Coarse-to-fine multiple testing strategies[*]

Kamel Lahouel, Donald Geman and Laurent Younes

Center for Imaging Science
Johns Hopkins University
3400 N. Charles st.
Baltimore MD 21218, USA
e-mail: klahoue1@jhu.edu
e-mail: geman@jhu.edu
e-mail: laurent.younes@jhu.edu

Abstract: We analyze control of the familywise error rate (FWER) in a multiple testing scenario with a great many null hypotheses about the distribution of a high-dimensional random variable among which only a very small fraction are false, or "active". In order to improve power relative to conservative Bonferroni bounds, we explore a coarse-to-fine procedure adapted to a situation in which tests are partitioned into subsets, or "cells", and active hypotheses tend to cluster within cells. We develop procedures for a non-parametric case based on generalized permutation testing and a linear Gaussian model, and demonstrate higher power than Bonferroni estimates at the same FWER when the active hypotheses do cluster. The main technical difficulty arises from the correlation between the test statistics at the individual and cell levels, which increases the likelihood of a hypothesis being falsely discovered when the cell that contains it is falsely discovered (survivorship bias). This requires sharp estimates of certain quadrant probabilities when a cell is inactive.

MSC 2010 subject classifications: Primary 62G10; secondary 62G09.
Keywords and phrases: Multiple testing, FWER, hierarchical testing, permutation tests.

Received January 2018.

1. Introduction

We consider a multiple testing scenario encountered in many current applications of statistics. Given a large index set V and a family $(H_0(v), v \in V)$ of null hypotheses about the distribution of a high-dimensional random vector $U \in \mathbb{R}^d$, we wish to design a procedure, basically a family of test statistics and thresholds, to estimate the subset $A \subset V$ over which the null hypotheses are false. We shall refer to A as the "active set" and write $\hat{A} = \hat{A}(\mathbf{U})$ for our estimator of A based on a random sample $\mathbf{U}$ of size n from U. The hypotheses in $\hat{A}(\mathbf{U})$ (namely the ones for which the null is rejected) are referred to as "detections" or "discoveries." Naturally, the goal is to maximize the number $|A \cap \hat{A}(\mathbf{U})|$ of detected true positives while simultaneously controlling the number $|A^c \cap \hat{A}(\mathbf{U})|$ of false discoveries. We will assume that $\mathbf{U}$ is defined on a probability space

$(\Omega, \mathbb{P})$, that will, as usual, be assumed large enough to allow for the definition of other random variables used later in our randomization schemes, with the independence assumptions that will be made at that point.

There are two widely used criteria for controlling false positives:

FWER: The family-wise error rate (FWER) is

$$\text{FWER}(\hat{A}) = \mathbb{P}\left(\hat{A}(\mathbf{U}) \cap A^c \neq \emptyset\right),$$

which is the probability of making at least one false discovery. This is usually controlled using Bonferroni bounds and their refinements [11, 14, 13, 12], or using resampling methods or random permutations.

FDR: The false discovery rate (FDR) is the expected ratio between the number of false alarms $|A^c \cap \hat{A}(\mathbf{U})|$ and the number of discoveries $|\hat{A}(\mathbf{U})|$ [4, 6, 5].

In many cases, including the settings in computational biology which directly motivate this work, we find $|A| \ll |V|$, $n \ll d$ as well as small "effect sizes." This is the case, for example, in genome-wide association studies (GWAS) where $U = (Y, X_v, v \in V)$ and the dependence of the "phenotype" Y on the "genotype" $(X_v, v \in V)$ is often assumed to be linear; the active set A are those v with non-zero coefficients and effect size refers to the fraction of the total variance of Y explained by a particular X_v. Under these challenging circumstances, the methods used to guarantee the FWER criterion are usually very conservative and power is limited; that is, number of true positive detections is often very small (if not null) compared to $|A|$ (the "missing heritability"). This is why the less conservative FDR criterion is sometimes preferred: it allows for a higher number of true detections, but of course at the expense of false positives. However, there are situations, such as GWAS, in which this tradeoff is unacceptable; for example, collecting more data and doing follow-up experiments may be too labor intensive or expensive, and therefore having even one false discovery may be deemed undesirable.

To set the stage for our proposal, suppose we are given a family $\{T_v = T_v(\mathbf{U}), v \in V\}$ of test statistics and assume that deviations from the null are captured by small values of $T_v(\mathbf{U})$ (e.g., p-values). Assume that individual rejection regions are of the form $\{u \in \mathcal{U} : T_v(u) \leq \theta\}$ for a constant θ independent of v. Defining $\hat{A}(\mathbf{U}) = \{v : T_v(\mathbf{U}) \leq \theta\}$, the Bonferroni upper-bound is

$$\text{FWER} \leq \sum_{v \in A^c} \mathbb{P}(T_v(\mathbf{U}) \leq \theta) \leq |V| \max_{v \in A^c} \mathbb{P}(T_v(\mathbf{U}) \leq \theta).$$

To ensure that $\text{FWER} \leq \alpha$, $\theta = \theta_B$ is selected such that $\mathbb{P}(T_v(\mathbf{U}) \leq \theta_B) \leq \alpha/|V|$ whenever $v \in A^c$. The Bonferroni bound can only be marginally improved (see, in particular the estimator in [14], which will be referred to as Bonferroni-Holm in the rest of the paper) in the general case. While alternative procedures (including permutation tests) can be designed to take advantage of correlations among tests, the bound is sharp when $|V| \gg |A|$ and tests are independent.

Coarse-to-fine Testing: Clearly some additional assumptions or domain-specific knowledge are necessary to ameliorate the reduction in power resulting from controlling the FWER. Motivated by applications in genomics, we suppose the set V has a natural hierarchical structure. In principle, it should then be possible to gain power if the active hypotheses are not randomly distributed throughout V but rather have a tendency to cluster within cells of the hierarchy. In fact, we shall consider the simplest example consisting of only two levels corresponding to individual hypotheses indexed by $v \in V$ and a partition of V into non-overlapping subsets $(g \subset V, g \in G)$, which we call "cells." We will propose a particular multiple testing strategy which is coarse-to-fine with respect to this structure, controls the FWER, and whose power will exceed that of the standard Bonferroni-Holm approach for typical models and realistic parameters when a minimal degree of clustering is present. It is important to note that the clustering property is not a condition for a correct control of the FWER at a given level using our coarse-to-fine procedure, but only for its increased efficiency in discovering active hypotheses.

Our estimate of A is now based on two families of test statistics: $\{T_v(\mathbf{U}), v \in V\}$, as above, and $\{T_g(\mathbf{U}), g \in G\}$. The cell-level test T_g is designed to assume small values only when g is "active," meaning that $g \cap A \neq \emptyset$. Our estimator of A is now

$$\hat{A}(\mathbf{U}) = \{v : T_g(\mathbf{U}) \leq \theta_G, \, T_v(\mathbf{U}) \leq \theta_V\}.$$

One theoretical challenge of this method is to derive a tractable method for controlling the FWER at a given level α. Evidently, this method can only outperform Bonferroni if $\theta_V > \theta_B$; otherwise, the coarse-to-fine active set is a subset of the Bonferroni discoveries. A key parameter is J, the number of indices belonging to active cells, and in the next section we will derive an FWER bound

$$\mathrm{FWER}(\hat{A}(\mathbf{U})) \leq \Phi(\theta_G, \theta_V, J)$$

under an appropriate compound null hypothesis. While J is not known in general, we can use or estimate an upper bound, $\hat{J}$. The smaller the value of $\hat{J}$ (implying a stronger clustering of active hypotheses in cells) the greater is the gain in power compared with the Bonferroni bound. In particular, as soon as $\hat{J} \ll |V|$, the coarse-to-fine strategy will lead to a considerably less conservative score threshold for individual hypotheses relative to the Bonferroni estimate and the coarse-to-fine procedure will yield an increase in power for a given FWER. Again, our assumptions about clustering are only expressed through an upper bound on J; no other assumptions about the distribution of A are made and the FWER is controlled in all cases.

The main technical difficulty arises from the correlation between the test statistics T_g and T_v for $v \in g$. This must be taken into account since it increases the likelihood of an individual index v being falsely declared active when the cell $g(v)$ that contains it is falsely discovered (survivorship bias). More specifically, we require estimates of quadrant probabilities under the *joint distribution* of $T_{g(v)}(\mathbf{U})$ and $T_v(\mathbf{U})$ when $g(v)$, the cell containing v, is inactive. We will design for this a non-parametric procedure based on generalized permutation

resampling and invariance assumptions (section 3). This will be combined with a method for estimating J, also non-parametric and based on permutation sampling.

We will also analyze the standard linear model with Gaussian data, which is an important parametric model. In this case Φ is expressed in terms of centered chi-square distributions and the power is expressed in terms of non-centered chi-square distributions. The efficiency of the coarse-to-fine method in detecting active hypotheses will depend on effect sizes, both at the level of cells and individual v, among other factors. This model, which has its own interest, will also be used to validate the non-parametric approach. Extensive simulations comparing the power of coarse-to-fine methods and Bonferroni-Holm appear throughout.

Applications and Related Work: As indicated above, our work (and some of our notation) is inspired by statistical issues arising in GWAS [8, 10, 3] and related areas in computational genomics. In the most common version of GWAS, the "genotype" of an individual is represented by the genetic states X_v at a very large family of genomic locations $v \in V$; variations at these locations are called single nucleotide polymorphisms or SNPs. In any given study the objective is to find those SNPs $A \subset V$ "associated" with a given "phenotype", for example a measurable trait Y such as height or blood pressure. The null hypothesis for SNP v is that Y and X_v are independent random variables, and whereas $|V|$ may run into the millions, the set A of active variants is expected to be in the hundreds. (Ideally, one seeks the "causal" variants, an even smaller set, but separating correlation and causality is notoriously difficult.) Control of the FWER is the gold standard and the linear model is common. If the considered variants are confined to coding regions, then the set of genes provides a natural partition of V (and the fact that genes are organized into pathways provides a natural three-level hierarchy) [15].

Another application of large-scale multiple testing is variable filtering in high-dimensional prediction: the objective is to predict a categorical or continuous variable Y based on a family of potentially discriminating features $X_v, v \in V$. Learning a predictor $\hat{Y}$ from i.i.d. samples of $U = (Y, X_v, v \in V)$ is often facilitated by limiting *a priori* the set of features utilized in training $\hat{Y}$ to a subset $A \subset V$ determined by testing the features one-by-one for dependence on Y and setting a significance threshold. In most applications of machine learning to artificial perception, no premium is placed on pruning A to a highly distinguished subset; indeed, the particular set of selected features is rarely examined or considered of significance. In contrast, the identities of the particular features selected and appearing in decision rules are often of keen interest in computational genomics, e.g., discovering cancer biomarkers, where the variables X_v represent "omics" data (e.g., gene expression), and Y codes for two possible cellular or disease phenotypes. Obtaining a "signature" $\hat{A}$ devoid of false positives can be beneficial in understanding the underlying biology and interpreting the decision rules. In this case the Gene Ontology (GO) [2] provides a very rich hierarchical structure, but one example being the organization of genes in pathways. Indeed,

building predictors to separate "driver mutations" from "passenger mutations" in cancer would appear to be a promising candidate for coarse-to-fine testing due to the fact that drivers are known to cluster in pathways.

There is a literature on coarse-to-fine pattern recognition (see, e.g., [7] and the references therein), but the emphasis has traditionally been on computational efficiency rather than error control. Computation is not considered in this paper. Moreover, in most of that work, especially applications to vision and speech, the emphasis is on detecting true positives (e.g., patterns of interest such as faces) at the expense of false positives. Simply "reversing" the role of true positives and negatives is not feasible due to the loss of reasonable invariance assumptions; in effect, every pattern of interest is unique.

Finally, in [16], a hierarchical testing approach is used in the context of the FWER. However, the intention in that paper is to improve the power of detection relative to the Bonferroni-Holm methods only at level of clusters of hypotheses; in contrast to our method, there is limited improvement at the level of individual hypotheses.

Organization of the Paper: The paper is structured as follows: In section 2 we present a Bonferroni-based inequality that will be central for controlling the FWER using the coarse-to-fine method in different models. In section 3 we will propose our main, non-parametric, coarse-to-fine procedure that will control the FWER under general invariance assumptions. In section 4, we then present an estimator for an upper bound on the number of active cells. We then focus on a parametric model (section 5) on which we will derive a parametric version of the coarse-to-fine procedure, specific to the model. In this special case, we are able to obtain power estimates for the coarse-to-fine procedure and optimize some of its free parameters. This estimator will be used, in particular, in our simulations (section 6), in which we will compare its power to that of Bonferroni-Holm, and of our non-parametric coarse-to-fine procedure. Finally, some concluding remarks are made in the discussion.

2. Coarse-to-fine framework

The finite family of null hypotheses will be denoted by $(H_0(v), v \in V)$, where H_0 is either true or false. We are interested in the active set of indices, $A = \{v \in V : H_0(v) = \text{false}\}$ and will write $V_0 = A^c$ for the set of inactive indices. Suppose our data $\mathbf{U}$ takes values in $\mathcal{U}$. The set $\hat{A}(\mathbf{U})$ is commonly designed based on individual rejection regions $\Gamma_v \subset \mathcal{U}$, with $\hat{A}(\mathbf{U}) = \{v : \mathbf{U} \in \Gamma_v\}$. As indicated in the previous section, in the conservative Bonferroni approach, the FWER is controlled at level α by assuming $|V| \max_{v \in V_0} \mathbb{P}(\mathbf{U} \in \Gamma_v) \leq \alpha$. If the rejection regions are designed so that this probability is independent of v whenever $H_0(v)$ is true, then the condition boils down to $\mathbb{P}(\mathbf{U} \in \Gamma_v) \leq \alpha/|V|$ for $v \in V_0$. Generally, $\Gamma_v = \{u \in \mathcal{U} : T_v(u) \leq t\}$ for a constant t (or t_v) for some family of test statistics $(T_v, v \in V)$.

While there is not much to do in the general case to improve on the Bonferroni method, it is possible to improve power if V is structured and one has prior

knowledge about the way the active hypotheses are organized relative to this structure. In this paper, we consider a coarse-to-fine framework in which V is provided with a partition G, so that $V = \bigcup_{g \in G} g$, where the subsets $g \subset V$ (which we will call cells) are non-overlapping. For $v \in V$, we let $g(v)$ denote the unique cell g that contains it. The "coarse" step selects cells likely to contain active indices, followed by a "fine" step in which a Bonferroni or equivalent procedure is applied only to hypotheses included in the selected cells. More explicitly, we will associate a rejection region Γ_g to each $g \in G$ and consider the discovery set

$$\hat{A}(\mathbf{U}) = \{v \in V : \mathbf{U} \in \Gamma_{g(v)} \cap \Gamma_v\}. \tag{1}$$

We will say that a cell g is active if and only if $g \cap A \neq \emptyset$, which we shall also express as $H_0(g) = $ false, implicitly defining $H_0(g)$ as the logical "and" of all $H_0(v), v \in g$. We will also consider the double null hypothesis $H_{00}(v) = H_0(g(v))$ of v belonging in an inactive cell (which obviously implies that v is inactive too), and we will let $V_{00} \subset V_0$ be the set of such v's.

Let $|g|$ denote the size of each cell g in G, G_0 and G_0^c respectively the set of non-active cells and active cells. Then, define $J = \sum_{g \in G_0^c} |g|$, the number of active indices contained in active cells. We will develop our procedure under the assumption that J is known, or, at least bounded from above. While this can actually be a plausible assumption in practice, we will relax it in section 3 in which we will design a procedure to estimate a bound on J.

With this notation, we have the following result:

Proposition 2.1. *With $\hat{A}$ defined by* (1)*:*

$$\mathrm{FWER}(\hat{A}) \leq |V| \max_{v \in V_{00}} \mathbb{P}\left(\mathbf{U} \in \Gamma_{g(v)} \cap \Gamma_v\right) + J \max_{v \in V_0} \mathbb{P}\left(\mathbf{U} \in \Gamma_v\right).$$

Notice that the result will obviously still be valid if we replace J by an upper bound.

Proof. This is just the Bonferroni bound applied to the decomposition

$$
\begin{aligned}
(\hat{A}(\mathbf{U}) \cap V_0 \neq \emptyset) &= \bigcup_{v \in V_{00}} (\mathbf{U} \in \Gamma_{g(v)} \cap \Gamma_v) \cup \bigcup_{v \in V_0 \setminus V_{00}} (\mathbf{U} \in \Gamma_{g(v)} \cap \Gamma_v) \\
&\subset \bigcup_{v \in V_{00}} (\mathbf{U} \in \Gamma_{g(v)} \cap \Gamma_v) \cup \bigcup_{v \in V_0 \setminus V_{00}} (\mathbf{U} \in \Gamma_v)
\end{aligned}
$$

so that

$$P(\hat{A}(\mathbf{U}) \cap V_0 \neq \emptyset) \leq |V_{00}| \max_{v \in V_{00}} \mathbb{P}\left(\mathbf{U} \in \Gamma_{g(v)} \cap \Gamma_v\right) + |V_0 \setminus V_{00}| \max_{v \in V_0} \mathbb{P}(\mathbf{U} \in \Gamma_v)$$

and the proposition results from $|V_{00}| \leq |V|$ and $|V_0 \setminus V_{00}| \leq J$. $\qquad\square$

The sets Γ_g and Γ_v will be designed using statistics $T_g(\mathbf{U})$ and $T_v(\mathbf{U})$ setting $\Gamma_g = (T_g(\mathbf{U}) \leq \theta_G)$ and $\Gamma_v = (T_v(\mathbf{U}) \leq \theta_V)$ for some constants θ_G and θ_V. Letting $p_{00}(\theta_G, \theta_V)$ be an upper-bound of $\mathbb{P}\left((T_{g(v)}(\mathbf{U}) \leq \theta_G) \cap (T_v(\mathbf{U}) \leq \theta_V)\right)$

for $v \in V_{00}$ and $p_0(\theta_V)$ of $\mathbb{P}\left(T_v(\mathbf{U}) \leq \theta_V\right)$ for $v \in V_0$, the previous upper bound becomes

$$\mathrm{FWER}(\hat{A}) \leq |V| \, p_{00}(\theta_G, \theta_V) + J \, p_0(\theta_V). \tag{2}$$

In the following sections our goal will be to design θ_G and θ_V such that this upper bound is smaller than a predetermined level α. Controlling the second term will lead to less conservative choices of the constant θ_V (compared to the Bonferroni estimate), as soon as $J \ll |V|$, depending on the degree of clustering, the probability p_{00} of false detection in the two-step procedure can be made much smaller than p_0 without harming the true detection rate and the coarse-to-fine procedure will yield an increase in power for a given FWER. We require tight estimates of p_{00} and taking into account the correlation between $T_{g(v)}(\mathbf{U})$ and $T_v(\mathbf{U})$ is necessary to deal with "survivorship bias."

3. Non-parametric coarse-to-fine testing

3.1. Notation

Recall that $\mathbf{U}$ denotes the random variable representing all the data, taking values in $\mathcal{U}$. We will build our procedure from user-defined *scores*, denoted ρ_v (at the locus level) and ρ_g (at the cell level), both defined on $\mathcal{U}$, i.e., functions of the observed data.

Moreover, we assume that there exists a group action of some group $\mathfrak{S}$ on $\mathcal{U}$, which will be denoted

$$(\xi, \mathbf{u}) \mapsto \xi \odot \mathbf{u}.$$

The product in $\mathfrak{S}$ will be denoted $(\xi, \xi') \mapsto \xi\xi'$. For example, if the observation is a realization of an i.i.d. family of random variables $U = ((Y^k, X^k), k = 1, \ldots n)$ where the Y's are real-valued and the variables $X^k = (X_v^k, v \in V)$ is a high-dimensional family of variables indexed by the set V, one will take $\mathbf{U} = \left((Y^k, X^k), k = 1, \ldots, n\right)$. $\mathfrak{S}$ will be the permutation group of $\{1, \ldots, n\}$ with

$$\xi \odot \mathbf{U} = \left((Y^{\xi_k}, X^k), k = 1, \ldots, n\right).$$

To simplify the discussion, we will assume that $\mathfrak{S}$ is finite and denote by μ the uniform probability measure on $\mathfrak{S}$, so that

$$\int_{\mathfrak{S}} f(\xi) d\mu(\xi) = \frac{1}{|\mathfrak{S}|} \sum_{\xi \in \mathfrak{S}} f(\xi).$$

Our running assumption will be that,

1. For any $v \in V_{00}$, the joint distribution of $(\rho_{g(v)}((\xi'\xi) \odot \mathbf{U}), \rho_v((\xi'\xi) \odot \mathbf{U}))_{\xi' \in \mathfrak{S}}$ is independent of $\xi \in \mathfrak{S}$.
2. For any $v \in V_0$, the joint distribution of $(\rho_v((\xi'\xi) \odot \mathbf{U}))_{\xi' \in \mathfrak{S}}$ is independent of $\xi \in \mathfrak{S}$.

We will also use the following well-known result.

Lemma 3.1. *Let X be a random variable and let $F_X(x) = P(X \leq x)$ denote its cumulative distribution function, with left limit $F_X^-(x) = P(X < x)$. Define, for $z \in [0, 1]$*

$$\bar{F}_X(x, z) := (1 - z)\left(1 - F_X^-(x)\right) + z\left(1 - F_X(x)\right) = \mathbb{P}(X > x) + (1 - z)\mathbb{P}(X = x).$$

Then, if $Z : \Omega \to [0, 1]$ is uniformly distributed and independent from X, one has, for $t \in [0, 1]$,

$$\mathbb{P}\left(\bar{F}_X(X, Z) \leq t\right) = t.$$

The reader can refer, for example, to [9] for a proof of this lemma in the case of discrete variable X (which will suffice for our purposes).

3.2. *Asymptotic resampling scores*

Let $Z : \Omega \to [0, 1]$ be uniformly distributed and independent of $\mathbf{U}$. We define the asymptotic scores at the cell and variable level by

$$T_g(\mathbf{U}, Z) = \mu\left(\xi : \rho_g(\mathbf{U}) < \rho_g(\xi \odot \mathbf{U})\right) + Z\ \mu\left(\xi : \rho_g(\mathbf{U}) = \rho_g(\xi \odot \mathbf{U})\right) \quad (3)$$

and

$$T_v(\mathbf{U}, Z) = \mu\left(\xi : \rho_v(\mathbf{U}) < \rho_v(\xi \odot \mathbf{U})\right) + Z\ \mu\left(\xi : \rho_v(\mathbf{U}) = \rho_v(\xi \odot \mathbf{U})\right) \quad (4)$$

$T_g(\mathbf{U}, Z)$ and $T_v(\mathbf{U}, Z)$ are the typical statistics used in permutation tests, estimating the proportion of scores that are higher than the observed one after randomizing the sample using the group action, while counting ties with a uniformly distributed weight.

For the coarse-to-fine procedure, we will need one more "conditional" statistic. For a given constant θ_G and a uniform random variable $\tilde{Z}$ independent of $\mathbf{U}$ and Z, we define

$$N_g^{\theta_G}(\mathbf{U}, Z) = \mu\left(\xi : T_g(\xi \odot \mathbf{U}, Z) \leq \theta_G\right). \quad (5)$$

We then let

$$T_v^{\theta_G}(\mathbf{U}, Z, \tilde{Z}) = \frac{1}{N_{g(v)}^{\theta_G}(\mathbf{U}, Z)} \mu\left(\xi : \rho_v(\mathbf{U}) < \rho_v(\xi \odot \mathbf{U}); T_{g(v)}(\xi \odot \mathbf{U}, Z) \leq \theta_G\right)$$

$$+ \frac{\tilde{Z}}{N_{g(v)}^{\theta_G}(\mathbf{U}, Z)} \times \mu\left(\xi : \rho_v(\mathbf{U}) = \rho_v(\xi \odot \mathbf{U}); T_{g(v)}(\xi \odot \mathbf{U}, Z) \leq \theta_G\right). \quad (6)$$

We call our scores asymptotic in this section because exact expectations over μ cannot be computed in general, and can only be obtained as limits of Monte-Carlo samples. The practical finite-sample case will be handled in the next section.

With this notation, we let

$$\hat{A} = \{v : T_{g(v)}(\mathbf{U}, Z) \leq \theta_G \text{ and } T_v^{\theta_G}(\mathbf{U}, Z, \tilde{Z}) \leq \theta_V \text{ and } T_v(\mathbf{U}, Z) \leq \theta_V'\}$$

which depends on the choice of three constants, θ_V, θ_G and θ_V'. We then have:

Theorem 3.1. *For all $v \in V_0$:*

$$\mathbb{P}\left(v \in \hat{A}\right) \leq \theta'_V \tag{7}$$

and for all $v \in V_{00}$,

$$\mathbb{P}\left(v \in \hat{A}\right) \leq \theta_G \theta_V \tag{8}$$

This result tells us how to control the FWER for a two-level permutation test based on any scores in the (generally intractable) case in which we can exactly compute the test statistics, when we declare an index v active if and only if $T_g(\mathbf{U}, Z) \leq \theta_G$ and $T_v^{\theta_G}(\mathbf{U}, Z, \tilde{Z}) \leq \theta_V$ and $T_v(\mathbf{U}, Z) \leq \theta'_V$ (or $\max\left(T_v^{\theta_G}(\mathbf{U}, Z, \tilde{Z}), \frac{\theta_V}{\theta'_V} T_v(\mathbf{U}, Z)\right) \leq \theta_V$ if one wants to use a single v-indexed statistic as considered in section 2).

Proof. For (7), we use a standard argument justifying randomization tests, that we provide here for completeness. If $v \in V_0$, we have

$$\mathbb{P}\left(v \in \hat{A}\right) = \mathbb{P}\left(T_g(\mathbf{U}, Z) \leq \theta_G; T_v^{\theta_G}(\mathbf{U}, Z, \tilde{Z}) \leq \theta_V; T_v(\mathbf{U}, Z) \leq \theta'_V\right)$$
$$\leq \mathbb{P}\left(T_v(\mathbf{U}, Z) \leq \theta'_V\right).$$

From the invariance assumption, we have

$$\begin{aligned}
\mathbb{P}\left(T_v(\mathbf{U}, Z) \leq \theta'_V\right) &= \mathbb{P}\left(T_v(\xi \odot \mathbf{U}, Z) \leq \theta'_V\right) \text{ for all } \xi \in \mathfrak{S} \\
&= \int_{\mathfrak{S}} \mathbb{P}\left(T_v(\xi \odot \mathbf{U}, Z) \leq \theta'_V\right) d\mu(\xi) \\
&= \mathbb{E}\left(\mathbb{E}\left(\mu\left(\xi : T_v(\xi \odot \mathbf{U}, Z) \leq \theta'_V\right) | \mathbf{U}\right)\right)
\end{aligned}$$

It now remains to remark that, for fixed $\mathbf{U}$, $T_v(\xi \odot \mathbf{U}, Z) = \bar{F}_{\zeta_{\mathbf{U}}}(\zeta_{\mathbf{U}}(\xi), Z)$ with $\zeta_{\mathbf{U}}(\xi') = \rho_v(\xi' \odot \mathbf{U})$ for $x' \in \mathfrak{S}$. Therefore, by Lemma 3.1,

$$\mathbb{E}\left(\mu\left(\xi : T_v(\xi \odot \mathbf{U}, Z) \leq \theta'_V\right) | \mathbf{U}\right) = \mathbb{E}\left(\mu\left(\xi : \bar{F}_{\zeta_{\mathbf{U}}}(\zeta_{\mathbf{U}}(\xi), Z) \leq \theta'_V\right) | \mathbf{U}\right) = \theta'_V, \tag{9}$$

which proves (7). Similarly, one has

$$\mathbb{E}(N_g^{\theta_G}(\mathbf{U}, Z) | \mathbf{U}) = \mathbb{P}\left(T_g(\mathbf{U}, Z) \leq \theta_G | \mathbf{U}\right) = \theta_G. \tag{10}$$

Let us now prove (8), assuming $v \in V_{00}$ and letting $g = g(v)$. We write

$$\mathbb{P}\left(v \in \hat{A}\right) \leq \mathbb{P}\left(T_{g(v)}(\mathbf{U}, Z) \leq \theta_G; T_v^{\theta_G}(\mathbf{U}, Z, \tilde{Z}) \leq \theta_V\right).$$

and find an upper bound for the right-hand side of the inequality. Using the invariance assumption, we have

$$\mathbb{P}\left(T_g(\mathbf{U}, Z) \leq \theta_G; T_v^{\theta_G}(\mathbf{U}, Z) \leq \theta_V\right)$$

$$= \int_{\mathfrak{S}} \mathbb{P}\left(T_g(\xi' \odot \mathbf{U}, Z) \le \theta_G; T_v^{\theta_G}(\xi' \odot \mathbf{U}, Z, \tilde{Z}) \le \theta_V\right) d\mu(\xi')$$

$$= \mathbb{E}\left(\mu\left(\xi' : T_g(\xi' \odot \mathbf{U}, Z) \le \theta_G; T_v^{\theta_G}(\xi' \odot \mathbf{U}, Z, \tilde{Z}) \le \theta_V\right)\right)$$

$$= \mathbb{E}\left(\mathbb{E}\left(\mu\left(\xi' : T_g(\xi' \odot \mathbf{U}, Z) \le \theta_G; T_v^{\theta_G}(\xi' \odot \mathbf{U}, Z, \tilde{Z}) \le \theta_V\right) | \mathbf{U}, Z\right)\right)$$

Notice that, since μ is right-invariant, we have $N_g^{\theta_G}(\xi' \odot \mathbf{U}, Z) = N_g^{\theta_G}(\mathbf{U}, Z)$ for all ξ' and

$$T_v^{\theta_G}(\xi' \odot \mathbf{U}, Z, \tilde{Z})$$
$$= \frac{1}{N_g^{\theta_G}(\xi' \odot \mathbf{U}, Z)}\mu(\xi : \rho_v(\xi' \odot \mathbf{U}) < \rho_v((\xi \circ \xi') \odot \mathbf{U})$$
$$; T_g((\xi \circ \xi') \odot \mathbf{U}, Z) \le \theta_G)$$
$$+ \frac{\tilde{Z}}{N_g^{\theta_G}(\xi' \odot \mathbf{U}, Z)}\mu(\xi : \rho_v(\xi' \odot \mathbf{U}) = \rho_v((\xi \circ \xi') \odot \mathbf{U});$$
$$T_g((\xi \circ \xi') \odot \mathbf{U}, Z) \le \theta_G)$$
$$= \frac{1}{N_g^{\theta_G}(\mathbf{U}, Z)}\mu\left(\xi : \rho_v(\xi' \odot \mathbf{U}) < \rho_v(\xi \odot \mathbf{U}); T_g(\xi \odot \mathbf{U}, Z) \le \theta_G\right)$$
$$+ \frac{\tilde{Z}}{N_g^{\theta_G}(\mathbf{U}, Z)}\mu\left(\xi : \rho_v(\xi' \odot \mathbf{U}) = \rho_v(\xi \odot \mathbf{U}); T_g(\xi \odot \mathbf{U}, Z) \le \theta_G\right)$$

Let $\tilde{\mu}$ denote the probability μ conditional to the event $(T_g(\xi \odot \mathbf{U}, Z) \le \theta_G)$ ($\mathbf{U}$, Z and $\tilde{Z}$ being fixed). Then

$$\frac{1}{N_g^{\theta_G}(\mathbf{U}, Z)}\mu\left(T_g(\xi' \odot \mathbf{U}, Z) \le \theta_G; T_v^{\theta_G}(\xi' \odot \mathbf{U}, Z, \tilde{Z}) \le \theta_V\right)$$
$$= \tilde{\mu}\left(\xi' : p(\xi', \tilde{Z}) \le \theta_V\right),$$

where

$$p(\xi', \tilde{Z}) = \tilde{\mu}\left(\xi : \rho_v(\xi \odot \mathbf{U}) > \rho_v(\xi' \odot \mathbf{U})\right) + \tilde{Z}\tilde{\mu}\left(\xi : \rho_v(\xi \odot \mathbf{U}) = \rho_v(\xi' \odot \mathbf{U})\right)$$

Hence, Lemma 3.1 implies that:

$$\mathbb{E}\left(\frac{1}{N_g^{\theta_G}(\mathbf{U}, Z)}\mu\left(\xi' : T_g(\xi' \odot \mathbf{U}, Z) \le \theta_G; T_v^{\theta_G}(\xi' \odot \mathbf{U}, Z, \tilde{Z}) \le \theta_V\right) \middle| \mathbf{U}, Z\right)$$
$$= \theta_V.$$

Hence,

$$\mathbb{P}\left(T_g(\mathbf{U}, Z) \le \theta_G; T_v^{\theta_G}(\mathbf{U}, Z, \tilde{Z}) \le \theta_V\right)$$
$$= \mathbb{E}\left(N_g^{\theta_G}(\mathbf{U}, Z)\theta_V\right) = \theta_V\mathbb{E}\left(N_g^{\theta_G}(\mathbf{U}, Z)\right) = \theta_V\theta_G. \qquad \square$$

Note that Theorem 3.1 is still true if one takes $Z = \tilde{Z} = 1$ in the definition of the test statistics, because the obtained detection set would then be a subset of $\hat{A}$. This would have resulted in a simpler expression in which ties are fully counted, with very little practical loss because the probability of ties in over such permutations is typically minuscule. However, equality in equations such as (9) will be needed in the proof of Theorem 3.2.

As an immediate corollary, we have:

Corollary 3.1.
$$\text{FWER}(\hat{A}) \leq |V|\theta_G\theta_V + J\theta'_V.$$

As mentioned above, this result does not have practical interest because it requires applying all possible permutations to the data. In practice, a random subset of permutations is picked instead, and we develop the related theory in the next section.

3.3. Finite resampling scores

We now replace T_g, $T_v^{\theta_G}$ and T_v with Monte-Carlo estimates and describe how the upper bounds in Theorem 3.1 need to be modified. We therefore introduce an i.i.d. random sample $\boldsymbol{\xi} = (\xi_1, ..., \xi_K) : \Omega \to \mathfrak{S}^K$, where $\xi_k \sim \mu$ and K is a positive integer. We also introduce the empirical measure:

$$\hat{\mu}_{\boldsymbol{\xi}} = \frac{1}{K} \sum_{k=1}^{K} \delta_{\xi_k}.$$

With this notation, we let:

$$\hat{T}_g(\mathbf{U}, \boldsymbol{\xi}) = \hat{\mu}_{\boldsymbol{\xi}} \left(\xi' : \rho_g(\mathbf{U}) \leq \rho_g(\xi' \odot \mathbf{U}) \right),$$
$$\check{T}_g^-(\mathbf{U}, \boldsymbol{\xi}, \xi') = \hat{\mu}_{\boldsymbol{\xi}} \left(\xi' : \rho_g(\xi' \odot \mathbf{U}) < \rho_g(\xi_k \odot \mathbf{U}) \right),$$
$$\hat{T}_v(\mathbf{U}, \boldsymbol{\xi}) = \hat{\mu}_{\boldsymbol{\xi}} \left(\xi' : \rho_v(\mathbf{U}) \leq \rho_v(\xi' \odot \mathbf{U}) \right),$$

and

$$\hat{T}_v^{\theta_G, \varepsilon_G}(\mathbf{U}, \boldsymbol{\xi}) = \frac{1}{\theta_G} \hat{\mu}_{\boldsymbol{\xi}} \left(\xi' : \rho_v(\mathbf{U}) \leq \rho_v(\xi' \odot \mathbf{U}); \check{T}_g^-(\mathbf{U}, \boldsymbol{\xi}, \xi') \leq \theta_G + \varepsilon_G \right).$$

We denote by $G_\beta(x, a, b)$ the c.d.f. of a beta distribution with parameters a and b evaluated at $x \in [0, 1]$, i.e.,

$$G_\beta(x, a, b) = \frac{1}{\beta(a, b)} \int_0^x t^{a-1}(1 - t)^{b-1} dt$$

with $\beta(a, b) = \Gamma(a)\Gamma(b)/\Gamma(a+b)$. We recall that if X is binomial with parameter n and p ($X \sim \text{Bin}(n, p)$) then, for an integer $t \in \{0, \ldots, n\}$

$$P(X \leq t) = G_\beta(1 - p, n - t, t + 1).$$

We can now define

$$\hat{A} = \left\{ v : \hat{T}_{g(v)}(\mathbf{U}, \boldsymbol{\xi}) \le \theta_G - \varepsilon_G \text{ and } \hat{T}_v^{\theta_G, \varepsilon_G}(\mathbf{U}, \boldsymbol{\xi}) \le \theta_V \text{ and } \hat{T}_v(\mathbf{U}, \boldsymbol{\xi}) \le \theta_V' \right\}$$

and state:

Theorem 3.2. *For $v \in V_0$,*

$$\mathbb{P}\left(v \in \hat{A}\right) \le \frac{\lfloor K\theta_V' \rfloor + 1}{K + 1}. \tag{11}$$

and, for $v \in V_{00}$ and $g = g(v)$,

$$\mathbb{P}\left(v \in \hat{A}\right) \le c_K(\theta_G, \varepsilon_G) + \theta_G \theta_V \tag{12}$$

where

$$c_K(\theta_G, \varepsilon_G) = \frac{\lfloor K(\theta_G - \varepsilon_G) \rfloor + 1}{K + 1} G_\beta(1 - \theta_G, K - \lfloor K(\theta_G - \varepsilon_G) \rfloor, \lfloor K(\theta_G - \varepsilon_G) \rfloor + 2) \tag{13}$$

$$- \theta_G G_\beta(1 - \theta_G, K - \lfloor K(\theta_G - \varepsilon_G) \rfloor, \lfloor K(\theta_G - \varepsilon_G) \rfloor + 1)$$
$$+ \theta_G G_\beta\big(\theta_G, \lfloor K(\theta_G + \varepsilon_G) \rfloor, K - \lfloor K(\theta_G + \varepsilon_G) \rfloor\big).$$

Here, $\lfloor x \rfloor$ denotes the integer part of x.

Corollary 3.2. *The FWER using the finite resampling scores is controlled by :*

$$\text{FWER} \le |V| c_K(\theta_G, \varepsilon_G) + |V| \theta_G \theta_V + J \frac{\lfloor K\theta_V' \rfloor + 1}{K + 1}.$$

Theorem 3.2 is proved in the appendix. Neglecting the rounding error in the last term (letting $(\lfloor K\theta_V' \rfloor + 1)(K + 1) \simeq \theta_V'$), this theorem therefore adds the finite-sample correction $c_K(\theta_G, \varepsilon_G)$ to the asymptotic upper bound (theorem 3.1). Figure 1 plots the level curves of the logarithm of this correction as a function of K and ε_G, fixing θ_G to values that are used in our simulations.

4. Estimating the number of indices inside active cells

We now focus on the issue of estimating from observed data the number J of indices belonging to active cells.

4.1. Asymptotic resampling scores

Recall that:

$$J = \sum_{g \in G_0^c} |g|,$$

where G_0 is the set of inactive cells. Our estimation will be made based on cell statistics $(T_g(\mathbf{U}), g \in G)$ under the following assumption. We will assume that T_g takes small values when g is active, so that, for a suitable non conservative threshold t_0, we have $P(T_g(\mathbf{U}) \le t_0) \simeq 1$. To simplify the argument, we will actually make the approximation that:
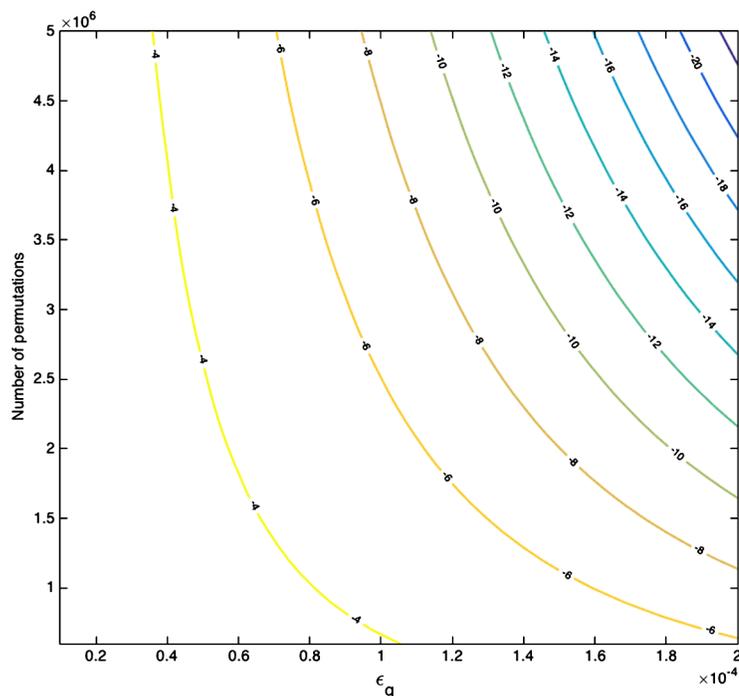
FIG 1. *Level curves of the logarithm of $c_K(\theta_G, \varepsilon_G)$, with $\theta_G = 2.1 \times 10^{-3}$*

[A] There exists $t_0 \in (0,1)$ such that $\mathbb{P}(T_g(\mathbf{U}) \leq t_0) = 1$ if $g \cap A \neq \emptyset$.

Notice that if we denote by:

$$G_0(\mathbf{U}) = \{g \in G : T_g(\mathbf{U}) > t_0\},$$

then assumption [A] implies that $G_0(\mathbf{U}) \subset G_0$. This in turn implies that:

$$N_0(\mathbf{U}) := \sum_{g \in G_0(\mathbf{U})} |g| \leq \sum_{g \in G_0} |g| = |V| - J.$$

Assumption [A] therefore implies an estimator for a lower bound for $|V| - J$ and therefore an upper bound for J, with holds with probability one. However, since the choice of t_0 will not be conservative (typically greater than 0.25), this upper bound will not be sharp enough to be able to take advantage of the clustering assumption. The purpose of this part is to use the set $G_0(\mathbf{U})$ to derive a less obvious and sharper upper bound of J. We start by defining the statistics that will be used to derive the estimator. We will as usual denote our group of transformations by $\mathfrak{S}$, the elements of the group by ξ and the group action by $\odot$. We furthermore define:

- for each $\xi \in \mathfrak{S}$, $N_1(\mathbf{U}, \xi) = \sum_{g \in G_0(\mathbf{U})} \mathbf{1}_{T_g(\xi \odot \mathbf{U}) \leq t_0} |g|$,

- $q_1(\mathbf{U}, \epsilon) = \sup \{n \in \mathbb{N} : \mu\left(\xi : N_1(\mathbf{U}, \xi) \geq n\right) > 1 - \epsilon\}$.

- Finally,

$$\hat{J}(\mathbf{U}, \epsilon) = |V| - (N_0(\mathbf{U}) + q_1(\mathbf{U}, \epsilon)). \tag{14}$$

With this notation, we have the following main result.

Theorem 4.1. *Assuming [A], we have:*

$$\mathbb{P}\left(\hat{J}(\mathbf{U}, \epsilon) < J\right) \leq \epsilon. \tag{15}$$

Proof. First, let us remark that proving (15) is equivalent to proving:

$$\mathbb{P}\left(N_0(\mathbf{U}) + q_1(\mathbf{U}, \epsilon) \geq \sum_{g \in G_0} |g|\right) \leq \epsilon, \tag{16}$$

and that we can write

$$\mathbb{P}\left(N_0(\mathbf{U}) + q_1(\mathbf{U}, \epsilon) \geq \sum_{g \in G_0} |g|\right) = \mathbb{P}\left(q_1(\mathbf{U}, \epsilon) \geq \sum_{g \in G_0} \mathbf{1}_{T_g(\mathbf{U}) \leq t_0} |g|\right).$$

Now define:

$$\tilde{N}_1(\mathbf{U}) = \sum_{g \in G_0} \mathbf{1}_{T_g(\mathbf{U}) \leq t_0} |g|.$$

Since $G_0(\mathbf{U}) \subset G_0$, we have $\tilde{N}_1(\xi \odot \mathbf{U}) \geq N_1(\mathbf{U}, \xi)$ for every $\xi \in \mathfrak{S}$, and $\tilde{q}_1(\mathbf{U}, \epsilon)$ defined as:

$$\tilde{q}_1(\mathbf{U}, \epsilon) = \sup \left\{n \in \mathbb{N} : \mu\left(\xi : \tilde{N}_1(\xi \odot \mathbf{U}) \geq n\right) > 1 - \epsilon\right\}$$

satisfies $\tilde{q}_1(\mathbf{U}, \epsilon) \geq q_1(\mathbf{U}, \epsilon)$ so that

$$\mathbb{P}\left(q_1(\mathbf{U}, \epsilon) \geq \sum_{g \in G_0} \mathbf{1}_{T_g(\mathbf{U}) \leq t_0} |g|\right) \leq \mathbb{P}\left(\tilde{q}_1(\mathbf{U}, \epsilon) \geq \sum_{g \in G_0} \mathbf{1}_{T_g(\mathbf{U}) \leq t_0} |g|\right).$$

By noticing that $\tilde{q}_1(\mathbf{U}, \epsilon) = \tilde{q}_1(\xi \odot \mathbf{U}, \epsilon)$ for every $\xi \in \mathfrak{S}$ by definition of $\tilde{q}_1$, and that the distribution of $\sum_{g \in G_0} \mathbf{1}_{T_g(\mathbf{U}) \leq t_0} |g|$ is invariant under the action of any element $\xi \in \mathfrak{S}$, we have the following:

$$\mathbb{P}\left(\tilde{q}_1(\mathbf{U}, \epsilon) \geq \sum_{g \in G_0} \mathbf{1}_{T_g(\mathbf{U}) \leq t_0} |g|\right) = \mathbb{E}\left(\mu\left(\xi : \tilde{q}_1(\mathbf{U}, \epsilon) \geq \tilde{N}_1(\xi \odot \mathbf{U})\right)\right).$$

But

$$\mu\left(\xi : \tilde{q}_1(\mathbf{U}, \epsilon) \geq \tilde{N}_1(\xi \odot \mathbf{U})\right) \leq \epsilon$$

by definition of $\tilde{q}_1$ and (16) is proved. □

Remark about another version/relaxation of assumption [A]. A relaxation of assumption [A] could be the following:

- $[\tilde{A}]$: Assume that there exists $0 \le r \le 1$ such that:

$$\mathbb{P}\left(|G_0(\boldsymbol{U}) \cap G_0^c| > rJ\right) = 0$$

where $G_0^c = V \setminus G_0$.

The last probability can be non-zero but negligible when compared to the FWER. Assuming $[\tilde{A}]$, one can replace the estimator $\hat{J}(\boldsymbol{U}, \epsilon)$ by the following estimator:

$$\tilde{J}(\boldsymbol{U}, \epsilon) := \frac{1+r}{1-r} \left(|V| - N_0(\boldsymbol{U})\right) - q_1(\boldsymbol{U}, \epsilon),$$

and have the same conclusion as theorem (4.1), namely:

$$\mathbb{P}\left(\tilde{J}(\boldsymbol{U}, \epsilon) < J\right) \le \epsilon.$$

The reason why the theorem is true, assuming $[\tilde{A}]$, is that $N_0(\boldsymbol{U}) = |G_0(\boldsymbol{U}) \cap G_0^c| + |G_0(\boldsymbol{U}) \cap G_0| \le rJ + |G_0(\boldsymbol{U}) \cap G_0|$. Therefore, $|V| - N_0(\boldsymbol{U}) \ge (1-r)J + |G_0^c(\boldsymbol{U}) \cap G_0|$ and $\frac{1+r}{1-r}\left(|V| - N_0(\boldsymbol{U})\right) \ge (1+r)J + |G_0^c(\boldsymbol{U}) \cap G_0|$ with probability 1. Also, we have that $q_1(\boldsymbol{U}, \epsilon) \le \tilde{q}_1(\boldsymbol{U}, \epsilon) + rJ$ with probability 1. As a result:

$$\mathbb{P}\left(\tilde{J}(\boldsymbol{U}, \epsilon) < J\right) \le \mathbb{P}\left(\tilde{q}_1(\boldsymbol{U}, \epsilon) \ge |G_0^c(\boldsymbol{U}) \cap G_0|\right).$$

The proof that the last probability is less than ϵ is now identical to theorem (4.1).

Let us now compare the estimators $\hat{J}(\boldsymbol{U}, \epsilon)$ and $\tilde{J}(\boldsymbol{U}, \epsilon)$. Let us recall that:

$$\hat{J}(\boldsymbol{U}, \epsilon) = \left(|V| - N_0(\boldsymbol{U})\right) - q_1(\boldsymbol{U}, \epsilon).$$

The difference between the two estimators is the first term of the difference, which is $\frac{1+r}{1-r}\left(|V| - N_0(\boldsymbol{U})\right)$ for $\tilde{J}(\boldsymbol{U}, \epsilon)$. On "average", one will get this term for $\hat{J}(\boldsymbol{U}, \epsilon)$ with $\frac{1+r}{1-r}t_0$ instead of t_0 for the assumption $[A]$. One can therefore view $\hat{J}(\boldsymbol{U}, \epsilon)$ as an estimator that is at least as conservative as the one using assumption $[\tilde{A}]$ with $\frac{1-r}{1+r}t_0$ instead of t_0. In our simulations, we used $t_0 = 0.3$ for the derivation of $\hat{J}(\boldsymbol{U}, \epsilon)$. Suppose for example that $r = 0.1$ in assumption $[\tilde{A}]$. This implies that our estimator $\hat{J}(\boldsymbol{U}, \epsilon)$ is at least as conservative as the estimator $\tilde{J}(\boldsymbol{U}, \epsilon)$ with $r = 0.1$ and $t_0 = \frac{0.9}{1.1}0.3 = 0.2454$ for $[\tilde{A}]$.

4.2. Finite resampling scores

We now discuss how the previous estimation of J can be modified when the uniform measure on $\mathfrak{S}$ is approximated by random sampling. Assuming two independent groups of i.i.d. samples of μ, $\boldsymbol{\xi} = (\xi_1, \xi_2, ..., \xi_K)$ and $\boldsymbol{\xi}' = (\xi_1', \xi_2', ...\xi_{K'}')$, and using the notation of section 3.3, we define:

$$\hat{T}_g(\mathbf{U}, \boldsymbol{\xi}) = \hat{\mu}_{\boldsymbol{\xi}}\left(\xi' : \rho_g(\mathbf{U}) \le \rho_g(\xi' \odot \mathbf{U})\right) = \frac{1}{K} \sum_{k=1}^{K} \mathbf{1}_{\rho_g(\mathbf{U}) \le \rho_g(\xi_k \odot \mathbf{U})}$$

and for every $i \in \{1, 2, ..., K'\}$:

$$\check{T}_g(\mathbf{U}, \boldsymbol{\xi}, \xi_i') = \hat{T}_g(\xi' \odot \mathbf{U}, \boldsymbol{\xi} \circ {\xi_i'}^{-1}) = \frac{1}{K} \sum_{k=1}^{K} \mathbf{1}_{\rho_g(\xi_i' \odot \mathbf{U}) \leq \rho_g(\xi_k \odot \mathbf{U})}, \cdot$$

We replace the assumption [A] of the previous section by the assumption [$\hat{A}$]:

[$\hat{A}$] There exists $t_0 \in (0, 1)$ such that $\mathbb{P}(\hat{T}_g(\mathbf{U}, \boldsymbol{\xi}) \leq t_0) = 1$ if $g \cap A \neq \emptyset$.

Notice that is possible to keep the previous assumption [A] and replace t_0 by $t_0 + \epsilon$ in [$\hat{A}$] and the probability 1 by $1 - \exp(-2K\epsilon^2)$, using a Hoeffding bound. We now provide an upper bound $\hat{J}$ of the number of indices belonging to active cells using the finite resampling scores. We will use the following notation.

- Let $\hat{G}_0(\mathbf{U}, \boldsymbol{\xi}) = \{g \in G : \hat{T}_g(\mathbf{U}, \boldsymbol{\xi}) > t_0\}$. (Notice that assumption [$\hat{A}$] implies that $\hat{G}_0(\mathbf{U}, \boldsymbol{\xi}) \subset G_0$).

- Let $\hat{N}_0(\mathbf{U}, \boldsymbol{\xi}) = \sum_{g \in G} |g| \mathbf{1}_{\hat{T}_g(\mathbf{U}, \boldsymbol{\xi}) > t_0}$ and $\hat{N}_1(\mathbf{U}, \boldsymbol{\xi}) = \sum_{g \in G_0} |g| \mathbf{1}_{\hat{T}_g(\mathbf{U}, \boldsymbol{\xi}) \leq t_0}$, so that
$$\hat{N}_0(\mathbf{U}, \boldsymbol{\xi}) + \hat{N}_1(\mathbf{U}, \boldsymbol{\xi}) = |V| - J.$$

- For each $i \in \{1, 2, ..., K'\}$, $\check{N}_1(\mathbf{U}, \boldsymbol{\xi}, \xi_i') = \sum_{g \in \hat{G}_0(\mathbf{U}, \boldsymbol{\xi})} |g| \mathbf{1}_{\check{T}_g(\mathbf{U}, \boldsymbol{\xi}, \xi_i') \leq t_0}$.

- The order statistics of the K' random variables
$$\check{N}_1(\mathbf{U}, \boldsymbol{\xi}, \xi_1'), \check{N}_1(\mathbf{U}, \boldsymbol{\xi}, \xi_2'), ..., \check{N}_1(\mathbf{U}, \boldsymbol{\xi}, \xi_{K'}'),$$
will be denoted by:
$$\check{N}_1(\mathbf{U}, \boldsymbol{\xi}, \xi_{(1)}'), \check{N}_1(\mathbf{U}, \boldsymbol{\xi}, \xi_{(2)}'), ..., \check{N}_1(\mathbf{U}, \boldsymbol{\xi}, \xi_{(K')}').$$
($\check{N}_1(\mathbf{U}, \boldsymbol{\xi}, \xi_{(1)}')$ being the smallest statistic).

- Finally, define
$$\hat{J}(\mathbf{U}, \boldsymbol{\xi}, \boldsymbol{\xi}', p) = |V| - \left(\hat{N}_0(\mathbf{U}, \boldsymbol{\xi}) + \check{N}_1(\mathbf{U}, \boldsymbol{\xi}, \xi_{(p)}')\right).$$

Notice that the computation of $\hat{J}(\mathbf{U}, \boldsymbol{\xi}, \boldsymbol{\xi}', p)$ requires the computation of just $K + K'$ scores.

We have the following result.

Theorem 4.2.
$$\mathbb{P}\left(\hat{J}(\mathbf{U}, \boldsymbol{\xi}, \boldsymbol{\xi}', p) < J\right) \leq \frac{p-1}{K'}.$$

Proof. We first notice that $\hat{J}(\mathbf{U}, \boldsymbol{\xi}, \boldsymbol{\xi}', p) = J + \hat{N}_1(\mathbf{U}, \boldsymbol{\xi}) - \check{N}_1(\mathbf{U}, \boldsymbol{\xi}, \xi_{(p)}')$, so that

$$\mathbb{P}\left(\hat{J}(\mathbf{U}, \boldsymbol{\xi}, \boldsymbol{\xi}', p) < J\right) = \mathbb{P}\left(\check{N}_1(\mathbf{U}, \boldsymbol{\xi}, \xi_{(p)}') > \hat{N}_1(\mathbf{U}, \boldsymbol{\xi})\right)$$
$$\leq \mathbb{P}\left(\check{N}_1(\mathbf{U}, \boldsymbol{\xi}, \xi_{(p)}') > \hat{N}_1(\mathbf{U}, \boldsymbol{\xi})\right)$$

where for each $i \in \{1, 2, ..., K'\}$,

$$\tilde{N}_1(\mathbf{U}, \boldsymbol{\xi}, \xi_i') = \sum_{g \in G_0} |g| \mathbf{1}_{\check{T}_g(\mathbf{U}, \boldsymbol{\xi}, \xi_i') \leq t_0}$$

and $\tilde{N}_1(\mathbf{U}, \boldsymbol{\xi}, \xi_{(p)}')$ is the corresponding pth order statistic. The last inequality holds because, for every j, $\tilde{N}_1(\mathbf{U}, \boldsymbol{\xi}, \xi_j) \geq \check{N}_1(\mathbf{U}, \boldsymbol{\xi}, \xi_j')$, which implies

$$\tilde{N}_1(\mathbf{U}, \boldsymbol{\xi}, \xi_{(p)}') \geq \check{N}_1(\mathbf{U}, \boldsymbol{\xi}, \xi_{(p)}')$$

We then have:

$$\boldsymbol{P}\left(\tilde{N}_1(\mathbf{U}, \boldsymbol{\xi}, \xi_{(p)}') > \hat{N}_1(\mathbf{u}, \boldsymbol{\xi})\right) = \boldsymbol{P}\left(|\{i : \tilde{N}_1(\mathbf{U}, \boldsymbol{\xi}, \xi_i') < \hat{N}_1(\mathbf{U}, \boldsymbol{\xi})\}| < p\right).$$

Notice that given $\boldsymbol{\xi}$ and $\mathbf{U}$, the variable $|\{i : \tilde{N}_1(\mathbf{U}, \boldsymbol{\xi}, \xi_i') < \hat{N}_1(\mathbf{u}, \boldsymbol{\xi})\}|$ follows a Binomial distribution with K' number of trials and a probability of success that is equal to:

$$\mu\left(\xi' : \tilde{N}_1(\mathbf{U}, \boldsymbol{\xi}, \xi') < \hat{N}_1(\mathbf{U}, \boldsymbol{\xi})\right).$$

Using the fact that $\check{T}_g(\mathbf{U}, \boldsymbol{\xi}, \xi') = \check{T}_g(\xi' \odot \mathbf{U}, \boldsymbol{\xi} \circ \xi'^{-1})$, we have:

$$\tilde{N}_1(\mathbf{U}, \boldsymbol{\xi}, \xi') = \tilde{N}_1(\xi' \odot \mathbf{U}, \boldsymbol{\xi} \circ \xi'^{-1}),$$

and

$$\mu\left(\xi' : \tilde{N}_1(\mathbf{U}, \boldsymbol{\xi}, \xi') < \hat{N}_1(\mathbf{U}, \boldsymbol{\xi})\right) = \mu\left(\xi' : \tilde{N}_1(\xi' \odot \mathbf{U}, \boldsymbol{\xi} \circ \xi'^{-1}) < \hat{N}_1(\mathbf{U}, \boldsymbol{\xi})\right).$$

At this point, we notice that $\mathfrak{S}$ acts on $\mathcal{U} \times \mathfrak{S}^K$ via the group action:

$$(\xi', (\mathbf{U}, \boldsymbol{\xi})) \to (\xi' \odot \mathbf{U}, \boldsymbol{\xi} \circ \xi'^{-1}),$$

and this group action leaves invariant the joint distribution of $(\mathbf{U}, \boldsymbol{\xi})$. Therefore, the distribution of $\mu\left(\xi' : \tilde{N}_1(\mathbf{U}, \boldsymbol{\xi}, \xi') < \tilde{N}_1(\mathbf{U}, \boldsymbol{\xi})\right)$ is dominated by the distribution of a uniform random variable on $[0, 1]$ and the proof of the 4.2 follows immediately using the same argument used to prove inequality (19) in Theorem 3.2. □

4.3. Application to the coarse-to-fine algorithm

Subsection 4.2 provided us with an estimator $\hat{J} = \hat{J}_\varepsilon$ in (14) such that $\hat{J} > J$ with probability larger than $1 - \varepsilon$, which implies that

$$\text{FWER}(\hat{A}) \leq |V| p_{00}(\theta_G, \theta_V) + \hat{J} p_0(\theta_V'),$$

with probability $1 - \varepsilon$ at least.

In section 3, we provided a nonparametric coarse-to-fine procedure controlling the FWER by choosing constants θ_G, θ_V and θ'_V controlling the upper-bound at a significance level α. This was done using a deterministic upper-bound of J, but cannot be directly applied with a data-based estimation of J because this would define data-dependent constants, which cannot be plugged into the definition of the set $\hat{A}$ without invalidating our estimation of the FWER. In other terms, if, for a fixed number J', one defines $\hat{A}_{J'}$ to be the discovery set obtained by optimizing θ_G and θ_V subject to $|V| p_{00}(\theta_G, \theta_V) + J' p_0(\theta'_V) \leq \alpha$, our previous results imply that $\mathrm{FWER}(\hat{A}_{J'}) \leq \alpha$ for all $J' \geq J$, but not necessarily that $\mathrm{FWER}(\hat{A}_{\hat{J}}) \leq \alpha + \varepsilon$.

A simple way to address this issue is to replace $\hat{A}_{\hat{J}}$ with

$$\tilde{A} = \bigcap_{J' \leq \hat{J}} \hat{A}_{J'}.$$

Because $\tilde{A} \subset \hat{A}_J$ with probability at least $1 - \varepsilon$, we have

$$\mathrm{FWER}(\tilde{A}) = P(\tilde{A} \cap V_0 \neq \emptyset) \leq P(\hat{A}_J \cap V_0 \neq \emptyset) + \varepsilon = \mathrm{FWER}(\hat{A}_J) + \varepsilon,$$

so that $\tilde{A}$ controls the FWER at level $\alpha + \varepsilon$ as intended.

4.4. Suggested coarse-to-fine procedure

The coarse-to-fine estimator relies on the choices of the constants θ_G, θ_V, θ'_V and ε_G and on the number of simulations, K. They were determined as follows in our experiments, for a control of the FWER at level α.

    i. Fix $\epsilon < \alpha$ and compute $\hat{J}_\epsilon$, the estimated upper bound of J. We took $\epsilon = \frac{\alpha}{10}$ in our experiments.

    ii. Fix a small $\delta > 0$ ($\delta = 10^{-4}$ in our experiments), and select $\theta_V = \theta_G$ and $\theta'_V = |V| \theta_G^2 / \hat{J}_\epsilon$ such that $2|V| \theta_G^2 \leq \alpha - \epsilon - \delta$.

    iii. We choose any K and ε_G such that $|V| c_K(\theta_G, \varepsilon_G) \leq \delta$, for some small $\delta > 0$.

These choices, which have the merit to be simple, albeit non-optimal, were found to perform well in our simulations (see section 6).

5. Model-based analysis

In this section, we propose an alternate coarse-to-fine testing procedure, adapted to a specific regression model. In this framework, it is possible to obtain estimates for the power of the obtained test, and optimize its parameters on this basis. We will use this analysis as a benchmark to compare with the general non-parametric approach provided in the previous sections. We will assume, for simplicity, that all the cells have the same size, so $|g|$ is constant for $g \in G$.

5.1. Regression model

We assume that the observation is a realization of an i.i.d. family of random variables $U = ((Y^k, X^k), k = 1, \ldots n)$ where the Y's are real-valued and $X^k = (X_v^k, v \in V)$ is a high-dimensional family of variables indexed by the set V. We also assume that X_v^k, $v \in V$, are independent and centered Gaussian, with variance σ_v^2, and that

$$Y^k = a_0 + \sum_{v \in A} a_v X_v^k + \psi^k$$

where $\psi^1, \ldots, \psi^n$ are i.i.d. Gaussian with variance σ^2, and a_v, $v \in A$, are unknown real coefficients. We will denote by $\mathbf{Y}$ the vector $(Y^1, \ldots, Y^n)$ and let $\bar{\mathbf{Y}} = \left(\sum_{k=1}^n Y^k / n \right) \mathbf{1}_n$ where $\mathbf{1}_n$ is the vector composed by ones repeated n times. We also let $\mathbf{X}_v = (X_v^1, \ldots, X_v^n)$ and $\boldsymbol{\psi} = (\psi^1, \ldots, \psi^n)$, so that

$$\mathbf{Y} = \sum_{v \in A} a_v \mathbf{X}_v + \boldsymbol{\psi}.$$

Finally, we will denote by $\sigma_{\mathbf{Y}}^2$ the common variance of $Y^1, \ldots, Y^n$ and assume that it is known (or estimated from the observed data).

5.2. Scores

For $v \in V$, we denote by π_v the orthogonal projection on the subspace S_v spanned by the two vectors $\mathbf{X}_v$ and $\mathbf{1}_n$. We will also denote by π_g $(g \in G)$ the orthogonal projection on the subspace S_g spanned by the vectors $\mathbf{X}_v$, $v \in g$, and $\mathbf{1}_n$, and let

$$T_g(U) = \frac{\|\pi_g \mathbf{Y}\|^2 - \|\bar{\mathbf{Y}}\|^2}{\sigma_{\mathbf{Y}}^2},$$

$$T_v(U) = \frac{\|\pi_v \mathbf{Y}\|^2 - \|\bar{\mathbf{Y}}\|^2}{\sigma_{\mathbf{Y}}^2}.$$

(The projections are simply obtained by least-square regression of $\mathbf{Y}$ on $\mathbf{X}_v$, $v \in g$, for π_g and on $\mathbf{X}_v$ for π_v.) We now provide estimates of

$$p_{00}(\theta_G, \theta_V) = \mathbb{P}\left(\frac{\|\pi_g \mathbf{Y}\|^2 - \|\bar{\mathbf{Y}}\|^2}{\sigma_{\mathbf{Y}}^2} > \theta_G; \frac{\|\pi_v \mathbf{Y}\|^2 - \|\bar{\mathbf{Y}}\|^2}{\sigma_{\mathbf{Y}}^2} > \theta_V \right)$$

for $v \in V_{00}$ and $g = g(v)$ and

$$p_0(\theta_V) = \mathbb{P}\left(\frac{\|\pi_v \mathbf{Y}\|^2 - \|\bar{\mathbf{Y}}\|^2}{\sigma_{\mathbf{Y}}^2} > \theta_V \right)$$

for $v \in V_0$. Note that, because we consider residual sums of squares, we here use large values of the scores in the rejection regions (instead of small values in the introduction and other parts of the paper), hopefully without risk of confusion.

Proposition 5.1. *For all θ_G and θ_V and $g \in G_0$:*

$$p_{00}(\theta_G, \theta_V) \leq C(|g|) \exp\left(-\frac{\theta_G}{2}\right) \theta_G^{\frac{|g|}{2}} \left(1 - G_\beta\left(\frac{\theta_V}{\theta_G}, \frac{1}{2}, \frac{|g|+1}{2}\right)\right)$$
$$+ \left(1 - F_1(\theta_G - |g| + 1)\right),$$

where F_1 is the c.d.f. of a chi-squared distribution with one degree of freedom and

$$C(|g|) = \frac{\exp\left(\frac{|g|-1}{2}\right)}{\sqrt{2}(|g|-1)^{\left(\frac{|g|-1}{2}\right)}} \frac{\Gamma(\frac{|g|}{2} + \frac{1}{2})}{\Gamma(\frac{|g|}{2} + 1)}.$$

Moreover

$$p_0(\theta_V) \leq 1 - F_1(\theta_V).$$

Note that the upper-bound for p_{00} is larger than 1 when $\theta_G \leq |g| - 1$, so that this estimate is useful only when $\theta_G > |g| - 1$.

Proof. For $v \in V_{00}$ and $g = g(v)$, we have

$$\sigma_{\mathbf{Y}}^2 = \sum_{v \in A} a_v^2 \sigma_v^2 + \sigma^2 = \sum_{v \in A \cap g^c} a_v^2 \sigma_v^2 + \sigma^2$$

because $A \cap g^c = A$. Consider the conditional probability:

$$\mathbb{P}\left(\frac{\|\pi_g \mathbf{Y}\|^2 - \|\bar{\mathbf{Y}}\|^2}{\sigma_{\mathbf{Y}-g}^2} > \theta_G; \frac{\|\pi_v \mathbf{Y}\|^2 - \|\bar{\mathbf{Y}}\|^2}{\sigma_{\mathbf{Y}-g}^2} > \theta_V \,\bigg|\, (\mathbf{X}_v)_{v \in g}\right).$$

The conditional distribution of $\mathbf{Y}$ given $(\mathbf{X}_v)_{v \in g}$ is Gaussian $\mathcal{N}(0, \sigma_{\mathbf{Y}}^2 I_n)$ (where I_n is the n-dimensional identity matrix). Denote by π_v' the projection on the orthogonal complement of $\mathbf{J}$ in S_v and by π_g' the projection on the orthogonal complement of S_v in S_g, so that

$$\|\pi_g \mathbf{Y}\|^2 - \|\bar{\mathbf{Y}}\|^2 = \|\pi_g' \mathbf{Y}\|^2 + \|\pi_v' \mathbf{Y}\|^2$$

and

$$\|\pi_v \mathbf{Y}\|^2 - \|\bar{\mathbf{Y}}\|^2 = \|\pi_v' \mathbf{Y}\|^2.$$

This implies that:

$$\mathbb{P}\left(\frac{\|\pi_g \mathbf{Y}\|^2 - \|\bar{\mathbf{Y}}\|^2}{\sigma_{\mathbf{Y}}^2} > \theta_G; \frac{\|\pi_v \mathbf{Y}\|^2 - \|\bar{\mathbf{Y}}\|^2}{\sigma_{\mathbf{Y}}^2} > \theta_V \,\bigg|\, (\mathbf{X}_v)_{v \in g}\right) =$$

$$\mathbb{P}\left(\frac{\|\pi_g' \mathbf{Y}\|^2 + \|\pi_v' \mathbf{Y}\|^2}{\sigma_{\mathbf{Y}}^2} > \theta_G; \frac{\|\pi_v' \mathbf{Y}\|^2}{\sigma_{\mathbf{Y}}^2} > \theta_V \,\bigg|\, (\mathbf{X}_v)_{v \in g}\right)$$

At this stage, one can apply Cochran's theorem to $P_g'(\mathbf{Y}/\sigma_{\mathbf{Y}})$ and $P_v'(\mathbf{Y}/\sigma_{\mathbf{Y}})$, which are conditionally independent given $\mathbf{X}_v$, $v \notin G$, to reduce the problem to finding an upper bound for:

$$\mathbb{P}\left(\eta + \zeta \geq \theta_G; \zeta \geq \theta_V\right),$$

where η is $\chi^2(|g| - 1)$ and ζ is $\chi^2(1)$, and the two variables are independent.

Assume that $\theta'_G = \theta_G - |g| + 1 > 0$ and write this probability as

$$\mathbb{P}\left(\eta + \zeta \geq \theta_G, \theta_V \leq \zeta < \theta'_G\right) + \mathbb{P}\left(\eta + \zeta \geq \theta_G, \zeta \geq \theta_V, \zeta \geq \theta'_G\right),$$

which is less than:

$$\mathbb{P}\left(\eta + \zeta \geq \theta_G, \theta_V \leq \zeta < \theta_G\right) + (1 - F_1(\theta'_G)).$$

The first term in the sum can be re-written as

$$\mathbb{E}\left(\mathbb{P}(\eta \geq \theta_G - \zeta | \zeta) \mathbf{1}_{\theta_V \leq \zeta < \theta'_G}\right).$$

We will use the following tail inequality for the c.d.f., F_k, of a $\chi^2(k)$ random variable, stating that

$$1 - F_k(zk) \leq (z \exp(1 - z))^{\frac{k}{2}},$$

for any $z > 1$. We apply this result to $k = |g| - 1$ and $z = \frac{\theta_G - V}{|g| - 1}$ and write (using the fact that the p.d.f. of a $\chi^2(1)$ is $z^{-1/2} e^{z/2}/(\sqrt{2}\Gamma(1/2))$ for $z > 0$)

$$\mathbb{E}\left(\mathbb{P}(\eta \geq \theta_G - \zeta | \zeta) \mathbf{1}_{\theta_V \leq \zeta < \theta'_G}\right)$$

$$\leq \mathbb{E}\left(\left(\frac{\theta_G - \zeta}{|g| - 1} \exp\left(1 - \frac{\theta_G - \zeta}{|g| - 1}\right)\right)^{\frac{|g| - 1}{2}} \mathbf{1}_{\theta_V \leq \zeta < \theta_G}\right)$$

$$= \int_{\theta_V}^{\theta_G} \frac{(\theta_G - z)^{|g|/2 - 1/2}}{\Gamma(1/2)\sqrt{2}(|g| - 1)^{|g|/2 - 1/2}} z^{-1/2} e^{-\frac{tg'}{2}} dz$$

$$\leq \frac{\theta_G^{\frac{|g|}{2}} e^{-\frac{\theta'_G}{2}}}{\Gamma(1/2)\sqrt{2}(|g| - 1)^{|g|/2 - 1/2}} \int_{\theta_V/tg}^{1} (1 - z)^{|g| - 1/2} z^{-1/2} dz$$

$$= C(|g|)\theta_G^{\frac{|g|}{2}} e^{-\frac{\theta_G}{2}} \left(1 - G_\beta\left(\frac{\theta_V}{\theta_G}, \frac{1}{2}, \frac{|g| + 1}{2}\right)\right).$$

The second upper-bound, for $p_0(\theta_V)$, is easily obtained, and left to the reader. $\qquad\square$

This leads us immediately to the following corollary:

Corollary 5.1. *With the thresholds θ_G and θ_V, an upper bound of the FWER is:*

$$\mathrm{FWER}(\hat{A}) \leq |V| C(|g|) \exp\left(-\frac{\theta_G}{2}\right) \theta_G^{\frac{|g|}{2}} \left(1 - G_\beta\left(\frac{\theta_V}{\theta_G}, \frac{1}{2}, \frac{|g| + 1}{2}\right)\right)$$

$$+ |V|\left(1 - F_1(\theta_G - |g| + 1)\right) + J|g|\left(1 - F_1(\theta_V)\right). \quad (17)$$

Figure 2 provides an illustration of the level curves associated to the above FWER upper bound. More precisely, it illustrates the tradeoff between the conservativeness at the cell level and at the individual index level. In the next
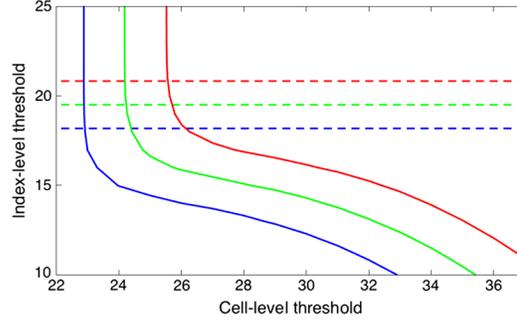
FIG 2. *Level curves of the upper bound of the FWER for the levels 0.2 (blue), 0.1 (green) and 0.05 (red). The horizontal dashed lines represent the thresholds at the individual level for a Bonferroni-Holm test, with corresponding colors. For this figure, $V = 10^4$, $J = 600$ and $g = 10$*

section, the optimization for power will be made along these level lines. Figure 2 also provides the value of the Bonferroni-Holm threshold. For the coarse-to-fine procedure to be less conservative than the Bonferroni-Holm approach, we need the index-level threshold to be smaller, i.e., the optimal point on the level line to be chosen below the corresponding dashed line.

The derivation of (17) is based on the assumption that we have a fixed cell size (across all the cells). If this is not true, it is easy to generalize the previous upper bound. Letting

$$\phi(|g|, \theta_G, \theta_V) = C(|g|) \exp\left(-\frac{\theta_G}{2}\right) \theta_G^{\frac{|g|}{2}} \left(1 - G_\beta\left(\frac{\theta_V}{\theta_G}, \frac{1}{2}, \frac{|g|+1}{2}\right)\right),$$

it suffices to replace $|V|\phi(|g|, \theta_G, \theta_V)$ in (17) with $\sum_{g \in G} |g|\phi(|g|, \sqrt{|g|}\theta_G, \theta_V)$ where θ_G does not depend on the cell g.

5.3. Optimal thresholds

Equation (17) provides a constraint on the pair (θ_G, θ_V) to control the FWER at a given level. We now show how to obtain "optimal" thresholds (θ_G^*, θ_V^*) that maximize the probability of detection subject to this constraint. The discussion will also help understanding how active indices clustering in cells improves the power of the coarse-to-fine procedure.

The conditional distribution of $\mathbf{Y}$ given $(\mathbf{X}_v, v \in g)$ is $\mathcal{N}(\sum_{v \in g \cap A} a_v \mathbf{X}_v, \sigma^2_{\bar{\mathbf{Y}}-g})$ with $\sigma^2_{\bar{\mathbf{Y}}-g} = \sum_{v \in A \cap g^c} a_v^2 \sigma_v^2 + \sigma^2$. It follows from this that, conditionally to these variables, $(\|P_g \mathbf{Y}\|^2 - \|\bar{\mathbf{Y}}\|^2)/\sigma^2_{\bar{\mathbf{Y}}-g}$ follows a non-central chi-square distribution $\chi^2(\rho_g(\mathbf{X}_v, v \in g), |g|)$, with

$$\rho_g(\mathbf{X}_v, v \in g) = \frac{\|\sum_{v \in g \cap A} a_v(\mathbf{X}_v - \bar{X}_v)\|^2}{\sigma^2_{\bar{\mathbf{Y}}-g}}$$

where $\bar{X}_v = \frac{1}{n}\sum_{k=1}^n X_v^k \mathbf{1}_n$. Using the fact that $\rho_g(\mathbf{X}_v, v \in g)/n$ converges to

$$\rho_g := \frac{\sum_{v \in g \cap A} a_v^2 \sigma_v^2}{\sigma_{\mathbf{Y}-g}^2},$$

we will work with the approximation

$$\frac{\|P_g \mathbf{Y}\|^2 - \|\bar{\mathbf{Y}}\|^2}{\sigma_{\mathbf{Y}-g}^2} \sim \chi^2(n\rho_g, |g|).$$

With a similar analysis, and letting for $v \in A$, $\sigma_{\mathbf{Y}-v}^2 = \sum_{v' \in A \setminus v} a_{v'}^2 \sigma_{v'}^2 + \sigma^2$, we will assume that

$$\frac{\|P_v \mathbf{Y}\|^2 - \|\bar{\mathbf{Y}}\|^2}{\sigma_{\mathbf{Y}-v}^2} \sim \chi^2(n\rho_v, 1)$$

with

$$\rho_v := \frac{a_v^2 \sigma_v^2}{\sigma_{\mathbf{Y}-v}^2}.$$

Therefore, an approximation of a lower bound for the probability of detection of an active index v in a cell g will be:

$$\mathbb{P}\left(v \in \hat{A}\right) \geq 1 - F_{|g|}(\theta_G, n\rho_g) - F_1(\theta_V, n\rho_v), \tag{18}$$

where $F_k(x, \delta)$ is the c.d.f of a non-central chi-squared distribution with k degrees of freedom and δ as a non-centrality parameter evaluated at x.

We use the lastest result in the following way. One can fix a target effect size η (the ratio of the effect of $\mathbf{X}_v$ compared to the total variance of $\mathbf{Y}$), and a target cluster size, k, that represents the number of active loci that we expect to find in an active cell, and take $\rho_v = \eta$ and $\rho_g = k\eta$ to optimize the lower-bound in (18) subject to the FWER constraint (17). This provides optimal constants (θ_G, θ_V) for this target case. This is illustrated with numerical simulations in the next section.

6. Simulations and power comparison

In this section, we will first generate simulations under the model of section 5. The purpose of the simulations will be to first show the effect of the coarse-to-fine algorithm on the detection power. It will also illustrate the effect of optimizing the thresholds, assuming a parametric model. Of course, as mentioned in the introduction, the default coarse-to-fine algorithm that should be considered is the non-parametric version of section 3. Therefore, we will compare it with the Bonferroni-Holm approach but also the parametric coarse-to-fine when the data has been generated by the parametric model.

Our second set of experiments uses the software PLINK [17] to simulate case control genome-wide association studies, where the indices will corresponds to

SNPs and compare the (non-parametric) coarse-to-fine approach to the Bonferroni-Holm procedure. In this setting we will also allow the variables X_v to be correlated.

As a main expected observation, all these simulations will illustrate the idea that the more clustered the active indices, the more powerful the coarse-to-fine procedure will be compared to the Bonferroni-Holm procedure.

6.1. Simulations under the parametric model

We let $|V| = 10^4$ with 400 cells of size $|g| = 25$ each. We assume $n = 300$ observations for the model

$$Y^k = a_0 + \sum_{v \in A} a_v X_v^k + \psi^k, k = 1, \ldots, n$$

with $a_0 = 0$ and $a_v = 1$, for all $v \in A$. We also let the X_v's and ψ be i.i.d standard normals. We will control the FWER at level $\alpha = 0.1$.

We will consider two versions of the parametric coarse-to-fine procedure. The first one is a best-case scenario, run under the optimistic assumption that the true values of ρ_g and ρ_v are known in (18). The second is a more realistic, but sub-optimal, procedure in which the sum of the first two terms in (17), and the last term in the same equation are adjusted to both equal $\alpha/2$. Both will be compared to the Bonferroni-Holm procedure.

The first simulation illustrates the effect of optimization over the thresholds on the parametric version of the coarse-to fine algorithm. More precisely, we consider 3 scenarios of a fixed active index contained in a cell having respectively 0,1 and 2 other active indices. For each of these scenarios, we will be interested in the probability of detection of such an index. This is illustrated in figures 3, 4 and 5. Unsurprisingly, the procedure using optimized parameters outperforms the other two but the coarse-to-fine approach using default parameters significantly improves on Bonferroni-Holm when the number of active indices in the cell is more than 1. We also note that the lower bound computed in (18) is most of the time quite close to the true probability of detection.

In the second set of simulations, we consider five scenarios varying the number of active cells and indices, namely (1) 20 active cells with 1 active index each; (2) 10 active cells with 2 active indices; (3) 4 active cells with 5 active indices; (4) 2 active cells with 10 active indices; (5) 1 active cell with 20 active indices. In each case, we ran 100 simulations from which we computed the average number of true detections. The results are provided in table 1 and also include the non-parametric coarse-to-fine method. We found that the fully-informed parametric methods outperforms all others with some margin, the parametric method with default parameters is only slightly better than the non-parametric one. All three outperform Bonferroni-Holm as soon as the number of active indexes in cells is more than 1.

Figure 6 provides the average estimated upper-bound for the number of active cells used in the coarse-to-fine methods. Even is this upper-bound is conserva-
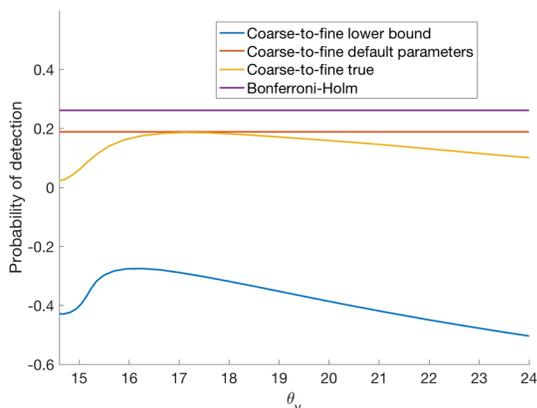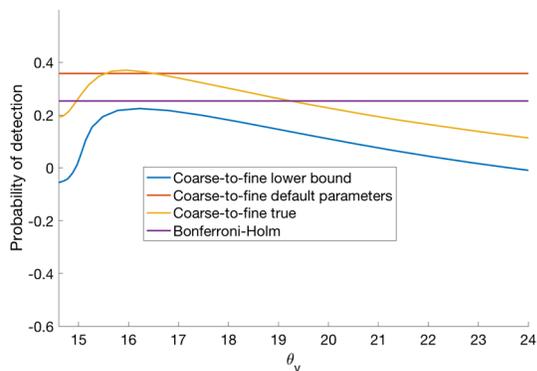
FIG 3. *Probability of detection as a function of θ_V in the admissible space, using the parametric coarse-to-fine procedure for an active cell with 1 active index. The value of θ_G is determined by the implicit equation $FWER(\theta_V, \theta_G) = \alpha$. Coarse-to-fine true represents the estimated true probability of detection via Monte Carlo simulation. Coarse-to-fine lower bound represents the lower bound of the probability of detection obtained via* (18). *We fixed $\hat{J}_{0.01}$ to the value 40×25, which is an upper bound of $\hat{J}$ for all the simulations performed. As expected, the Bonferroni-Holm procedure is better, given that the clustering assumption is not true.*

FIG 4. *Probability of detection as a function of θ_V in the admissible space, using the parametric coarse-to-fine procedure for an active cell with 2 active indices. The CTF procedure outperforms Bonferroni-Holm in this case (even when using the default choice for the thresholds). See Fig.* 3 *for additional details.*

tive and estimate about 20 more cells that their real number, the number of detections is only slightly affected.

6.2. Simulations using the PLINK software

In a second set of simulations, we use the PLINK software to generate case control studies, where the indexes $v \in V$ represent SNPs. The variable X_v takes
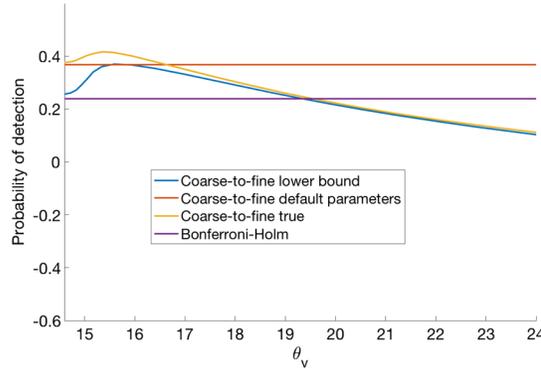
FIG 5. *Probability of detection as a function of θ_V in the admissible space, using the parametric coarse-to-fine procedure for an active cell with 3 active indices. The CTF procedure outperforms Bonferroni-Holm. See Fig. 3 for additional details.*

TABLE 1
Average number of true detections for each of the 4 methods, from left to right: coarse-to-fine using the optimized thresholds, parametric coarse-to-fine using the default thresholds, non parametric coarse-to-fine using default parameters, and Bonferroni-Holm. The total number of active indices is 20 in all cases.

Active indices per active cell	Optimal CTF	Parametric CTF	Nonparametric CTF	BH
1	3.88	3.88	3.85	5.11
2	7.82	7.35	6.09	5.11
4	9.69	8.56	8.13	5.11
5	11.03	8.68	8.15	5.11
20	11.52	9.34	9.12	5.11

ternary values: 0 if both alleles in the SNP are wild-type (the major allele in the population), 1 if one of the alleles is a variant and 2 if both alleles are. The major allele frequency range is $[0.8, 0.95]$. The total number of SNPs is $|V| = 10^4$. The X_v's will either be simulated as independent variables, of with some "linkage disequilibrium" (LD) in which case each SNP is paired is another with a correlation equal to 0.8.

From these SNPs, a binary phenotype Y (cases vs. controls) is generated, yielding $n = 600$ samples, 300 cases ($Y = 1$) and 300 controls ($Y = 0$). The generative model for Y is logistic

$$P(Y = 1 | X_v, v \in V) \propto \exp\left(a_0 + \sum_{v \in A} a_v X_v Y\right),$$

with $a_v = \log 2$ for $v \in A$. This sets the odds ratio for active SNPs is set to 2,
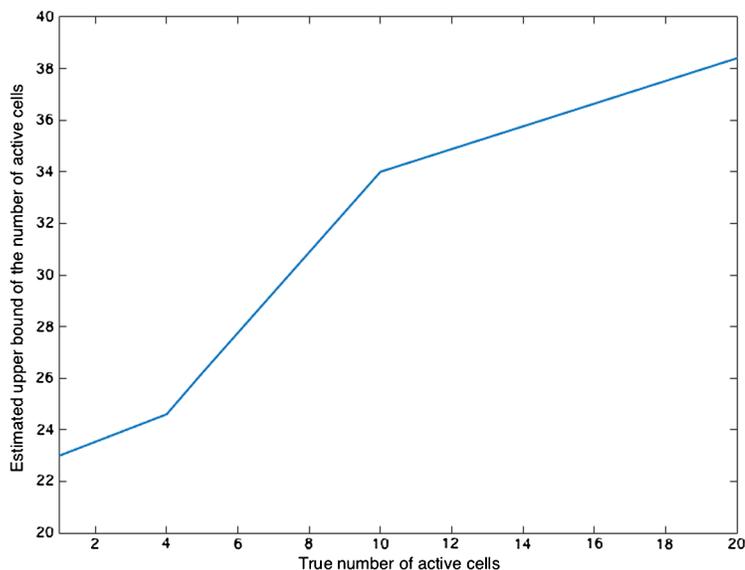
FIG 6. *Plot of the average upper bound of the number of active cells as a function the true number. Even though this upper bound is not particularly tight, it will be sufficient to ensure that coarse-to-fine outperforms the Bonferroni-Holm procedure.*

where

$$\text{odds ratio} = \frac{P(Y=1|X_v=1, X_{v'}, v' \neq v)/P(Y=0|X_v=1, X_{v'}, v' \neq v)}{P(Y=1|X_v=0, X_{v'}, v' \neq v)/P(Y=0|X_v=0, X_{v'}, v' \neq v)} = e^{a_v}.$$

We consider cells (loosely interpreted as "genes") of fixed size, ν_G, with a random assignment of active SNPs to cells based on a variant of a Chinese restaurant process [1]. More precisely, assume that the active indices are $1, 2, ..., |A|$, and denote by $C_1, C_2, \ldots, C_{|A|}$ the random variables representing the cells to which each active index is assigned. The sequence $C_1, \ldots, C_{|A|}$ is generated as follows.

i) $C_1 = 1$
ii) Iterate over $k = 1, 2, \ldots, |A|$ ($|A| = 25$ for all cases). For a given k, let $N_k = \max_k\{C_1, \ldots, C_k\}$ and for $i = 1, \ldots, N_k$, let $n_i = \sum_{j=1}^k \mathbf{1}_{C_j = i}$ be the number of indices assigned to cell i. Then

$$\boldsymbol{P}\left(C_{k+1} = i\right) \propto \begin{cases} \dfrac{\alpha}{k+\alpha} & \text{if } i = N_k + 1 \\ \dfrac{n_i}{k+\alpha} & \text{if } i \leq N_k \text{ and } n_i < \nu_G \\ \quad 0 & \text{otherwise} \end{cases}$$

Here, α is a parameter controlling the clustering of the indices. The smaller α, the more clustered the active indices will be within active cells (see figure 7).
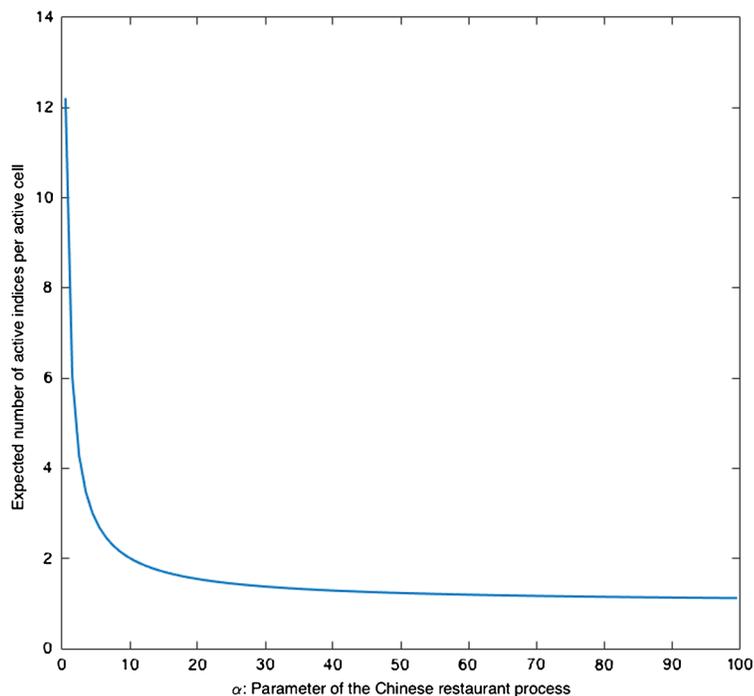
FIG 7. *Expected number of active indices per active cell as a function of α, the clustering parameter of our assignment process. In this case where the size of a cell is greater or equal than the number of active indices, our clustering process corresponds exactly to a Chinese restaurant process.*

We generated datasets with the previous parameters, iterating over $\alpha = 0.5$, 1, 5, 10, 20, 30, 40 and 50 in the Chinese restaurant process and considering four cases: (I) $\nu_G = 10$, no LD; (II) $\nu_G = 25$, no LD; (III) $\nu_G = 10$, LD $= 0.8$; (III) $\nu_G = 25$, LD $= 0.8$. In each case, we took the average over 50 simulations.

Since we are interested in the effect of clustering on the performance of the coarse-to-fine algorithm compared to the Bonferroni-Holm procedure, and not the effect of α itself, we excluded the rare events where we generated a random clustering where the number of active cells decreased after increasing the parameter α. Table 2 illustrates the estimation of the upper bound of the number of active cells for the independent and correlated datasets. Tables 3 compares the performance of the non-parametric coarse-to-fine procedure with the Bonferroni-Holm procedure.

7. Discussion

Given a partition of the space of hypotheses, the basic assumption which allows the coarse-to-fine multiple testing algorithm to obtain greater power than the Bonferroni-Holm approach at the same FWER level is that the distribution of

TABLE 2

Comparison between the true number of indices in active cells (J) and the estimated upper bound ($\hat{J}$) averaged over 50 simulations, as a function of the clustering parameter of the Chinese restaurant process (α), for cell sizes $|g| = 25$ or 10, in the independent and correlated cases.

| | Independent SNPs | | | | Correlated SNPs | | | |
| | $|g| = 25$ | | $|g| = 10$ | | $|g| = 25$ | | $|g| = 10$ | |
α	J	$\hat{J}$	J	$\hat{J}$	J	$\hat{J}$	J	$\hat{J}$
0.5	50	595	30.4	344.2	85	535	34.2	184.6
1	132.5	658	51.2	373.8	128	556.5	50.6	192
5	232	711	94	398.4	218.5	600	89.8	217.2
10	316	754.5	124.4	426.6	298	632.5	119.4	232.8
20	390.5	791	154.4	447.2	376	690	151.2	255.6
30	454	809	178.8	463.2	441	704.5	177.6	271.2
40	502	821.5	192.6	481	500.5	730.5	197	282
50	530.5	830	202.2	482.4	525	750	212.2	285.8

TABLE 3

Comparison between the average number of true detections for the coarse-to-fine (CTF) and Bonferroni-Holm (BH) procedures averaged over 50 simulations, as functions of the clustering parameter of the Chinese restaurant process (α), for cell sizes $|g| = 25$ or 10, in the independent and correlated cases.

| | Independent SNPs | | | | Correlated SNPs | | | |
| | $|g| = 25$ | | $|g| = 10$ | | $|g| = 25$ | | $|g| = 10$ | |
α	CTF	BH	CTF	BH	CTF	BH	CTF	BH
0.5	10.6	5.2	9.4	5.08	10.6	4.76	11.6	3.7
1	9.94	5.2	8.66	5.08	10.14	4.76	10.08	3.7
5	9.18	5.2	7.76	5.08	9.52	4.76	10.02	3.7
10	8.14	5.2	7.36	5.08	8	4.76	9.68	3.7
20	7.22	5.2	7.04	5.08	7.28	4.76	8.24	3.7
30	6.26	5.2	6.88	5.08	6.68	4.76	7.64	3.7
40	4.06	5.2	5.2	5.08	5.06	4.76	5.8	3.7
50	3.8	5.2	5	5.08	4.6	4.76	4.98	3.7

the numbers of active hypotheses across the cells of the partition is non-uniform. The gap in performance is then roughly proportional to the degree of skewness. The test derived for the parametric model can be seen as a generalization to coarse-to-fine testing of the F-test for determining whether a set of coefficients is zero in a regression model; the testing procedure derived for the non-parametric case is a generalization of permutation tests to a multi-level multiple testing.

This scenario was motivated by the situation encountered in genome-wide association studies, where the hypotheses are associated with genetic variations (e.g., SNPs), each having a location along the genome, and the cells are associated with genes. In principle, our coarse-to-fine procedure will then detect more active variants to the extent that these variants cluster in genes. Of course this extent will depend in practice on many factors, including effect sizes, the representation of the genotype (i.e., the choice of variants to explore) as well as the phenotype, and complex interactions within the genotype. It may be very difficult and uncommon to know anything specific about the expected nature of the combinatorics between genes and variants. In some sense, "the proof is in the pudding," in that one can simply try both the standard and coarse-to-fine approaches and compare the sets of variants detected. Given tight control of the FWER, everything found is likely to be real. Indeed, the analytical bounds obtained here make this comparison possible, at least under linear models commonly used in GWAS and in a general non-parametric model under invariance assumptions.

Looking ahead, we have only analyzed the coarse-to-fine approach for the simplest case of two-levels and a true partition, i.e., non-overlapping cells. The methods for controlling the FWER for both the parametric and non-parametric cases generalize naturally to multiple levels assuming nested partitions. The analytical challenge is to generalize the coarse-to-fine approach to overlapping cells, even for two levels: while our methods for controlling the FWER remain valid, they are likely to become overly conservative if cells overlap (however, one could artificially create the partitions by imposing the constraint of assigning at most one cell to an index). This case is of particular interest in applications, where genes are grouped into overlapping "pathways." For example, in Systems Biology, cellular phenotypes, especially complex diseases such as cancer, are studied in the context of these pathways and mutated genes and other abnormalities are in fact known to cluster in pathways; indeed, this is the justification for a pathway-based analysis. Hence the clustering properties may be stronger for variants or genes in pathways than for variants in genes.

Appendix A: Proof of Theorem 3.2

The proof of Theorem 3.2 is based on the introduction of randomized finite sampling scores, allowing us to use lemma 3.1 at multiple occurrences. These randomized scores will be less conservative (but significantly more complex) than the scores that were introduced before Theorem 3.2, which will therefore be obtained as a corollary of the present proof.

Let us first recall our notation and introduce some new one. For a positive integer K, we let μ be the uniform probability on $\mathfrak{S}$ and let $\boldsymbol{\xi} = (\xi_1, ..., \xi_K) \in \mathfrak{S}^K$ where $\xi_1, \ldots, \xi_K$ are independent and have distribution μ. Also, we define $\mathbf{Z} = (Z_1, ..., Z_K)$ and $\tilde{\mathbf{Z}} = (\tilde{Z}_1, ..., \tilde{Z}_K)$ where $Z_1, \ldots, Z_K$ and $\tilde{Z}_1, \ldots, \tilde{Z}_K$ are independent random variables uniformly distributed on $[0, 1]$, that are also independent of $\mathbf{U}$ and $\boldsymbol{\xi}$. We will also need two additional independent uniformly distributed variables, Z and $\tilde{Z}$, also independent of all other variables.

All these variables are assumed to be defined on a probability space $(\Omega, \mathbb{P})$. We also introduce the empirical measures

$$\hat{\mu}_{\boldsymbol{\xi}} = \frac{1}{K} \sum_{k=1}^{K} \delta_{\xi_k}, \quad \hat{\mu}_{\boldsymbol{\xi}, \mathbf{z}} = \frac{1}{K} \sum_{k=1}^{K} \delta_{\xi_k} \delta_{Z_k}, \text{ and } \hat{\mu}_{\boldsymbol{\xi}, \tilde{\mathbf{z}}} = \frac{1}{K} \sum_{k=1}^{K} \delta_{\xi_k} \delta_{\tilde{Z}_k}.$$

With these notations, we let:

$$\hat{T}_{g,r}(\mathbf{U}, \boldsymbol{\xi}, \mathbf{Z}, Z)$$
$$= \hat{\mu}_{\boldsymbol{\xi}} \left(\xi' : \rho_g(\mathbf{U}) < \rho_g(\xi' \odot \mathbf{U}) \right) + \hat{\mu}_{\boldsymbol{\xi}, \mathbf{z}} \left((\xi', z') : \rho_g(\mathbf{U}) = \rho_g(\xi' \odot \mathbf{U}); z' \leq Z \right)$$
$$= \frac{1}{K} \sum_{k=1}^{K} \left(\mathbf{1}_{\rho_g(\mathbf{U}) < \rho_g(\xi_k \odot \mathbf{U})} + \mathbf{1}_{\rho_g(\mathbf{U}) = \rho_g(\xi_k \odot \mathbf{U})} \mathbf{1}_{Z_k \leq Z} \right)$$

$$\check{T}_{g,r}(\mathbf{U}, \boldsymbol{\xi}, \mathbf{Z}_K, \xi', Z) = \hat{T}_g(\xi' \odot \mathbf{U}, \boldsymbol{\xi} \circ {\xi'}^{-1}, \mathbf{Z}, Z)$$
$$= \frac{1}{K} \sum_{k=1}^{K} \left(\mathbf{1}_{\rho_g(\xi' \odot \mathbf{U}) < \rho_g(\xi_k \odot \mathbf{U})} + \mathbf{1}_{\rho_g(\xi' \odot \mathbf{U}) = \rho_g(\xi_k \odot \mathbf{U})} \mathbf{1}_{Z_k \leq Z} \right),$$

$$\hat{T}_{v,r}(\mathbf{U}, \boldsymbol{\xi}, \mathbf{Z}, Z)$$
$$= \hat{\mu}_{\boldsymbol{\xi}} \left(\xi' : \rho_v(\mathbf{U}) < \rho_v(\xi' \odot \mathbf{U}) \right) + \hat{\mu}_{\boldsymbol{\xi}, \mathbf{z}} \left((\xi', z') : \rho_v(\mathbf{U}) = \rho_v(\xi' \odot \mathbf{U}); z' \leq Z \right)$$

and

$$\hat{T}_{v,r}^{\theta_G, \varepsilon_G}(\mathbf{U}, \boldsymbol{\xi}, \mathbf{Z}, \tilde{\mathbf{Z}}, Z, \tilde{Z})$$
$$= \frac{1}{\theta_G} \hat{\mu}_{\boldsymbol{\xi}} \left(\xi' : \rho_v(\mathbf{U}) < \rho_v(\xi' \odot \mathbf{U}); \check{T}_g(\mathbf{U}, \boldsymbol{\xi}, \mathbf{Z}, \xi', Z) \leq \theta_G + \varepsilon_G \right)$$
$$+ \frac{1}{\theta_G} \hat{\mu}_{\boldsymbol{\xi}, \tilde{\mathbf{z}}} \left((\xi', z') : \rho_v(\mathbf{U}) = \rho_v(\xi' \odot \mathbf{U}); z' \leq \tilde{Z}; \check{T}_g(\mathbf{U}, \boldsymbol{\xi}, \mathbf{Z}, \xi', Z) \leq \theta_G + \varepsilon_G \right)$$

We can now define:

$$\hat{A}_r = \left\{ v : \hat{T}_{g(v),r}(\mathbf{U}, \boldsymbol{\xi}, \mathbf{Z}, Z) \leq \theta_G - \varepsilon_G \text{ and } \hat{T}_{v,r}^{\theta_G, \varepsilon_G}(\mathbf{U}, \boldsymbol{\xi}, \mathbf{Z}, Z) \leq \theta_V \right.$$
$$\left. \text{and } \hat{T}_{v,r}(\mathbf{U}, \boldsymbol{\xi}, \mathbf{Z}, \tilde{\mathbf{Z}}, Z, \tilde{Z}) \leq \theta_V' \right\}$$

It is easy to see that $\hat{A} \subset \hat{A}_r$. Therefore, the following result implies Theorem 3.2.

Theorem A.1. *For $v \in V_0$,*

$$\mathbb{P}\left(v \in \hat{A}_r \right) \leq \frac{\lfloor K\theta_V' \rfloor + 1}{K + 1}. \tag{19}$$

and, for $v \in V_{00}$ and $g = g(v)$,

$$\mathbb{P}\left(v \in \hat{A}_r\right) \leq \frac{\lfloor K(\theta_G - \varepsilon_G)\rfloor + 1}{K+1} G_\beta(1-\theta_G, K-\lfloor K(\theta_G-\varepsilon_G)\rfloor, \lfloor K(\theta_G-\varepsilon_G)\rfloor+2)$$
$$- \theta_G G_\beta(1 - \theta_G, K - \lfloor K(\theta_G - \varepsilon_G)\rfloor, \lfloor K(\theta_G - \varepsilon_G)\rfloor + 1)$$
$$+ \theta_G G_\beta\big(\theta_G, \lfloor K(\theta_G + \varepsilon_G)\rfloor, K - \lfloor K(\theta_G + \varepsilon_G)\rfloor + 1\big) + \theta_G \theta_V. \quad (20)$$

Proof of Theorem A.1.
Step 1. We start with (19) which is simpler and standard. Let $v \in V_0$. Conditionally to **U** and Z, $K\hat{T}_{v,r}(\mathbf{U}, \boldsymbol{\xi}, \mathbf{Z}, Z)$ follows a binomial distribution $\text{Bin}(K, T_v(\mathbf{U}, Z))$ (with T_v defined by equation (4)), so that

$$\mathbb{P}(v \in \hat{A}_r) \leq \mathbb{P}(\hat{T}_{v,r}(\mathbf{U}, \boldsymbol{\xi}, \mathbf{Z}, Z) \leq \theta'_V)$$
$$= \mathbb{E}\left(G_\beta(1 - T_v(\mathbf{U}, Z), K - \lfloor K\theta'_V\rfloor, \lfloor K\theta'_V\rfloor + 1)\right).$$

Theorem 3.1 states that $T_v(\mathbf{U}, Z)$ follows a uniform distribution on $[0, 1]$. Therefore,

$$\mathbb{P}(v \in \hat{A}_r) \leq \int_0^1 G_\beta(t, K - \lfloor K\theta'_V\rfloor, \lfloor K\theta'_V\rfloor + 1)dt = \frac{\lfloor K\theta'_V\rfloor + 1}{K+1}.$$

Step 2. We now consider (20) and take $v \in V_{00}$, $g = g(v)$. We first prove that:

$$\mathbb{P}\left(\hat{T}_{g,r}(\mathbf{U}, \boldsymbol{\xi}, \mathbf{Z}, Z) \leq \theta_G - \varepsilon_G; \hat{T}_{v,r}^{\theta_G, \varepsilon_G}(\mathbf{U}, \boldsymbol{\xi}, \mathbf{Z}, \tilde{\mathbf{Z}}, Z, \tilde{Z}) \leq \theta_V\right) \quad (21)$$
$$\leq \mathbb{P}\left(T_g(\mathbf{U}, Z) \leq \theta_G; \hat{T}_{v,r}^{\theta_G, \varepsilon_G}(\mathbf{U}, \boldsymbol{\xi}, \mathbf{Z}, \tilde{\mathbf{Z}}, Z, \tilde{Z}) \leq \theta_V\right)$$
$$+ \frac{\lfloor K(\theta_G - \varepsilon_G)\rfloor + 1}{K+1} G_\beta(1 - \theta_G, K - \lfloor K(\theta_G - \varepsilon_G)\rfloor, \lfloor K(\theta_G - \varepsilon_G)\rfloor + 2)$$
$$- \theta_G G_\beta(1 - \theta_G, K - \lfloor K(\theta_G - \varepsilon_G)\rfloor, \lfloor K(\theta_G - \varepsilon_G)\rfloor + 1)$$

Notice that:

$$\mathbb{P}\left(\hat{T}_{g,r}(\mathbf{U}, \boldsymbol{\xi}, \mathbf{Z}, Z) \leq \theta_G - \varepsilon_G; \hat{T}_{v,r}^{\theta_G, \varepsilon_G}(\mathbf{U}, \boldsymbol{\xi}, \mathbf{Z}, \tilde{\mathbf{Z}}, Z, \tilde{Z}) \leq \theta_V\right) -$$
$$\mathbb{P}\left(T_g(\mathbf{U}, Z) \leq \theta_G; \hat{T}_{v,r}^{\theta_G, \varepsilon_G}(\mathbf{U}, \boldsymbol{\xi}, \mathbf{Z}, \tilde{\mathbf{Z}}, Z, \tilde{Z}) \leq \theta_V\right)$$
$$\leq \mathbb{P}\left(\hat{T}_{g,r}(\mathbf{U}, \boldsymbol{\xi}, \mathbf{Z}, Z) \leq \theta_G - \varepsilon_G; T_g(\mathbf{U}, Z) \geq \theta_G\right)$$

Now, write:

$$\mathbb{P}\left(\hat{T}_{g,r}(\mathbf{U}, \boldsymbol{\xi}, \mathbf{Z}, Z) \leq \theta_G - \varepsilon_G; T_g(\mathbf{U}, Z) \geq \theta_G\right)$$
$$= \mathbb{E}\left(\mathbb{P}\left(\hat{T}_{g,r}(\mathbf{U}, \boldsymbol{\xi}, \mathbf{Z}, Z) \leq \theta_G - \varepsilon_G; T_g(\mathbf{U}, Z) \geq \theta_G \,|\, \mathbf{U}, Z\right)\right).$$

By the same remark used to prove (19), we have that:

$$\mathbb{P}\left(\hat{T}_{g,r}(\mathbf{U},\boldsymbol{\xi},\mathbf{Z},Z) \leq \theta_G - \varepsilon_G; T_g(\mathbf{U},Z) \geq \theta_G \,|\, \mathbf{U},Z\right)$$
$$= \mathbf{1}_{T_g(\mathbf{U},Z)\geq\theta_G} G_\beta(1 - T_g(\mathbf{U},Z), K - \lfloor K(\theta_G - \varepsilon_G)\rfloor, \lfloor K(\theta_G - \varepsilon_G)\rfloor + 1).$$

Taking the expectation over $\mathbf{U}$ and Z, and using the fact that $1 - T_g(\mathbf{U},Z)$ is uniformly distributed over $[0,1]$ under the "double null" hypothesis:

$$\mathbb{P}\left(\hat{T}_{g,r}(\mathbf{U},\boldsymbol{\xi},\mathbf{Z}_K,Z) \leq \theta_G - \varepsilon_G; T_g(\mathbf{U},Z) \geq \theta_G\right)$$
$$= \int_{\theta_G}^1 G_\beta(1 - t, K - \lfloor K(\theta_G - \varepsilon_G)\rfloor, \lfloor K(\theta_G - \varepsilon_G)\rfloor + 1)dt$$
$$= \frac{\lfloor K(\theta_G - \varepsilon_G)\rfloor + 1}{K+1} G_\beta(1 - \theta_G, K - \lfloor K(\theta_G - \varepsilon_G)\rfloor, \lfloor K(\theta_G - \varepsilon_G)\rfloor + 2)$$
$$- \theta_G G_\beta(1 - \theta_G, K - \lfloor K(\theta_G - \varepsilon_G)\rfloor, \lfloor K(\theta_G - \varepsilon_G)\rfloor + 1)$$

Step 3. Let

$$\tilde{T}_{v,r}^{\theta_G}(\mathbf{U},\boldsymbol{\xi},\mathbf{Z},\tilde{\mathbf{Z}},Z,\tilde{Z})$$
$$= \frac{1}{K\theta_G}\sum_{k=1}^K \mathbf{1}_{\rho_v(\mathbf{U})<\rho_v(\xi_k\odot\mathbf{U})}\mathbf{1}_{T_g(\xi_k\odot\mathbf{U},Z)\leq\theta_G}$$
$$+ \frac{1}{K\theta_G}\sum_{k=1}^K \mathbf{1}_{\rho_v(\mathbf{U})=\rho_v(\xi_k\odot\mathbf{U})}\mathbf{1}_{\tilde{Z}_k\leq\tilde{Z}}\mathbf{1}_{T_g(\xi_k\odot\mathbf{U},Z)\leq\theta_G}$$
$$= \frac{1}{K\theta_G}\hat{\mu}_{\boldsymbol{\xi}}\left(\xi' : \rho_v(\xi'\odot\mathbf{U}) > \rho_v(\mathbf{U}); T_g(\xi'\odot\mathbf{U},Z) \leq \theta_G\right)$$
$$+ \frac{1}{K\theta_G}\hat{\mu}_{\boldsymbol{\xi},\tilde{\mathbf{z}}}\left((\xi',z') : \rho_v(\xi'\odot\mathbf{U}) = \rho_v(\mathbf{U}); z' \leq \tilde{Z}\, T_g(\xi'\odot\mathbf{U},Z) \leq \theta_G\right)$$

We now prove that

$$\mathbb{P}\left(T_g(\mathbf{U},Z) \leq \theta_G; \hat{T}_{v,r}^{\theta_G,\varepsilon_G}(\mathbf{U},\boldsymbol{\xi},\mathbf{Z},\tilde{\mathbf{Z}},Z,\tilde{Z}) \leq \theta_V\right)$$
$$\leq \mathbb{P}\left(T_g(\mathbf{U},Z) \leq \theta_G; \tilde{T}_{v,r}^{\theta_G}(\mathbf{U},\boldsymbol{\xi},\mathbf{Z},\tilde{\mathbf{Z}},Z,\tilde{Z}) \leq \theta_V\right)$$
$$+ \theta_G G_\beta(\theta_G, \lfloor K(\theta_G + \varepsilon_G)\rfloor + 1, K+1). \quad (22)$$

Notice that:

$$\mathbb{P}\left(T_g(\mathbf{U},Z) \leq \theta_G; \hat{T}_{v,r}^{\theta_G,\varepsilon_G}(\mathbf{U},\boldsymbol{\xi},\mathbf{Z},\tilde{\mathbf{Z}},Z,\tilde{Z}) \leq \theta_V\right)$$
$$- \mathbb{P}\left(T_g(\mathbf{U},Z) \leq \theta_G; \tilde{T}_{v,r}^{\theta_G}(\mathbf{U},\boldsymbol{\xi},\mathbf{Z},\tilde{\mathbf{Z}},Z,\tilde{Z}) \leq \theta_V\right)$$
$$\leq \mathbb{P}\left(T_g(\mathbf{U},Z) \leq \theta_G; \hat{T}_{v,r}^{\theta_G,\varepsilon_G}(\mathbf{U},\boldsymbol{\xi},\mathbf{Z},\tilde{\mathbf{Z}},Z,\tilde{Z}) \leq \tilde{T}_{v,r}^{\theta_G}(\mathbf{U},\boldsymbol{\xi},\mathbf{Z},\tilde{\mathbf{Z}},Z,\tilde{Z})\right)$$

Conditioning on $\mathbf{U}$ and Z taking the expected value:

$$\mathbb{P}\left(T_g(\mathbf{U},Z) \le \theta_G;\, \hat{T}_{v,r}^{\theta_G,\varepsilon_G}(\mathbf{U},\boldsymbol{\xi},\mathbf{Z},\tilde{\mathbf{Z}},Z,\tilde{Z}) \le \tilde{T}_{v,r}^{\theta_G}(\mathbf{U},\boldsymbol{\xi},\mathbf{Z},\tilde{\mathbf{Z}},Z,\tilde{Z})\right)$$

$$= \mathbb{E}\left(\mathbf{1}_{T_g(\mathbf{U},Z)\le\theta_G}\mathbb{P}\left(\hat{T}_{v,r}^{\theta_G,\varepsilon_G}(\mathbf{U},\boldsymbol{\xi},\mathbf{Z},\tilde{\mathbf{Z}},Z,\tilde{Z}) \le \tilde{T}_{v,r}^{\theta_G}(\mathbf{U},\boldsymbol{\xi},\mathbf{Z},\tilde{\mathbf{Z}},Z,\tilde{Z})\,|\,\mathbf{U},Z\right)\right)$$

We rewrite the last expectation as:

$$\mathbb{E}\left(\mathbf{1}_{T_g(\mathbf{U},Z)\le\theta_G}\mathbb{P}\left(\sum_{k=1}^{K}(\mathbf{1}_{\rho_v(\mathbf{U})<\rho_v(\xi_k\odot\mathbf{U})} + \mathbf{1}_{\rho_v(\mathbf{U})=\rho_v(\xi_k\odot\mathbf{U})}\mathbf{1}_{\tilde{Z}_k\le\tilde{Z}})\right.\right.$$

$$\left.\left.\times\,(\mathbf{1}_{\check{T}_{g,r}(\mathbf{U},\boldsymbol{\xi},\xi_k,\mathbf{Z},Z)\le\theta_G+\varepsilon_G} - \mathbf{1}_{T_g(\xi_k\odot\mathbf{U},Z)\le\theta_G}) \le 0\,\Big|\,\mathbf{U},Z\right)\right).$$

In order that

$$\sum_{k=1}^{K}(\mathbf{1}_{\rho_v(\mathbf{U})<\rho_v(\xi_k\odot\mathbf{U})} + \mathbf{1}_{\rho_v(\mathbf{U})=\rho_v(\xi_k\odot\mathbf{U})}\mathbf{1}_{\tilde{Z}_k\le\tilde{Z}})$$

$$\times\,(\mathbf{1}_{\check{T}_{g,r}(\mathbf{U},\boldsymbol{\xi},\xi_k,\mathbf{Z},Z)\le\theta_G+\varepsilon_G} - \mathbf{1}_{T_g(\xi_k\odot\mathbf{U},Z)\le\theta_G}) \le 0, \quad (23)$$

there must exist $k_0 \in \{1,2,...,K\}$ such that $\check{T}_{g,r}(\mathbf{U},\boldsymbol{\xi},\xi_k,\mathbf{Z},Z) > \theta_G + \varepsilon_G$ and $T_g(\xi_k \odot \mathbf{U},Z) \le \theta_G$. Letting $M_g(\boldsymbol{\xi},\mathbf{U},\mathbf{Z})$ denote the number of indexes k such that $T_g(\xi_k \odot \mathbf{U},Z_k) \le \theta_G$, the existence of such a k_0 implies that $M_g(\boldsymbol{\xi},\mathbf{U},\mathbf{Z}) > K(\theta_G + \varepsilon_G)$. Indeed, we first notice that, for a every $0 \le j,k \le K$:

$$\mathbf{1}_{\rho_g(\xi_k\odot\mathbf{U})<\rho_g(\xi_j\odot\mathbf{U})} + \mathbf{1}_{\rho_g(\xi_k\odot\mathbf{U})=\rho_g(\xi_j\odot\mathbf{U})}\mathbf{1}_{Z_j\le Z} \le \mathbf{1}_{T_g(\xi_j\odot\mathbf{U},Z_j)\le T_g(\xi_k\odot\mathbf{U},Z)}.$$

This statement is obvious when $\rho_g(\xi_k \odot \mathbf{U}) = \rho_g(\xi_j \odot \mathbf{U})$ and $Z_j \le Z$ and can be checked by proving that $T_g(\xi_j \odot \mathbf{U},1) \le T_g(\xi_k \odot \mathbf{U},0)$ when $\rho_g(\xi_k \odot \mathbf{U}) < \rho_g(\xi_j \odot \mathbf{U})$. Therefore, if k_0 exists, we must have

$$\check{T}_{g,r}(\mathbf{U},\boldsymbol{\xi},\xi_{k_0},\mathbf{Z},Z) \le \frac{1}{K}\sum_{j=1}^{K}\mathbf{1}_{T_g(\xi_j\odot\mathbf{U},Z_j)\le T_g(\xi_{k_0}\odot\mathbf{U},Z)}$$

$$\le \frac{1}{K}\sum_{j=1}^{K}\mathbf{1}_{T_g(\xi_j\odot\mathbf{U},Z_j)\le\theta_G}.$$

Because $\mathbb{E}\left(\mu(T_g(\xi \odot \mathbf{U},Z) \le \theta_G)|\mathbf{U}\right) = \theta_G$, we can bound the probability of (23) conditional to $\mathbf{U}$ and Z by the probability that a binomial $\mathrm{Bin}(K,\theta_G)$ is larger than $\lceil K(\theta_G + \varepsilon_G)\rceil$, which is given by

$$G_\beta(\theta_G,\lfloor K(\theta_G+\varepsilon_G)\rfloor + 1, K - \lfloor K(\theta_G+\varepsilon_G)\rfloor).$$

We therefore have

$$\mathbb{P}\left(\hat{T}_{v,r}^{\theta_G,\varepsilon_G}(\mathbf{U},\boldsymbol{\xi},\mathbf{Z},\tilde{\mathbf{Z}},Z,\tilde{Z}) \le \tilde{T}_{v,r}^{\theta_G}(\mathbf{U},\boldsymbol{\xi},\mathbf{Z},\tilde{\mathbf{Z}},Z,\tilde{Z})\,|\,\mathbf{U},Z\right)$$

$$\le G_\beta(\theta_G,\lfloor K(\theta_G+\varepsilon_G)\rfloor + 1, K - \lfloor K(\theta_G+\varepsilon_G)\rfloor).$$

Now, finally notice that

$$\mathbb{P}\left(T_g(\mathbf{U}, Z) \le \theta_G\right) = \theta_G,$$

and we have proved (22).

Step 4. We finally show that

$$\mathbb{P}\left(T_g(\mathbf{U}, Z) \le \theta_G; \tilde{T}_{v,r}^{\theta_G}(\mathbf{U}, \boldsymbol{\xi}, \mathbf{Z}, \tilde{\mathbf{Z}}, Z, \tilde{Z}) \le \theta_V\right) \le \theta_V \theta_G \qquad (24)$$

For this, we note that

$$K\theta_G \tilde{T}_{v,r}^{\theta_G}(\mathbf{U}, \boldsymbol{\xi}, \mathbf{Z}, \tilde{\mathbf{Z}}, Z, \tilde{Z}) = \sum_{k=1}^{K} \mathbf{1}_{\rho_v(\mathbf{U}) < \rho_v(\xi_k \odot \mathbf{U})} \mathbf{1}_{T_g(\xi_k \odot \mathbf{U}, Z) \le \theta_G}$$

$$+ \sum_{k=1}^{K} \mathbf{1}_{\rho_v(\mathbf{U}) = \rho_v(\xi_k \odot \mathbf{U})} \mathbf{1}_{\tilde{Z}_k \le \tilde{Z}} \mathbf{1}_{T_g(\xi_k \odot \mathbf{U}, Z) \le \theta_G}.$$

Conditionally to $\mathbf{U}$, $\mathbf{Z}$ and $\tilde{Z}$, this variable follows a binomial distribution with probability of success $N_g^{\theta_G}(\mathbf{U}, Z) T_v^{\theta_G}(\mathbf{U}, Z, \tilde{Z})$. Therefore:

$$\mathbb{P}\left(T_g(\mathbf{U}, Z) \le \theta_G; \tilde{T}_{v,r}^{\theta_G}(\mathbf{U}, \boldsymbol{\xi}, \mathbf{Z}, \tilde{\mathbf{Z}}, Z, \tilde{Z}) \le \theta_V\right)$$

$$= \mathbb{E}\Big(\mathbf{1}_{T_g(\mathbf{U}, Z) \le \theta_G} G_\beta(1 - N_g^{\theta_G}(\mathbf{U}, Z) T_v^{\theta_G}(\mathbf{U}, Z, \tilde{Z}),$$

$$K - \lfloor K\theta_G \theta_V \rfloor, \lfloor K\theta_G \theta_V \rfloor + 1)\Big)$$

We now use the fact that The distribution of $\mathbf{U}$ is invariant under the action of the group $\mathfrak{S}$ and that $N_g^{\theta_G}(\xi \odot \mathbf{U}, Z) = N_g^{\theta_G}(\mathbf{U}, Z)$ for all $\xi \in \mathfrak{S}$ to write, introducing a new random variable $\bar{\xi}$ independent from the others in the expectation

$$\mathbb{P}\left(T_g(\mathbf{U}, Z) \le \theta_G; \tilde{T}_{v,r}^{\theta_G}(\mathbf{U}, \boldsymbol{\xi}, \mathbf{Z}, \tilde{\mathbf{Z}}, Z, \tilde{Z}) \le \theta_V\right)$$

$$= \mathbb{E}\Big(\mathbf{1}_{T_g(\bar{\xi} \odot \mathbf{U}, Z) \le \theta_G} G_\beta(1 - N_g^{\theta_G}(\mathbf{U}, Z) T_v^{\theta_G}(\bar{\xi} \odot \mathbf{U}, Z, \tilde{Z}),$$

$$K - \lfloor K\theta_G \theta_V \rfloor, \lfloor K\theta_G \theta_V \rfloor + 1)\Big)$$

Now, using lemma 3.1, we notice that, given $\mathbf{U}$ and Z, the random variable $T_v^{\theta_G}(\xi \odot \mathbf{U}, Z, \tilde{Z})$ is, conditionally to $T_g(\xi \odot \mathbf{U}, Z) \le \theta_G$, uniformly distributed over $[0, 1]$. Recall also that $N_g^{\theta_G}(\mathbf{U}, Z)$ is, by definition, equal to $\mathbb{P}(T_g(\bar{\xi} \odot \mathbf{U}, Z) \le \theta_G | U, Z)$. From this, it follows that

$$\mathbb{P}\left(T_g(\mathbf{U}, Z) \le \theta_G; \tilde{T}_{v,r}^{\theta_G}(\mathbf{U}, \boldsymbol{\xi}, \mathbf{Z}, \tilde{\mathbf{Z}}, Z, \tilde{Z}) \le \theta_V\right)$$

$$= \mathbb{E}\Big(N_g^{\theta_G}(\mathbf{U}, Z) \int_0^1 G_\beta(1 - N_g^{\theta_G}(\mathbf{U}, Z)t, Z, \tilde{Z}),$$

$$K - \lfloor K\theta_G \theta_V \rfloor, \lfloor K\theta_G \theta_V \rfloor + 1)dt\Big)$$

$$= \mathbb{E}\left(N_g^{\theta_G}(\mathbf{U}, Z) \int_0^1 \int_0^{1-N_g^{\theta_G}(\mathbf{U},Z)t} s^{K-\lfloor K\theta_G\theta_V \rfloor - 1}(1-s)^{\lfloor K\theta_G\theta_V \rfloor} ds dt \right)$$

$$= \mathbb{E}\left(N_g^{\theta_G}(\mathbf{U}, Z) \int_0^1 s^{K-\lfloor K\theta_G\theta_V \rfloor - 1}(1-s)^{\lfloor K\theta_G\theta_V \rfloor} \int_0^{\min\left(\frac{1-s}{N_g^{\theta_G}(\mathbf{U},Z)}, 1\right)} dt ds \right)$$

$$\leq \mathbb{E}\left(N_g^{\theta_G}(\mathbf{U}, Z) \int_0^1 s^{K-\lfloor K\theta_G\theta_V \rfloor - 1}(1-s)^{\lfloor K\theta_G\theta_V \rfloor} \int_0^{\frac{1-s}{N_g^{\theta_G}(\mathbf{U},Z)}} dt ds \right)$$

$$= \mathbb{E}\left(\int_0^1 s^{K-\lfloor K\theta_G\theta_V \rfloor - 1}(1-s)^{\lfloor K\theta_G\theta_V \rfloor + 1} ds \right)$$

$$= \frac{\lfloor K\theta_G\theta_V \rfloor + 1}{K + 1}.$$

Hence we proved (24), which finishes the proof of lemma A.1 and of theorem 3.2. □

References

[1] David J Aldous. Exchangeability and related topics. In *École d'Été de Probabilités de Saint-Flour XIII, 1983*, pages 1–198. Springer, 1985. MR0883646

[2] Michael Ashburner, Catherine A Ball, Judith A Blake, David Botstein, Heather Butler, J Michael Cherry, Allan P Davis, Kara Dolinski, Selina S Dwight, Janan T Eppig, et al. Gene ontology: tool for the unification of biology. *Nature Genetics*, 25(1):25–29, 2000.

[3] David J Balding. A tutorial on statistical methods for population association studies. *Nature Reviews Genetics*, 7(10):781–791, 2006.

[4] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300, 1995. MR1325392

[5] Yoav Benjamini, Abba M. Krieger, and Daniel Yekutieli. Adaptive linear step-up procedures that control the false discovery rate. *Biometrika*, 93(3):491, 2006. MR2261438

[6] Yoav Benjamini and Daniel Yekutieli. The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics*, pages 1165–1188, 2001. MR1869245

[7] Gilles Blanchard and Donald Geman. Hierarchical testing designs for pattern recognition. *Annals of Statistics*, pages 1155–1202, 2005. MR2195632

[8] Aiden Corvin, Nick Craddock, and Patrick F Sullivan. Genome-wide association studies: a primer. *Psychological Medicine*, 40(07):1063–1077, 2010.

[9] Thorsten Dickhaus, Klaus Straßburger, Daniel Schunk, Carlos Morcillo-Suarez, Thomas Illig, and Arcadi Navarro. How to analyze many contingency tables simultaneously in genetic association studies. *Statistical applications in genetics and molecular biology*, 11(4), 2012. MR2958611

[10] Brooke L Fridley. Bayesian variable and model selection methods for genetic association studies. *Genetic Epidemiology*, 33(1):27–37, 2009.

[11] K. Ruben Gabriel. Simultaneous test procedures–some theory of multiple comparisons. *The Annals of Mathematical Statistics*, 40(1):pp. 224–250, 1969. MR0240931

[12] Phillip Good. *Permutation tests: a practical guide to resampling methods for testing hypotheses.* Springer Science & Business Media, 2013. MR2103758

[13] Yosef Hochberg. A sharper Bonferroni procedure for multiple tests of significance. *Biometrika*, 75(4):800, 1988. MR0995126

[14] Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979. MR0538597

[15] Hailiang Huang, Pritam Chanda, Alvaro Alonso, Joel S Bader, and Dan E Arking. Gene-based tests of association. *PLoS Genetics*, 7(7):e1002177, 2011.

[16] Nicolai Meinshausen. Hierarchical testing of variable importance. *Biometrika*, 95(2):265–278, 2008. MR2521583

[17] Shaun Purcell, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, Manuel AR Ferreira, David Bender, Julian Maller, Pamela Sklar, Paul IW De Bakker, Mark J Daly, et al. Plink: a tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics*, 81(3):559–575, 2007.