que les méthodes proposées offrent de bonnes performances dans des situations variées.

**Title :** Dependence handling for global testing and prediction problems

**Keywords :** dependence, global testing, prediction, regression

**Abstract :** In various types of data, including genomic or functional data, explanatory variables are characterized by a strong dependence structure. For diverse problems, such as global testing or for the construction of prediction rules, handling this dependence structure remains an open issue. Several authors recommend to properly take the dependence into account, whereas others propose to completely ignore it. It appears that the best choice depends on both the dependence structure itself and the pattern of the association signal between the explanatory variables and the response.

In this thesis, adaptive approaches are proposed, aiming at determining the best way to handle dependence. An adaptive global testing method is therefore constructed, namely for gene testing problems in genome-wide association studies. Similarly, an adaptive regression rule is developed. In both cases, the flexible handling of dependence is performed by the introduction of weights. An explicit expression for the optimal weights is then derived, which depends on both the dependence structure of the explanatory variables and the association signal. The results obtained in simulation studies and on non-simulated datasets demonstrate that the proposed methods perform well in various situations.