An adaptive decorrelation procedure for signal detection

Florian Hébert, David Causeur, Mathieu Emily

An Adaptive Decorrelation Procedure for Signal Detection

Florian Hébert*, David Causeur, Mathieu Emily

Agrocampus Ouest, CNRS, IRMAR - UMR 6625, F-35000 Rennes, France

Abstract

In global testing, where a large number of pointwise test statistics are aggregated to simultaneously test for a collection of null hypotheses, the handling of dependence is a crucial issue. In various fields, more particularly in genetic epidemiology and functional data analysis, many testing methods for detecting an association signal between a response and explanatory variables have been proposed. Some aggregation procedures ignore 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 between the structure of dependence across pointwise test statistics and the pattern of the association signal. Within a large class of test statistics covering a continuum of decorrelation approaches, an optimal procedure is introduced. This procedure is based on the maximization of an *ad-hoc* cumulant generating function-based distance between the null and nonnull distributions of a global test statistic, in order to adapt the aggregation of the pointwise statistics to the pattern of the association signal. A comparative study including simulations and applications to genetic association studies demonstrates that the ability of this test to detect a signal is more robust to the dependence structure than existing methods.

Keywords: Decorrelation, Dependent tests, Global testing, Signal detection.

1. Introduction

In many research fields, signal detection is viewed as the simultaneous tests 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,

*Corresponding author - Postal address: Agrocampus Ouest, Batiment 24, 65 rue de Saint Brieuc, 35000 Rennes, France. Email address: `florian.hebert@agrocampus-ouest.fr`

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 [31] for fANOVA and [12] for GWAS issues) reflects the difficulty of identifying a method that would show a good detection performance in a wide scope of situations. As reported by [9] 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 comparison of aggregation procedures. Besides, several studies also investigate the influence of the pattern of the association signal, especially its sparsity rate [13, 2, 33], on the power of global testing methods. As demonstrated hereafter, the best way of aggregating pointwise test statistics actually depends on the interplay between the dependence structure and the association signal.

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, [22] suggest using the maximum absolute pointwise test statistics, which turns out to be analogous to the famous `minP` procedure, proposed by [10] to test for the 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 [30], whereas similar weighted or unweighted L^2-norm statistics are recommended by many authors [20, 28, 12] 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 [2]. The former paper focuses on the impact of the sparsity rate of the association signal on the choice between the L^2-norm based test statistics of standard Analysis of Variance and the higher criticism [13] in a wide variety of correlation patterns. The former higher criticism (HC) test statistics 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 signal, [14] show that HC reaches the optimal detection bounds obtained by [18].

It is commonly observed that, whatever the aggregation method, detection performance for a given association signal is affected by dependence across pointwise test statistics. A growing number of studies therefore suggest that signal detection procedures are improved by aggregating decorrelated pointwise test statistics, as for instance in [16] and [17] for HC and [1] for the slightly different feature selection issue in two-group classification models. Indeed, the innovated higher criticism (iHC) proposed by [17] first performs a whitening transformation of the original test statistics using the Cholesky decomposition of the correlation matrix and calculates the HC statistic on decorrelated pointwise test statistics. Similarly, [1] introduce Correlation-Adjusted t-scores based on a James-Stein shrinkage estimate of correlations.

However, as discussed in [6] in the closely related two-group classification issue, the potential gain in detection performance that can be expected from decorrelation remains unclear (see [29, 4] in the GWAS context). Indeed, arguing that decorrelation may generate noise and weaken the signal, [4] introduces the generalized higher criticism (GHC) procedure in which aggregation is performed on the raw pointwise test statistics and the impact of dependence is accounted for by an *ad-hoc* scaling of the HC statistic. Observing that the detection performance of the GHC procedure highly depends both on the pattern of dependence across pointwise statistics and on the sparsity rate of the signal, [4] propose to combine it with the maximum of the absolute pointwise test statistics and a weighted L^2-norm statistic in an omnibus testing approach. While the former test shows good detection performance in the simulation setup proposed in [4], 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, the global testing approach proposed hereafter aims at adapting the aggregation of pointwise tests to both the correlation structure and the pattern of the signal. For that purpose, a class of global test statistics defined as linear combinations of the squared decorrelated pointwise test statistics is introduced. The linear coefficients corresponding to each global test statistics within this class enable a flexible handling of dependence in the aggregation procedure and prevents from the dilution of the signal that can be induced by a complete whitening of the pointwise test statistics. A subclass of global test statistics is identified, which linear coefficients maximize a cumulant generating function-based distance between the null and non-null distributions of the test statistics. A global test statistic is deduced, by estimating the linear coefficients of the optimal test statistics within this subclass.

In Section 2, a general framework is introduced for the global testing of an association between a response and explanatory variables in a generalized linear model. In this framework, a list of popular signal detection methods by aggregation of pointwise test statistics is presented, covering a large scope of dependence handling strategies. The impact of the interplay between the patterns of dependence across pointwise statistics and association signal on the detection performance of these methods is demonstrated in an illustrative situation.

Section 3 introduces a class of global test statistics defined as linear combinations of squared decorrelated pointwise tests. A specific choice of linear coefficients gives the sum of squared pointwise test statistics, which ignores dependence in aggregation, whereas another choice gives the Hotelling's t-square test, which on the contrary accounts for dependence by introducing a preliminary whitening step. Other choices of linear coefficients leading to alternative dependence handling strategies, a procedure searching for an optimal choice of the linear coefficients is proposed. Finally, a comparative study of the proposed global testing procedure with a variety of alternative methods is conducted in Section 4, based on simulations under various assumptions on the dependence across explanatory variables and in the context of two genetic association studies.

2. Signal Detection by Aggregation of Pointwise Test Statistics

The global testing issue is usually presented in the standard linear regression model framework under normality assumptions [2, 12]. However, in order to enlarge the scope of applications to the association test between a binary response variable and a block of genetic markers in genetic epidemiology issues, the main existing aggregation methods for signal detection are introduced in a generalized linear model (GLM) framework [10, 28, 4].

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

Setup for the association signal.

In the following, Y denotes a response 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 $X_j = X(t_j),\ j = 1, \ldots, p$ of a curve $X :\ t \mapsto X(t)$ on a time grid $t_1, \ldots, t_p$ and Y can either be a continuous variable, usually assumed to be normally distributed given $\boldsymbol{X} = \boldsymbol{x}$ in scalar-on-function regression issues or a grouping variable for supervised classification of functional data. In GWAS issues, when related to genomic selection in animal or plant science, Y can also be a continuous variable measuring a yield whereas, in most genetic epidemiology issues, Y is a binary (case/control) variable. In the former genomic context, $\boldsymbol{X} = (X_1, \ldots, X_p)$ is a profile of three-level categorical genetic markers observed for a set of contiguous Single-Nucleotide Polymorphisms in a given region of the genome.

The former sample of typical situations for global testing applications is covered by the following GLM settings, as suggested by [10] and [4] for conditional distributions of Y in the exponential family [21]:

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

where h is a link function, $\boldsymbol{U} = (U_1, \ldots, U_q)'$ is an optional q-vector of covariates with corresponding regression coefficients $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ is the p-vector of regression coefficients for the explanatory variables.

The need for covariate adjustment is especially common in GWAS where a population structure identifying marked clusters or external environmental conditions are suspected to explain a part of the variations of the response. The canonical link function is often chosen for h, namely the identity function in continuous trait analysis or the logit function for case-control studies [29, 25, 32].

In the above framework, signal detection is viewed as a global test for the significance of the signal $\boldsymbol{\beta}$ of regression coefficients:

$$\begin{cases} \mathrm{H}_0 : & \boldsymbol{\beta} = \boldsymbol{0} \\ \mathrm{H}_1 : & \boldsymbol{\beta} \neq \boldsymbol{0}. \end{cases} \tag{2}$$

Signal detection.

Hereafter, $\boldsymbol{y} = (y_1, \ldots, y_n)'$ denotes the n-vector of observations of the response variable, $\mathbb{X}$ the $n \times p$ matrix whose jth 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 regular designs where $n \gg p + q$, under the GLM assumption (1), the testing issue (2) can be addressed using Analysis of Variance F-tests for the comparison of nested models if the conditional distribution of Y given $\boldsymbol{U}$ and $\boldsymbol{X}$ is normal or chi-square Likelihood-Ratio Tests under alternative assumptions on the distribution of Y within the exponential family. However, such well-proven tests cannot be implemented in large ($n \approx p + q$) or high ($n \ll p + q$) dimensional situations, where maximum likelihood (ML) estimation of model (1) is either numerically unstable or impossible. An alternative approach is to form a global test statistic from the aggregation of the pointwise test statistics for the marginal association between the response and each explanatory variable in the following models: for $j = 1, \ldots, p$,

$$h(\mathbb{E}[Y|\boldsymbol{U} = \boldsymbol{u}, X_j = x_j]) = \boldsymbol{u}'\boldsymbol{\alpha}_j^\star + \boldsymbol{x}_j\beta_j^\star,$$

where $\boldsymbol{\alpha}_j^\star$ is the q-vector of regression parameters for the covariates and $\beta_j^\star$ is the marginal association parameter between Y and X_j.

A usual choice for testing $\mathrm{H}_{0j}: \ \beta_j^\star = 0$ is the Student's t-test when Y is assumed to be normally distributed conditionally on $\boldsymbol{U}$ and $\boldsymbol{X}$ or the equivalent Wald's test under other assumptions on the conditional distribution of Y within the exponential family. In the GLM framework of model (1), [10] and [4] also propose a marginal testing procedure based a similar z-score. First, under the global null hypothesis H_0, model (1) becomes:

$$h(\mathbb{E}[Y|\boldsymbol{U} = \boldsymbol{u}, \boldsymbol{X} = \boldsymbol{x}]) = \boldsymbol{u}'\boldsymbol{\alpha}_0, \tag{3}$$

where $\boldsymbol{\alpha}_0$ is a q-vector of regression coefficients. Let $\hat{\boldsymbol{\alpha}}_{\boldsymbol{0}}$ be the Maximum Likelihood estimator of $\boldsymbol{\alpha}_0$. We denote $\hat{\boldsymbol{y}}_{\boldsymbol{0}}$ the n-vector whose i-th coordinate equals $h^{-1}(\boldsymbol{u_i}'\hat{\boldsymbol{\alpha}}_{\boldsymbol{0}})$, where $\boldsymbol{u_i}$ is the ith row of $\mathbb{U}$. For testing the significance of an association between Y and X_j, the following marginal test statistic is proposed by [10] and [4]:

$$Z_j = \frac{\boldsymbol{X_j}'(\boldsymbol{y} - \hat{\boldsymbol{y}}_{\boldsymbol{0}})}{\sqrt{\hat{\boldsymbol{\Gamma}}_{j,j}}}. \tag{4}$$

where $\hat{\boldsymbol{\Gamma}} = \hat{\sigma}_y^2(\mathbb{X}'\mathbb{X} - \mathbb{X}'\mathbb{U}(\mathbb{U}'\mathbb{U})^{-1}\mathbb{U}'\mathbb{X})$ is the estimated covariance matrix of the vector $\mathbb{X}'(\boldsymbol{y} - \hat{\boldsymbol{y}}_{\boldsymbol{0}})$, with $\hat{\sigma}_y^2 = (\boldsymbol{y} - \hat{\boldsymbol{y}}_{\boldsymbol{0}})'(\boldsymbol{y} - \hat{\boldsymbol{y}}_{\boldsymbol{0}})/n$.

It is clearly beyond the scope of the present paper to discuss in more detail the choice of a marginal testing procedure. As in the Rare-and-Weak paradigm introduced by [13] for the higher criticism procedure, it will be assumed hereafter that the p-vector $\boldsymbol{\mathcal{Z}} = (Z_1, \ldots, Z_p)'$ of pointwise 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 variance matrix $\boldsymbol{\Sigma}$:

$$\boldsymbol{\mathcal{Z}} = (Z_1, \ldots, Z_p)' \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}). \tag{5}$$

Although the pointwise association parameter $\boldsymbol{\mu}$ and the regression parameter $\boldsymbol{\beta}$ can show different patterns, the hypothesis testing issue (2) coincides with the test for the nullity of $\boldsymbol{\mu}$:

$$\begin{cases} \mathrm{H}_0: & \boldsymbol{\mu} = \mathbf{0} \\ \mathrm{H}_1: & \boldsymbol{\mu} \neq \mathbf{0}, \end{cases} \quad i.e. \quad \begin{cases} \mathrm{H}_0: & \forall j \in \{1, \ldots, p\},\ \mu_j = 0 \\ \mathrm{H}_1: & \exists j \in \{1, \ldots, p\},\ \mu_j \neq 0. \end{cases} \tag{6}$$

Let us first illustrate this point in the special case of model (1) under assumption of a joint multinormal distribution for Y and $\boldsymbol{X}$ given $\boldsymbol{U}$. Here, the expectation $\boldsymbol{\mu}$ of the vector of pointwise Student t-tests in models (3) is proportional to the p-vector $\boldsymbol{\beta}^\star = (\beta_1^\star, \ldots, \beta_p^\star)'$ of marginal association parameters:

$$\boldsymbol{\mu} \propto \boldsymbol{\beta}^\star = \boldsymbol{D}_{\sigma^2_{x|u}}^{-1} \boldsymbol{\sigma}_{x,y|u}, \tag{7}$$

where $\boldsymbol{D}_{\sigma^2_{x|u}}$ is the $p \times p$ diagonal matrix whose diagonal entries form the p-vector $\sigma^2_{x|u}$ of conditional variances of $\boldsymbol{X}$ given $\boldsymbol{U}$ and $\boldsymbol{\sigma}_{x,y|u}$ is the p-vector of conditional covariances between Y and each explanatory variables given $\boldsymbol{U}$.

Provided the conditional variance-covariance matrix $\boldsymbol{\Sigma}_{x|u}$ of $\boldsymbol{X}$ given $\boldsymbol{U}$ has rank p, $\boldsymbol{\beta} = \boldsymbol{\Sigma}_{x|u}^{-1}\boldsymbol{\sigma}_{x,y|u} = \boldsymbol{\Sigma}_{x|u}^{-1}\boldsymbol{D}_{\sigma^2_{x|u}}\boldsymbol{\beta}^\star$. Consequently, $\boldsymbol{\beta} = 0$ if and only if $\boldsymbol{\mu} \propto \boldsymbol{\beta}^\star = 0$. However, except in the situation where the explanatory variables are uncorrelated given $\boldsymbol{U}$, the set of nonzero coordinates in $\boldsymbol{\beta}$ is generally different from the set of nonzero coordinates in $\boldsymbol{\mu}$. This explains that the signal identification issue, consisting in finding the nonzero coordinates of $\boldsymbol{\beta}$, based on $\boldsymbol{\mathcal{Z}}$ under dependence is far more challenging than the global testing issue [17, 24].

In most papers addressing global testing issues, a rare signal usually refers to a sparse vector $\boldsymbol{\beta}$ in Equation (1), which, due to the correlation among the variables, does not imply the sparsity of $\boldsymbol{\mu}$. For association studies in genetic epidemiology, where genetic markers are dependent three-level categorical variables, this is illustrated using two data-driven simulation scenarios, one with a sparse $\boldsymbol{\beta}$ (`scenario 1`) and another one with a non-sparse $\boldsymbol{\beta}$ (`scenario 2`). Both simulation scenarios reproduce the conditions of an association study between a binary response variable Y and the region of the genome made by a block of $p = 64$ genetic markers forming a Linkage Disequilibrium (LD) block on chromosome 1. For each simulated dataset, 1,000 independent profiles of dependent genetic markers data, whose both marginal and joint distributions are estimated using a publicly available GWAS dataset [26], are generated using the `R` package `GenOrd` [3]. Then, consistently with model (1), for each simulated profile of genetic markers data, the response is simulated according to a logistic linear regression model, where the only nonzero coordinates of the vector $\boldsymbol{\beta}$ of regression coefficients in `scenario 1` are for the 5th and 10th genetic marker, whereas in `scenario 2`, a few other coefficients around the 5th and 10th genetic marker are also set to nonzero values. For each simulated dataset, the vector $\boldsymbol{\mathcal{Z}}$ of pointwise test statistics as defined by equation (4) is calculated, without any covariate adjustment.

Both the regression coefficients $\boldsymbol{\beta}$ and the corresponding pointwise association signal $\boldsymbol{\mu}$, deduced by averaging over 1,000 simulations of the vector $\boldsymbol{\mathcal{Z}}$ as

detailed above, are represented in Figure 1. As previously shown in the special case of a normal joint distribution of Y and $\boldsymbol{X}$, it can be remarked that the patterns of $\boldsymbol{\beta}$ and $\boldsymbol{\mu}$, especially their sparsity rates, are also very different in the present logistic regression framework.

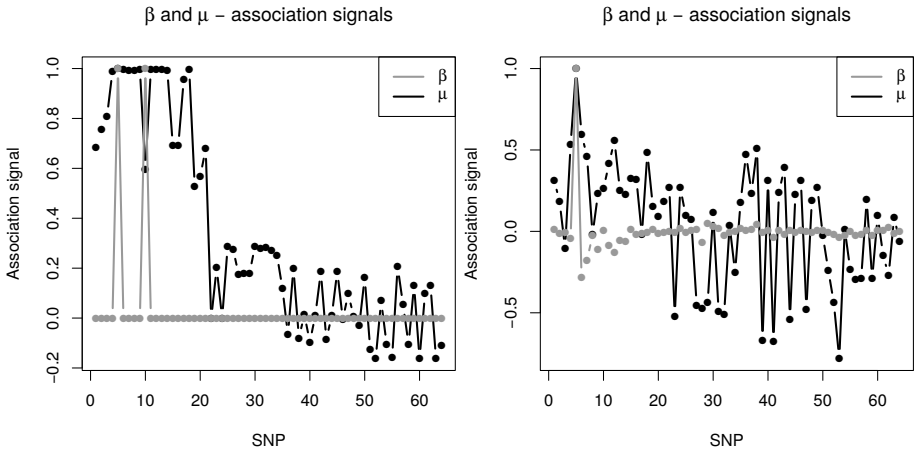

Figure 1: Regression coefficients ($\boldsymbol{\beta}$) and corresponding pointwise association signal ($\boldsymbol{\mu}$) in scenarios reproducing a genetic association study between a binary response variable and an LD block of 64 SNPs on chromosome 1. `Scenario 1` (Left): only the SNPs located at the 5th and 10th positions have non-zero regression coefficients. `Scenario 2` (Right): a few other coefficients around those SNPs are also nonzero.

2.2. Aggregation procedures of pointwise statistics

For the testing issue (6), a combined test statistic $T(\boldsymbol{\mathcal{Z}})$ is obtained by aggregating the pointwise statistics Z_j. The most commonly used aggregation methods are briefly introduced hereafter and summarized in Table 1.

One of the most used aggregation methods is the maximum squared pointwise test statistic (see [22] for fANOVA and [27, 10] for GWAS), for which $T(\boldsymbol{\mathcal{Z}}) = T_{\max}(\boldsymbol{\mathcal{Z}}) = \max_j Z_j^2$. The former method, called `minP` by [27], is indeed often preferred to alternatives because its ability to reveal a true signal turns out to be generally good and resistant to strong correlations between the coordinates of $\boldsymbol{\mathcal{Z}}$ [29].

The squared L^2-norm of $\boldsymbol{\mathcal{Z}}$, denoted $T_2(\boldsymbol{\mathcal{Z}})$ and defined as the sum of the squared Z_j is also proposed by [20] in the GWAS context and by [23] to define a functional F-test. Weighted versions of the L^2-norm statistic are also widely used when the goal is to identify an association with explanatory variables having a smaller variance. Especially designed for the identification of rare variants in the GWAS context, the Sequence Kernel Association Test (`SKAT`, [28]) is one of them, for which the jth weight depend on the minor allele frequency of the jth genetic marker via a Beta function (see Table 1).

The Hotelling's t-square test, denoted $T_{\mathrm{H}}(\boldsymbol{\mathcal{Z}})$ (see [12] for applications to GWAS issues), can just be viewed as a whitened version of the L^2-norm statistic. Indeed, provided $\hat{\boldsymbol{\Sigma}}$ is a consistent estimate of $\boldsymbol{\Sigma}$ and $\hat{\boldsymbol{\Sigma}}^{-1}$ exists, $T_{\mathrm{H}}(\boldsymbol{\mathcal{Z}}) = \boldsymbol{\mathcal{Z}}'\hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{\mathcal{Z}}$ is the L^2-norm of the decorrelated vector $\hat{\boldsymbol{\Sigma}}^{-1/2}\boldsymbol{\mathcal{Z}}$, where $\hat{\boldsymbol{\Sigma}}^{-1/2}$ is any $p \times p$ matrix satisfying $(\hat{\boldsymbol{\Sigma}}^{-1/2})'\hat{\boldsymbol{\Sigma}}^{-1/2} = \hat{\boldsymbol{\Sigma}}^{-1}$.

The higher criticism (HC) signal detection method introduced by [13] can also be cast into the above testing framework for signal detection. Indeed, the HC test statistic aggregates the pointwise test statistics by forming a Kolmogorov-Smirnov type distance between the empirical probability distribution function of the pointwise 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 $S(t)$, straightforwardly deduced from the binomial distribution of $S(t)$ when the pointwise test statistics are assumed to be independent.

However, when the former independence assumption is relaxed, the null distribution of $S(t)$ is no longer binomial, which requires to update the scaling factor in the HC statistic. Consistently, [4] propose to scale the HC statistic by the empirical standard error of $S(t)$ under dependence in the generalized higher criticism (`GHC`) method. In order to improve the robustness of the former test to various patterns of dependence, [4] also suggest an omnibus test statistic T_{Omn}, which p-value is derived by taking the smallest among the p-values obtained with `minP`, `SKAT` and `GHC`.

Table 1: A sample of frequently used aggregation methods. $\boldsymbol{D}_\beta$ is a diagonal matrix which jth diagonal entry is a weight depending on the Minor Allele Frequency of the jth SNP via a Beta function. Φ is the cumulative distribution function of the standard normal distribution. $S(t)$ is defined as $S(t) = \sum_{j=1}^{p} \mathbb{I}_{\{|Z_j| \geq t\}}$. p_{GHC}, p_{minP} and p_{SKAT} are the p-values for the aggregation methods `GHC`, `minP` and `SKAT` respectively.

Method	Statistic
`minP`	$T_{\max}(\boldsymbol{\mathcal{Z}}) = \max_{1 \leq j \leq p} Z_j^2$
L^2-norm	$T_2(\boldsymbol{\mathcal{Z}}) = \boldsymbol{\mathcal{Z}}'\boldsymbol{\mathcal{Z}} = \sum_{j=1}^{p} Z_j^2$
Hotelling	$T_{\mathrm{H}}(\boldsymbol{\mathcal{Z}}) = \boldsymbol{\mathcal{Z}}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mathcal{Z}}$
`SKAT`	$T_{\mathrm{SKAT}}(\boldsymbol{\mathcal{Z}}) = \boldsymbol{\mathcal{Z}}'\boldsymbol{D}_\beta\boldsymbol{\mathcal{Z}}$
`HC`	$T_{\mathrm{HC}}(\boldsymbol{\mathcal{Z}}) = \max_{1 \leq j \leq p/2} \frac{j/m - p_{(j)}}{\sqrt{p_{(j)}(1-p_{(j)})}}$
`GHC`	$T_{\mathrm{GHC}}(\boldsymbol{\mathcal{Z}}) = \sup_{t \geq t_0} \left\{ \frac{S(t) - 2p(1-\Phi(t))}{\sqrt{\widehat{\mathrm{var}}(S(t))}} \right\}$
Omnibus	$T_{\mathrm{Omn}} = \min\left(p_{\mathrm{GHC}}, p_{\mathrm{minP}}, p_{\mathrm{SKAT}}\right)$

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

Many comparative studies [12, 4] 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 to shed light on the fact that, even for a same correlation structure, the compared performance of signal detection methods is also markedly affected by the pattern of the pointwise association parameter $\boldsymbol{\mu}$ introduced in expression (5).

The simulation settings presented in Subsection 2.1 introduce two patterns for $\boldsymbol{\mu}$ displayed in Figure 1, corresponding each to a different scenario for the regression parameter $\boldsymbol{\beta}$. For 1,000 simulated datasets in these two scenarios, the p-value of the aggregation methods introduced in Table 1 is estimated using a Monte-Carlo estimation of the null distribution based on 1,000 random permutations of the observations of Y. Under both simulation scenarios, a sequence of regression signals $\boldsymbol{\beta} = \xi\boldsymbol{\beta}_{\mathbf{max}}$ with increasing amplitudes is generated by scaling a vector of regression coefficients $\boldsymbol{\beta}_{\mathbf{max}}$, where $\xi \in \{0, 0.1, 0.2, \ldots, 1\}$ is a scaling factor controlling the signal strength and $\boldsymbol{\beta}_{\mathbf{max}}$ corresponds to a signal amplitude for which the power of the most powerful test is close to 1. Finally, for each combination of a pattern of association signal $\boldsymbol{\mu}$ and a signal amplitude, the power of each global testing method is estimated by the proportion of significant tests, with p-values smaller than 0.05.

The resulting power curves are represented in Figure 2. It clearly shows that the ranking of aggregation methods depends on the pattern of the pointwise association signal $\boldsymbol{\mu}$. Focusing on the 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. In `scenario 2`, the ranking is reversed since Hotelling's test is by far the most powerful method while L^2-norm and `SKAT` perform as the worst methods.

Such a result confirms that no single aggregation test is the most powerful over the combinations of a correlation structure and a pattern for the association signal [4]. This also suggests that, whatever the pattern of $\boldsymbol{\mu}$, the scope of tests defined as quadratic forms in $\boldsymbol{\mathcal{Z}}$ may be large enough for the search of the most powerful method. Moreover, recalling that the Hotelling's t-square test is a whitened version of the L^2-norm test, this also demonstrates that the potential benefit of decorrelating the test statistics depends on the pattern of $\boldsymbol{\mu}$. In the next Section, a new class of aggregation methods allowing for a flexible whitening of the test statistics is therefore introduced.

3. Flexible Decorrelation of Pointwise Test Statistics

Suppose first that $\boldsymbol{\Sigma}$ is known, with eigendecomposition $\boldsymbol{V}\boldsymbol{D}_{\boldsymbol{\lambda}}\boldsymbol{V}'$ where $\boldsymbol{V}$ is the $p \times p$ matrix of eigenvectors such that $\boldsymbol{V}'\boldsymbol{V} = \mathbf{I}_{\boldsymbol{p}}$ and $\boldsymbol{D}_{\boldsymbol{\lambda}}$ is a $p \times p$ diagonal matrix whose diagonal entries are the positive eigenvalues λ_j, $j = 1, \ldots, p$. Let $\boldsymbol{\mathcal{Z}}^* = \boldsymbol{V}'\boldsymbol{\mathcal{Z}}$ denote the decorrelated version of $\boldsymbol{\mathcal{Z}}$. Under the asymptotic normality assumption on $\boldsymbol{\mathcal{Z}}$ introduced in expression (5), $\boldsymbol{\mathcal{Z}}^*$ is also asymptotically normally distributed with mean $\boldsymbol{V}'\boldsymbol{\mu}$ and variance $\mathrm{Var}(\boldsymbol{\mathcal{Z}}^*) = \boldsymbol{D}_{\boldsymbol{\lambda}}$.

Now, let us introduce the class $\mathcal{T}_2$ of aggregated test statistics $T_2(\boldsymbol{\mathcal{Z}}^*; \boldsymbol{h})$ for $\mathrm{H}_0 : \boldsymbol{\mu} = \mathbf{0}$ defined as linear combinations of the squared coordinates Z_j^* of $\boldsymbol{\mathcal{Z}}^*$:

$$\mathcal{T}_2 = \left\{ T_2(\boldsymbol{\mathcal{Z}}^*; \boldsymbol{h}) = \sum_{j=1}^{p} h_j [Z_j^*]^2, \boldsymbol{h} = (h_1, \ldots, h_p) \in \mathbb{R}^p, \ \sum_{j=1}^{p} h_j = p \right\}. \quad (8)$$

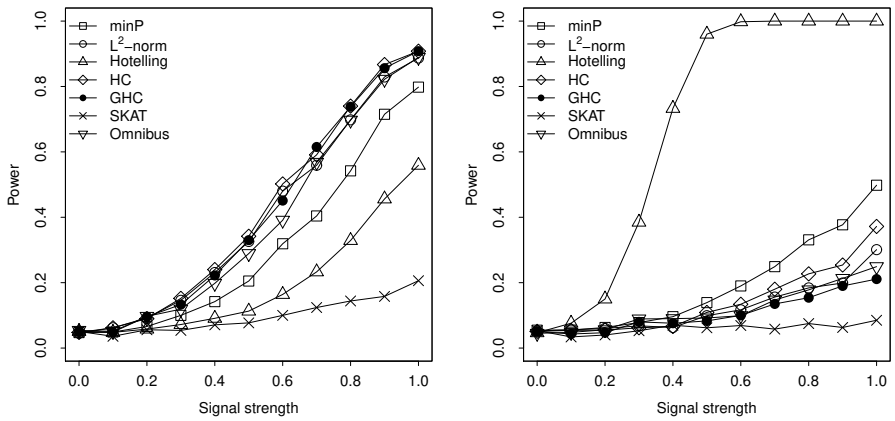

Figure 2: Detection power of the L^2-norm test, `minP`, `HC`, `GHC`, the Hotelling's t-square test, `SKAT` and the omnibus test (with type-I error level $\alpha = 0.05$) in the two simulation scenarios presented in Figure 1 (Left: `scenario 1`. Right: `scenario 2`).

The linear restriction $\sum_{j=1}^{p} h_j = p$ fixing the sum of linear coefficients to an arbitrary value is due to the invariance of the power of the global test statistics in $\mathcal{T}_2$ with respect to the multiplication by a scalar.

Interestingly, $\mathcal{T}_2$ contains two global test statistics with opposite strategies regarding the handling of dependence across pointwise test statistics:

- $T_2(\boldsymbol{\mathcal{Z}}) = T_2(\boldsymbol{\mathcal{Z}}^*; \mathbf{1}_p)$, for which dependence is ignored, where $\mathbf{1}_p$ is the p-vector whose all entries equal one,
- $T_{\text{H}}(\boldsymbol{\mathcal{Z}}) = T_2(\boldsymbol{\mathcal{Z}}^*; \boldsymbol{\omega}_{\text{H}})$, for which dependence is accounted for by whitening $\boldsymbol{\mathcal{Z}}$, where $\boldsymbol{\omega}_{\text{H}} = (\omega_{1\text{H}}, \ldots, \omega_{p\text{H}})'$ with $\omega_{j\text{H}} = p(1/\lambda_j)/(\sum_{k=1}^{p} 1/\lambda_k)$, for $j = 1, \ldots, p$.

Note that, neither with $T_2(\boldsymbol{\mathcal{Z}})$ nor with $T_{\text{H}}(\boldsymbol{\mathcal{Z}})$, the handling of dependence depends on the pattern of the association signal $\boldsymbol{\mu}$.

By an appropriate choice of the linear coefficients $\boldsymbol{h}$, the class $\mathcal{T}_2$ defines a large framework for a flexible handling of dependence. We propose hereafter to search for the vector of linear coefficients corresponding to an optimal global test statistic within $\mathcal{T}_2$.

3.1. Oracle Testing Procedure

For all $\boldsymbol{h} \in \mathbb{R}^p$, $T_2(\boldsymbol{\mathcal{Z}}^*; \boldsymbol{h})$ is distributed as $\sum_{j=1}^{p} h_j \lambda_j \chi_1^2(\gamma_j)$, where the χ_1^2 variables in the former sum are mutually independent and $\gamma_j = (\boldsymbol{v}_j' \boldsymbol{\mu})^2/\lambda_j$ are noncentrality parameters.

The optimal vector of linear coefficients is chosen to maximize a distance between the null and non-null distributions of $T_2(\boldsymbol{\mathcal{Z}}^*; \boldsymbol{h})$. In the present context

of a linear combination of noncentral chi-square distributions, using the most usual distribution-based distances, such as Cramér-Von Mises-type distances or Kullback-Leibler divergence, leads to untractable calculations of cumulative or probability distribution functions.

However, the moment generating function MGF, or equivalently the cumulant generating function CGF defined as follows:

$$\begin{aligned} \text{MGF}: &\quad t \mapsto \text{MGF}(t; \boldsymbol{h}, \boldsymbol{\lambda}, \boldsymbol{\gamma}) = \mathbb{E}[\exp(tT_2(\boldsymbol{\mathcal{Z}}^*; \boldsymbol{h}))], \\ \text{CGF}: &\quad t \mapsto \text{CGF}(t; \boldsymbol{h}, \boldsymbol{\lambda}, \boldsymbol{\gamma}) = \log \text{MGF}(t; \boldsymbol{h}, \boldsymbol{\lambda}, \boldsymbol{\gamma}) \end{aligned}$$

offer convenient alternatives to define a distance between linear combinations of independent noncentral chi-square distributions.

Indeed, for all $\boldsymbol{h} \in \mathbb{R}^p$ and t such that, for all $j = 1, \ldots, p$, $1 - 2th_j\lambda_j > 0$, the log-ratio of non-null and null moment generating functions of $T_2(\boldsymbol{\mathcal{Z}}^*; \boldsymbol{h})$ turns out to have a simple closed-form expression:

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

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 $T_2(\boldsymbol{\mathcal{Z}}^*; \boldsymbol{h})$. Since the former log-ratio is always larger than $-\sum_{j=1}^{p} \gamma_j/2$, then, for all $t < \min_j\{1/2h_j\lambda_j\}$,

$$\begin{aligned} \Delta(t; \boldsymbol{h}, \boldsymbol{\lambda}, \boldsymbol{\gamma}) &= \text{CGF}(t; \boldsymbol{h}, \boldsymbol{\lambda}, \boldsymbol{\gamma}) - \text{CGF}_0(t; \boldsymbol{h}, \boldsymbol{\lambda}) + \frac{1}{2} \sum_{j=1}^{p} \gamma_j, \\ &= \frac{1}{2} \sum_{j=1}^{p} \frac{\gamma_j}{1 - 2th_j\lambda_j} > 0. \end{aligned} \tag{9}$$

Moreover, $\Delta(t; \boldsymbol{h}, \boldsymbol{\lambda}, \boldsymbol{\gamma})$ depends jointly on the pointwise association signal and the dependence across pointwise test statistics through the parameters $(\boldsymbol{\lambda}, \boldsymbol{\gamma})$.

First, $\Delta(t; \boldsymbol{h}, \boldsymbol{\lambda}, \boldsymbol{\gamma})$ is maximized with respect to $\boldsymbol{h} \in \mathbb{R}^p$, under the linear restriction $\sum_{j=1}^{p} h_j = p$. For all t, the resulting vector $\boldsymbol{h}_t^\star(\boldsymbol{\gamma}, \boldsymbol{\lambda})$ of optimal linear coefficients, obtained using a Lagrange multiplier technique, has the following closed-form expression (details are given in the supplementary material, Appendix A):

$$\boldsymbol{h}_t^\star(\boldsymbol{\gamma}, \boldsymbol{\lambda}) = \frac{1}{\kappa_t} \boldsymbol{\omega}_{\text{H}} - (1 - \frac{1}{\kappa_t}) \boldsymbol{\omega},$$

where $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_p)'$ with $\omega_j = p\sqrt{\gamma_j/\lambda_j} / \sum_{k=1}^{p} \sqrt{\gamma_k/\lambda_k}$, for $j = 1, \ldots, p$ and $\kappa_t = 2pt/(\sum_{j=1}^{p} 1/\lambda_j)$. It is straightforwardly checked that, for all t and for all $(\boldsymbol{\lambda}, \boldsymbol{\gamma})$, $1 - 2th_{jt}^\star\lambda_j > 0$, for all $j = 1, \ldots, p$, where $h_{jt}^\star$ is the jth coordinate of $\boldsymbol{h}_t^\star(\boldsymbol{\gamma}, \boldsymbol{\lambda})$.

Consequently, and still considering that the association signal $\boldsymbol{\mu}$, hence also $\boldsymbol{\gamma}$, is known, the test statistics in $\mathcal{T}_2$ with linear coefficients $\boldsymbol{h}_t^\star(\boldsymbol{\gamma}, \boldsymbol{\lambda})$ can be

viewed as linear combinations of Hotelling's t-square statistic $T_2(\boldsymbol{\mathcal{Z}}^*;\boldsymbol{\omega}_{\mathrm{H}})$ and $T_2(\boldsymbol{\mathcal{Z}}^*;\boldsymbol{\omega})$:

$$T_2^\star(\boldsymbol{\mathcal{Z}}^*;\kappa) = \frac{1}{\kappa}T_2(\boldsymbol{\mathcal{Z}}^*;\boldsymbol{\omega}_{\mathrm{H}}) - (1-\frac{1}{\kappa})T_2(\boldsymbol{\mathcal{Z}}^*;\boldsymbol{\omega}).$$

The class $\mathcal{T}_2^\star = \{T_2^\star(\boldsymbol{\mathcal{Z}}^*;\kappa),\ \kappa\in\mathbb{R}\}$ of Oracle test statistics is a subclass of $\mathcal{T}_2$ indexed by a single scalar κ. The special case $\kappa = 1$ gives the Hotelling's t-square test and $\lim_{\kappa\to\pm\infty} T_2^\star(\boldsymbol{\mathcal{Z}}^*;\kappa) = \pm T_2(\boldsymbol{\mathcal{Z}}^*;\boldsymbol{\omega})$.

More generally, κ can be viewed as a tuning parameter to adapt the linear coefficients of the global test statistics within $\mathcal{T}_2^\star$ to the patterns of dependence and pointwise association signal defined by $(\boldsymbol{\lambda},\boldsymbol{\gamma})$. Again, this point is illustrated by comparing the two scenarios for $(\boldsymbol{\lambda},\boldsymbol{\gamma})$ on an LD block in chromosome 1 introduced in Section 2 (see Figure 1). For each signal strength, the power of the global test statistics within $\mathcal{T}_2^\star$ is calculated and the optimal value $\kappa^\star$ of κ is deduced. Similarly as in Figure 2 for the global testing methods summarized in Table 1, Figure 3 displays the detection rates of the optimal test statistic $T_2^\star(\boldsymbol{\mathcal{Z}}^*;\kappa^\star)$ along the signal strength in the two scenarios.

In `scenario 1`, whatever the signal strength, the best choice for κ is a large negative value whereas it is a small value in `scenario 2`. In both scenarios, it turns out that there exists global test statistics within $\mathcal{T}_2^\star$ showing a better detection rate than the L^2-norm test and the Hotelling's t-square test.

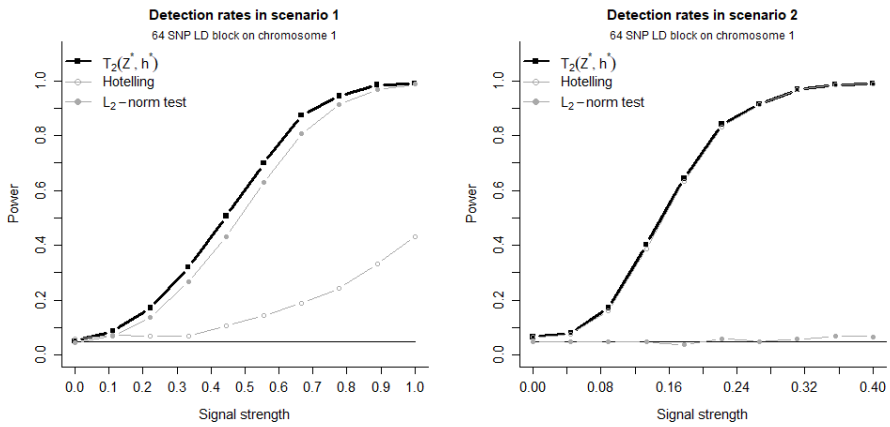

Figure 3: Power curves of the Oracle test statistics $T_2^\star(\boldsymbol{\mathcal{Z}}^*;\kappa^\star)$, the L^2-norm test and the Hotelling's t-square test in the two scenarios for an association signal $\boldsymbol{\mu}$ introduced in Section 2 (Left: `scenario 1`. Right: `scenario 2`).

3.2. Implementation of the `MGF-R` Procedure

A plug-in version $\hat{\boldsymbol{h}}_{\boldsymbol{t}}^\star = \boldsymbol{h}_{\boldsymbol{t}}^\star(\hat{\boldsymbol{\gamma}},\hat{\boldsymbol{\lambda}})$ of the Oracle vector of linear coefficients 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 $\hat{\boldsymbol{\Sigma}}$ of $\boldsymbol{\Sigma}$. Similarly, the coordinates of $\hat{\boldsymbol{\gamma}}$ are defined by $\hat{\gamma}_j = [Z_j^*]^2/\hat{\lambda}_j$, where $\hat{\boldsymbol{\mathcal{Z}}}^* = \hat{\boldsymbol{V}}'\boldsymbol{\mathcal{Z}}$ and $\hat{\boldsymbol{V}}$ is the $p \times p$ matrix of eigenvectors of $\hat{\boldsymbol{\Sigma}}$.

In the case without covariates $\boldsymbol{U}$ in model (1), the null distribution of $T_2^\star(\hat{\boldsymbol{\mathcal{Z}}}^*; \kappa)$ is approximated by a Monte-Carlo procedure based on a large number K of random permutations of $\boldsymbol{y}$. For each dataset obtained after random permutation, values $\left\{T_{2k}^{(0)}(\kappa),\ k = 1, \ldots, K,\ \kappa \in \mathbb{R}\right\}$ of $T_2^\star(\hat{\boldsymbol{\mathcal{Z}}}^*, \kappa)$ under the null hypothesis are calculated. An approximation $\hat{p}_\kappa^\star$ of the p-value for $T_2^\star(\hat{\boldsymbol{\mathcal{Z}}}^*, \kappa)$ is obtained by taking the proportion of random permutations for which $T_{2k}^{(0)}(\kappa)$ exceeds the observed value of $T_2^\star(\hat{\boldsymbol{\mathcal{Z}}}^*, \kappa)$. Finally, the smallest p-value in $\{\hat{p}_\kappa^\star,\ \kappa \in \mathbb{R}\}$ is denoted $\hat{p}^\star$.

The same procedure is applied to each null value $T_{2k}^{(0)}(\kappa)$ obtained after random permutation, leading to p-values $\hat{p}_k^\star$. Finally, the p-value $\hat{p}$ of the optimal global test statistics within $\mathcal{T}_2$ is obtained by taking the proportion of p-values $\hat{p}_k^\star$ smaller than $\hat{p}^\star$.

In the presence of covariates $\boldsymbol{U}$ in model (1), the above random permutation procedure has to be adapted to preserve the relationships between Y and $\boldsymbol{U}$ on the one hand and between $\boldsymbol{X}$ and $\boldsymbol{U}$ on the other hand. One solution consists in randomly permuting $\boldsymbol{y} - \hat{\boldsymbol{y}}_0$, as defined in expression (4) of the pointwise test statistics, instead of $\boldsymbol{y}$. Consistently, still in expression (4), $\boldsymbol{X}_j$ would also be replaced by the residuals of a regression model of X_j on the covariates. However, in the context of genomic association studies, [8, 15] advocate for the use of parametric resampling methods accounting for the suspected effect of confounding covariates, not explicitly included in the model.

4. Comparison of Detection Methods for Association Signals

4.1. Simulation Study

Both the ability of the `MGF-R` procedure to control the type-I error rate and its statistical power compared to other global testing procedures are investigated hereafter for a large scope of simulation scenarios. The general framework of those simulation scenarios is similar as in the study introduced in Subsection 2.1: reproducing the conditions of a genetic association study, the $n = 2,000$ profiles $\boldsymbol{X}$ of explanatory variables in each simulated dataset are independent p-vectors of three-level categorical variables, the response is a binary variable with equal prior probabilities and the association model is the logistic regression model (1) without additional covariates $\boldsymbol{U}$. The marginal and joint distributions of the profiles of explanatory variables are either set by parametric assumptions or reproduce the distribution estimated using the WTCCC dataset [26]. The `R` package `GenOrd` [3] is used to generate the profiles of dependent three-level explanatory variables.

The global testing methods summarized in Table 1 are included in the comparative study. All methods are implemented using `R`: the `SKAT` method via the package `SKAT` [19] and the `GHC` method via the package `GHC` [4]. The `MGF-R`

Table 2: Average computation time (in seconds) for each global testing method in the comparative study (see Table 1) for testing association between a binary variable and a set of 68 explanatory variables (with same within-gene dependence as in gene PDZRN4).

	`minP`	L^2-norm	Hotelling	`HC`	`GHC`	`SKAT`	`Omnibus`	`MGF-R`
Time	0.50	0.49	0.60	0.56	6.01	0.05	1139.10	0.68

procedure detailed in Subsection 3.2 is available in the `R` package `MGFRtest` (https://github.com/fhebert/MGFRTest). The p-values are calculated by a Monte-Carlo procedure based on 1,000 random permutations as detailed in Subsection 3.2. Table 2 gives the average computation times of these global testing methods over 100 simulated datasets for which the dependence pattern reproduces the correlation as estimated using the WTCCC dataset [26] across the 68 genetic markers within gene PDZRN4. Obviously, the computation time of the `MGF-R` procedure is not markedly larger than with the most trivial global testing methods, such as `minP` or the L^2-norm test and much smaller than the `GHC` and Omnibus test.

Type-I Error Rate Control. To study to which extent the `MGF-R` procedure controls the type-I error rate, two types of situations regarding the dependence across explanatory variables are considered, either based on parametric models for correlation or on more complex correlation matrices as observed within blocks of genetic markers in the WTCCC dataset [26]. The parametric models considered hereafter are limited to equicorrelation, where the correlation between any two explanatory variables X_j and X_k is ρ, and autocorrelation, where the correlation between two explanatory variables X_j and X_k is $\rho^{|k-j|}$. For each dependence scenario, a value of the binary response variable is randomly assigned to each profile of explanatory variables.

The number p of explanatory variables is set to 20, 50 and 100 and ρ to 0.2, 0.5 and 0.8. The marginal distributions are obtained assuming that genetic markers are in Hardy-Weinberg equilibrium with Minor Allele Frequency set to 0.4. The p-value for the global test of an association between the p-profile of explanatory variales and the response is obtained by implementing the `MGF-R` procedure as detailed in Subsection 3.2 on 1,000 simulated datasets. The empirical type I error rate is then estimated by the proportions of p-values smaller than the nominal level α over the 1,000 simulations, for different values of α.

Table S1 in the supplementary material reproduces the results for each combination of a parametric dependence model (equicorrelation and autocorrelation), a level of dependence across explanatory variables (ρ), a number p of explanatory variables and a nominal level α for the test. It confirms that in these situations the empirical type-I error rates are close to the level α.

For the simulations conducted under data-driven dependence patterns, three blocks of SNPs corresponding to genes `PDZRN4` (68 SNPs), `DTD1` (49 SNPs) and `KCNN3` (37 SNPs) are extracted from the WTCCC dataset [26] and used to estimate correlations across genetic markers. Figure S1 of the supplementary mate-

Table 3: Empirical type I error rates of the MGF-R test for different values of the nominal level α and for three within-gene correlation matrices estimated using the WTCCC dataset [26].

α	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

rial displays image plots of the corresponding within-gene correlation matrices, showing 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 of the MGF-R test, calculated as above by proportions of positive tests over 1,000 simulations, are given in Table 3 for different values of the nominal level α. Obviously, under these dependence structures also, the MGF-R test seems to properly control the type I error rate.

Power studies. The ability of the MGF-R test to detect an association signal is now compared to the detection performance of the global testing methods introduced in Subsection 2.2 and summarized in Table 1. As above for the control of the type-I error rate, the scope of dependence patterns considered in the simulation study includes both parametric correlation models and correlation matrices within blocks of genetic markers estimated using the WTCCC dataset [26].

The patterns of dependence within blocks of genetic markers are those introduced above for the three genes PDZRN4, DTD1 and KCNN3 (see also Figure S1 in supplementary material). As in the short simulation study in Subsection 2.1, the association signal between the binary response variable and the p-profile of explanatory variables is set by fixing the regression coefficients $\boldsymbol{\beta}$ in the logistic regression model (1) without additional covariates $\boldsymbol{U}$. In the two scenarios for an association signal considered for each gene, a vector $\boldsymbol{\beta}_{\max}$ with only two or three non-zero regression coefficients, given above each plot in Figure 4, is chosen so that the detection rate of at least one of the global testing methods is close to one. A sequence of regression coefficients $\boldsymbol{\beta} = \xi\boldsymbol{\beta}_{\mathbf{max}}$ with $\xi \in \{0.1, 0.2, \ldots, 0.9, 1\}$ is considered to define intermediate situations in terms of association signal strength.

For each gene and two scenarios of association signals, 1,000 datasets are simulated and the p-value for the test of $\boldsymbol{\beta} = \mathbf{0}$ is calculated by all the global testing methods introduced above. The corresponding proportions of p-values smaller than 0.05 over the 1,000 simulations are displayed in Figure 4. As previously mentioned, the first striking result is the strong inconsistency of the ranking of global testing methods for a given gene, or equivalently for a given dependence pattern. Interestingly, the MGF-R test is the only method with an acceptable power in all scenarios, close to the most powerful method in scenarios (a), (c), (e) and even the best method in scenarios (b), (d) and (f). In all scenarios, either the L^2-norm test or the Hotelling's t-square test

has high power thus confirming that the class $\mathcal{T}_2$ introduced in Section 3 covers an interesting scope of test statistics. All the other methods, including the omnibus test, have low power in at least one scenario. It is worth noticing that `minP` and `GHC` have close performance 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.

Two Toeplitz-type correlation matrices are used to generate weaker and more regular parametric dependence patterns than those observed within blocks of SNPs:

- a one-band correlation matrix, for which all correlations are zero except on the first off-diagonal band where the correlation between X_j and X_k is 0.5 if $|k-j|=1$,
- and a three-band correlation matrix, for which all correlations are zero except on the first three off-diagonal bands where the correlation between X_j and X_k is 0.75 if $|k-j|=1$, 0.5 if $|k-j|=2$ and 0.25 if $|k-j|=3$.

The association signal between the binary response variable and the p-profile of explanatory variables, with $p=100$ or $p=200$, is set by fixing the regression coefficients $\boldsymbol{\beta}$ in the logistic regression model (1) without additional covariates $\boldsymbol{U}$. For $p=100$ (resp. 200), only the explanatory variables located at positions 22, 37 and 74 (resp. 50, 100 and 150) take nonzero and equal values. As previously explained in Subsection 2.2, the former nonzero coefficients are first set to a maximal value for which at least one of the global testing methods shows a detection rate close to 1, which defines a reference vector $\boldsymbol{\beta}_{\max}$ of regression coefficients. Intermediate vectors $\boldsymbol{\beta}=\xi\boldsymbol{\beta}_{\max}$ of regression coefficients defining a continuum of association signal strengths are generated by multiplying $\boldsymbol{\beta}_{\max}$ by a reduction factor $\xi \in \{0.1, 0.2, \ldots, 0.9, 1\}$.

For each number p of explanatory variables, dependence pattern and vector of regression coefficients $\boldsymbol{\beta}$, 1,000 datasets are simulated and the p-value for the test of $\boldsymbol{\beta}=\mathbf{0}$ is calculated by all the global testing methods introduced above. The proportions of simulations for which the p-value is smaller than 0.05 are displayed in Figure 5 along the signal strength. First, it can be deduced that, due to the strong regularity of the correlation models, the ranking of global testing methods is not markedly different between the two dependence patterns, dominated by the two higher criticism methods and the max-based method `minP`. The sum-of-squared-based testing methods L^2-norm and `MGF-R` seem to be slightly less powerful, `SKAT` and the Hotelling's t-square test being markedly underpowered.

4.2. Application to real datasets

Using benchmark genetic datasets with corresponding binary phenotype, the `MGF-R` test is now used to test the association between the response variable and biologically validated genes. Since the `MGF-R` procedure involves the estimation of pairwise correlations between explanatory variables, as the other global testing methods handling dependence across pointwise test statistics, it

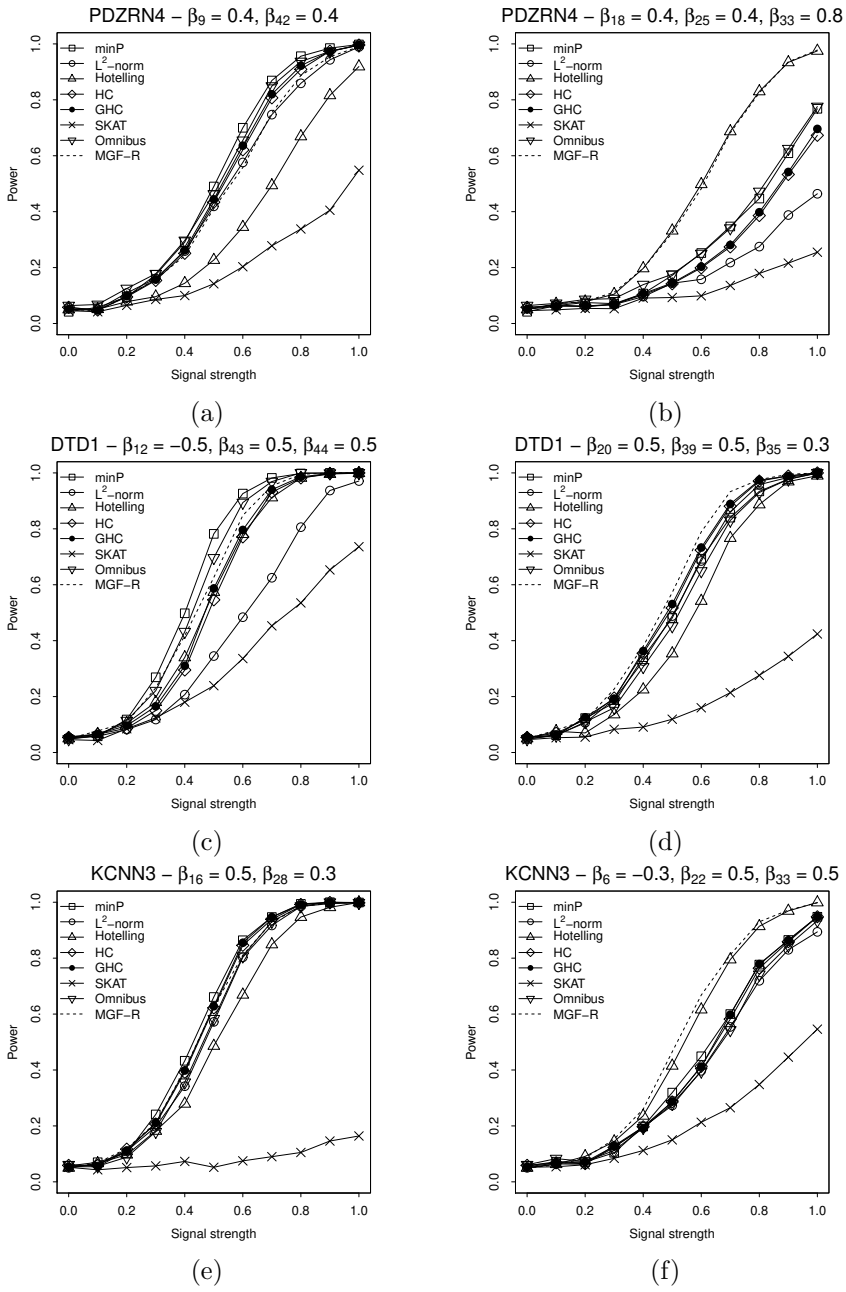

Figure 4: Detection rates along signal strength for all the global testing methods in the comparative study. The six simulation scenarios combine three genes (dependence patterns) and two vectors of regression coefficients $\boldsymbol{\beta}$.

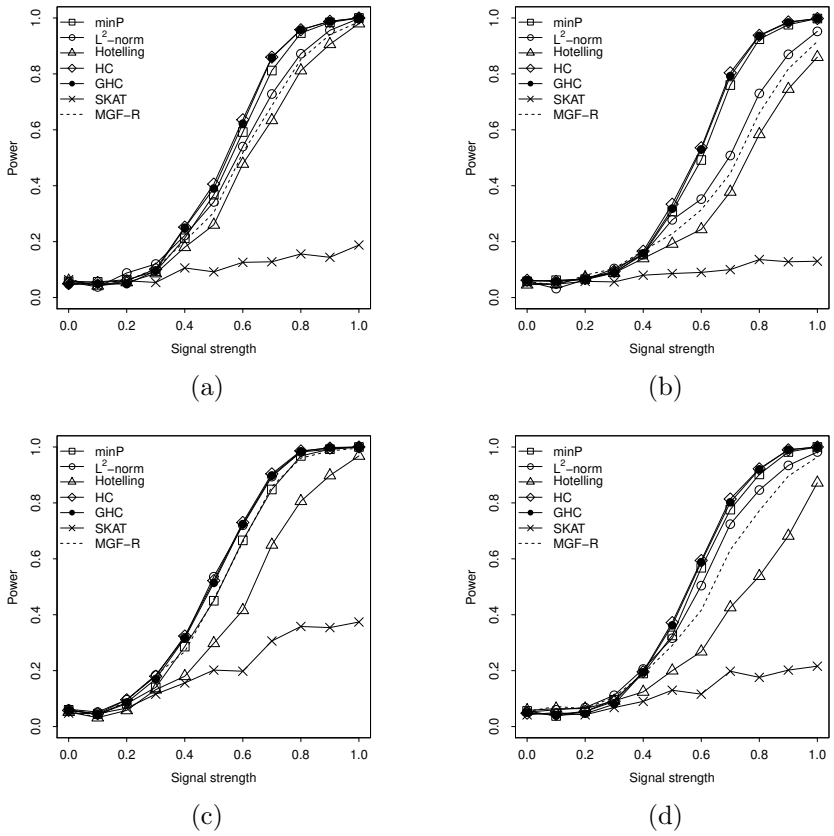

Figure 5: Detection rates along signal strength for all the global testing methods in the comparative study. The four simulation scenarios combine two Toeplitz-type correlation patterns and two numbers p of explanatory variables. (a): one-band structure, $p = 100$; (b): one-band structure, $p = 200$; (c): three-band structure, $p = 100$; (d): three-band structure, $p = 200$.

can be expected that it would suffer from instability in small sample designs. This explains why the first illustrative genetic association studies conducted on a very small number of dogs has been chosen. For each dataset, the `MGF-R` procedure is compared to the alternative methods introduced in Subsection 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.

Dogs Dataset. The first dataset consists of 28 dogs for which 5615 genes were sequenced to investigate the genetic background of the furnishing phenotype that is a characteristic pattern of a moustache and eyebrows [7]. 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` [5]. The p-values are calculated by a Monte-Carlo method with 1,000,000 random permutations for each gene. Due to extremely large computational time, the results for the omnibus test could not be computed.

According to [7], gene `RSPO2` in chromosome 13 (119 SNPs) is strongly associated to the furnishing phenotype. The gene `RSPO2` is correctly detected by the `MGF-R` test and the L^2-norm test, each having a p-value approximately equal to 0.0056 after a Bonferroni correction. However `minP`, HC, `GHC`, `SKAT`, Omnibus and Hotelling's tests fail at detecting `RSPO2` with corrected p-values equal to 0.48, 0.44, 1, 1, 1 and 1 respectively. In Table 4, the number of genes detected by each testing method and the number of genes commonly detected by each pair of methods are given. It can be seen that the number of detected associations is much larger for the L^2-norm test (3,268 significant genes) and the `MGF-R` test (3,077 significant genes) than the other tests. Hotelling's test only detects one gene as significant.

Table 4: Number of genes detected by each testing method and number of genes commonly detected by each pair of testing methods, for a nominal level 0.05 after a Bonferroni correction of the p-values.

	L^2-norm	`minP`	`HC`	`GHC`	`Hotelling`	`SKAT`	`MGF-R`
L^2-norm	3268	349	430	374	1	416	2902
`minP`	-	403	285	270	0	54	357
`HC`	-	-	468	398	0	103	433
`GHC`	-	-	-	410	0	85	375
`Hotelling`	-	-	-	-	1	0	1
`SKAT`	-	-	-	-	-	442	402
`MGF-R`	-	-	-	-	-	-	3077

Crohn's Disease Dataset. The second application to a genetic association testing issue 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 [11], a list of genes known to be involved in Crohn's disease is given. All the global testing methods listed in table 1 and the `MGF-R` procedure are applied to these 73 genes using the WTCCC dataset [26].

In total, eight genes are significantly detected by at least one method: `DENND1B`, `SCAMP3`, `MST1`, `CARD9`, `ZMIZ1`, `C10orf55`, `PTPN2` and `SBNO2`. The corresponding p-values after a Bonferroni correction are given in Table 5 for all global testing methods, except `SKAT` which does not detect any significant genes. The `MGF-R` test is the only method able to detect the eight 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 ability of the `MGF-R` test to detect a wide range of signals. This is also illustrated by the detection of a

Table 5: Bonferroni corrected p-values of genes detected by at least one of the testing methods introduced in the simulation scenarios of Subsection 4.1. `Htlg` stands for the Hotelling's t-square test.

Gene	L^2-norm	minP	HC	GHC	Htlg	Omnibus	MGF-R
DENND1B	**7e-03**	0.062	**0.008**	**0.007**	0.166	**0.014**	**0.022**
SCAMP3	1	0.651	0.633	0.679	**0.019**	0.226	**0.033**
MST1	**0.006**	**0.036**	**0.029**	**0.023**	0.208	**0.007**	**0.020**
CARD9	0.369	0.123	0.191	0.185	0.494	0.387	**0.033**
ZMIZ1	0.289	0.443	0.137	0.1172	**0.000**	0.081	**0.000**
C10orf55	1	1	1	1	**0.008**	1	**0.024**
PTPN2	**0.000**	**0.000**	**0.000**	**0.000**	**0.021**	**0.014**	**0.000**
SBNO2	**0.000**	**0.003**	**0.003**	**0.011**	**0.007**	**0.038**	**0.004**

significant association with gene `CARD9` when using the `MGF-R` test, whereas no other method finds it significant.

5. Discussion

The choice of a global testing procedure has a strong impact on the detection of an association signal in a regression framework especially when the same method is to be used in a wide variety of dependence and association signal patterns. SNP-set testing approaches of genome-wide association studies provide illustrations of such a diversity of situations. Comparative studies of a large panel of global testing procedures based on the aggregation of pointwise test statistics confirm that none of them show uniformly good detection rate over the possible combinations of a dependence pattern across explanatory variables and an association signal.

Based on these observations, a class of global test statistics defined as linear combinations of squared decorrelated pointwise test statistics is introduced to allow for an adaptive handling of dependence. Indeed, special choices of the linear coefficients leads either to a complete whitening of the pointwise test statistics or to ignore their mutual dependence. The closed-form expression of linear coefficients is provided, for which a moment generating function-based distance between the null and non-null distributions of the corresponding global test statistics is maximal. Interestingly, these optimal linear coefficients depend both on the dependence parameters and on the association signal. The `MGF-R` test consists in estimating the optimal linear coefficients to obtain a flexible global test statistics ensuring a reasonable power in most situations.

Simulation studies conducted with correlation structures estimated within blocks of genetic markers or under parametric models of correlation demonstrate the benefits of using `MGF-R` compared to a large panel of methods. A comparative study in the context of two Genome-Wide Association Studies confirms the good performance of `MGF-R`. Indeed, the flexibility of `MGF-R` allows the identification of genes that have been biologically assessed in very diverse situations while the other methods only detect significant associations in specific situations.

These results open new perspectives about the handling of dependence in global testing in a more general framework that would include max-based or higher criticism aggregation tests. Indeed, the simulation studies conducted with more regular correlation models, where the ranking of methods is less affected by the interplay between the patterns of dependence and association signal, reveals that the class of linear combinations of decorrelated pointwise test statistics may be too limited to reach the detection performance of max-based and higher criticism aggregation methods.

Many comparative studies of global testing approaches investigate the impact of the pattern of association signal through the sparsity rate of the vector of regression coefficients. However, the results obtained in Section 3, especially the closed-form expression of the optimal linear coefficients, suggest that the expectation of the decorrelated pointwise tests is more directly related to the performance of global testing procedures. Those results open new leads to establish guidelines for the handling of dependence.

6. Software

The `MGF-R` procedure is available in the `R` package `MGFRtest` (freely downloadable at https://github.com/fhebert/MGFRTest).

7. Supplementary Material

Supplementary material is available online at
https://www.journals.elsevier.com/computational-statistics-and-data-analysis/.

References

[1] Ahdesmäki, M., Strimmer, K., 03 2010. Feature selection in omics prediction problems using CAT scores and False Nondiscovery Rate control. Ann. Appl. Stat. 4 (1), 503–519.

[2] Arias-Castro, E., Candès, E. J., Plan, Y., 2011. Global testing under sparse alternatives: Anova, multiple comparisons and the Higher Criticism. The Annals of Statistics, 2533–2556.

[3] Barbiero, A., Ferrari, P. A., 2015. GenOrd: Simulation of Discrete Random Variables with Given Correlation Matrix and Marginal Distributions. R package version 1.4.0.
URL `https://CRAN.R-project.org/package=GenOrd`

[4] Barnett, I., Mukherjee, R., 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.

[5] Bates, D., Maechler, M., 2018. Matrix: Sparse and Dense Matrix Classes and Methods. R package version 1.2-14.
URL `https://CRAN.R-project.org/package=Matrix`

[6] Bickel, P. J., Levina, E., 12 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.

[7] 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).

[8] Buzkova, P., Lumley, T., Rice, K., 2011. Permutation and parametric bootstrap tests for gene–gene and gene–environment interactions. Annals of human genetics 75 (1), 36–45.

[9] Cai, T., Liu, W., 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.

[10] Conneely, K. N., 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.

[11] 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.

[12] Derkach, A., Lawless, J. F., Sun, L., 2014. Pooled association tests for rare genetic variants: a review and some new results. Statistical Science, 302–321.

[13] Donoho, D., Jin, J., 2004. Higher Criticism for detecting sparse heterogeneous mixtures. Annals of Statistics, 962–994.

[14] Donoho, D., 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.

[15] Epstein, M. P., Duncan, R., Jiang, Y., Conneely, K. N., Allen, A. S., 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.

[16] Hall, P., Jin, J., 2008. Properties of Higher Criticism under strong dependence. The Annals of Statistics 36 (1), 381–402.

[17] Hall, P., Jin, J., 2010. Innovated Higher Criticism for detecting sparse signals in correlated noise. The Annals of Statistics 38 (3), 1686–1732.

[18] Ingster, Y. I., 1997. Some problems of hypothesis testing leading to infinitely divisible distributions. Mathematical Methods of Statistics 6 (1), 47–69.

[19] Lee, S., with contributions from Larisa Miropolsky, Wu, M., 2017. SKAT: SNP-Set (Sequence) Kernel Association Test. R package version 1.3.2.1. URL `https://CRAN.R-project.org/package=SKAT`

[20] 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.

[21] McCullagh, P., Nelder, J., 1989. Generalized Linear Models, Second Edition. Chapman and Hall/CRC Monographs on Statistics and Applied Probability Series. Chapman & Hall.

[22] Ramsay, J., Hooker, G., Graves, S., 2009. Functional Data Analysis with R and MATLAB. Use R! Springer New York. URL `https://books.google.fr/books?id=fNKHa8eV7WYC`

[23] Shen, Q., Faraway, J., 2004. An F test for linear models with functional responses. Statistica Sinica, 1239–1257.

[24] 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.

[25] Vukcevic, D., Hechter, E., Spencer, C., Donnelly, P., 2011. Disease model distortion in association studies. Genetic epidemiology 35 (4), 278–290.

[26] 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.

[27] Westfall, P., Young, S., 1993. Resampling-Based Multiple Testing. New York: Wiley.

[28] Wu, M. C., Lee, S., Cai, T., Li, Y., Boehnke, M., 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.

[29] 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.

[30] Zhang, J.-T., 2013. Analysis of variance for functional data. CRC Press.

[31] Zhang, J.-T., Liang, X., 2014. One-way anova for functional data via globalizing the pointwise F-test. Scandinavian Journal of Statistics 41 (1), 51–71.

[32] Zhao, S. D., Cai, T. T., Cappola, T. P., Margulies, K. B., Li, H., 2017. Sparse simultaneous signal detection for identifying genetically controlled disease genes. Journal of the American Statistical Association 112 (519), 1032–1046.

[33] 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.