

NON-ASYMPTOTIC RESAMPLING-BASED CONFIDENCE REGIONS AND MULTIPLE TESTS IN HIGH DIMENSION

BY SYLVAIN ARLOT, GILLES BLANCHARD AND ETIENNE ROQUAIN

Universite Paris-Sud, Fraunhofer FIRST.IDA and Vrije Universiteit

We study generalized bootstrapped confidence regions for the mean of a random vector whose coordinates have an unknown dependence structure. The dimensionality of the vector can possibly be much larger than the number of observations and we focus on a non-asymptotic control of the confidence level. The random vector is supposed to be either Gaussian or to have a symmetric bounded distribution. We consider two approaches, the first based on a concentration principle and the second on a direct bootstrapped quantile. The first one allows us to deal with a very large class of resampling weights while our results for the second are specific to Rademacher weights. We present an application of these results to the one-sided and two-sided multiple testing problem, in which we derive several resampling-based step-down procedures providing a non-asymptotic FWER control. We compare our different procedures in a simulation study, and we show that they can outperform Bonferroni's or Holm's procedures as soon as the observed vector has sufficiently correlated coordinates.

1. Introduction.

1.1. *Goals and motivations.* In this work, we assume that we observe a sample $\mathbf{Y} := (\mathbf{Y}^1, \dots, \mathbf{Y}^n)$ of $n \geq 2$ i.i.d. observations of an integrable random vector $\mathbf{Y}^i \in \mathbb{R}^K$ with dimensionality K possibly much larger than n , with an unknown dependence structure of the coordinates. Let $\mu \in \mathbb{R}^K$ denote the common mean of the \mathbf{Y}^i ; our main goal is to find a non-asymptotic $(1-\alpha)$ -confidence region $\mathcal{G}(\mathbf{Y}, 1-\alpha)$ for μ , of the form:

$$(1) \quad \mathcal{G}(\mathbf{Y}, 1-\alpha) = \left\{ x \in \mathbb{R}^K \mid \phi(\bar{\mathbf{Y}} - x) \leq t_\alpha(\mathbf{Y}) \right\},$$

where $\phi : \mathbb{R}^K \rightarrow \mathbb{R}$ is a measurable function (measuring a kind of distance, for example an ℓ_p -norm for $p \in [1, \infty]$), $\alpha \in (0, 1)$, $t_\alpha : (\mathbb{R}^K)^n \rightarrow \mathbb{R}$ is a measurable data-dependent threshold, and $\bar{\mathbf{Y}} = \frac{1}{n} \sum_{i=1}^n \mathbf{Y}^i$ is the empirical mean of the sample \mathbf{Y} .

AMS 2000 subject classifications: Primary 62G15; secondary 62G09, 62G10

Keywords and phrases: confidence regions, family-wise error, multiple testing, high dimensional data, non-asymptotic error control, resampling, cross-validation, concentration inequalities, resampled quantile

As a particular application of such confidence regions, we will focus on the following multiple testing problem: suppose we want to test simultaneously for all vector coordinates $1 \leq k \leq K$ the null hypotheses $H_k : \mu_k \leq 0$ against $A_k : \mu_k > 0$. A classical type of procedure consists in rejecting the null hypotheses \mathcal{H}_k for indices $k \in R(\mathbf{Y}, \alpha)$ corresponding to those empirical means that are larger than a possibly data-dependent threshold t :

$$(2) \quad R(\mathbf{Y}, \alpha) = \left\{ 1 \leq k \leq K \mid \bar{\mathbf{Y}}_k > t(\mathbf{Y}, \alpha) \right\} .$$

The error of such a multiple testing procedure can be measured by the family-wise error rate (FWER) defined by the probability that at least one hypothesis is wrongly rejected. Denoting by $\mathcal{H}_0 = \{k \mid \mu_k \leq 0\}$ the set of coordinates corresponding to the true null hypotheses, the FWER of the procedure defined in (2) can be controlled as follows:

$$\mathbb{P} \left(\exists k \mid \bar{\mathbf{Y}}_k > t(\mathbf{Y}) \text{ and } \mu_k \leq 0 \right) \leq \mathbb{P} \left(\sup_{k \in \mathcal{H}_0} \left\{ \bar{\mathbf{Y}}_k - \mu_k \right\} > t(\mathbf{Y}) \right) .$$

Since μ_k is unknown under H_k , controlling the above probability by a level α is equivalent to establishing a $(1 - \alpha)$ -confidence region for μ of the form (1), using $\phi(x) = \sup_{k \in \mathcal{H}_0} (x_k)$. Similarly, the same reasoning with $\phi(x) = \sup_{k \in \mathcal{H}_0} |x_k|$ in (1) allows us to test $H_k : \mu_k = 0$ against $A_k : \mu_k \neq 0$, by choosing the rejection set $\left\{ 1 \leq k \leq K \mid \left| \bar{\mathbf{Y}}_k \right| > t_\alpha(\mathbf{Y}) \right\}$.

In the framework we consider in the present work, we emphasize that:

- we aim at obtaining a *non-asymptotical* result valid for any fixed K and n , with K possibly much larger than the number of observations n .
- we do not want to make any specific assumption on the dependency structure of the coordinates of \mathbf{Y}^i (although we will consider some general assumptions over the distribution of \mathbf{Y} , for example that it is Gaussian).

In the Gaussian case, a traditional parametric method based on the direct estimation of the covariance matrix to derive a confidence region would not be appropriate in the situation where $K \gg n$, unless the covariance matrix is assumed to belong to some parametric model of lower dimension, which we explicitly don't want to posit here. In this sense our approach is closer in spirit to non-parametric or semiparametric statistics.

This viewpoint is motivated by practical some applications, especially neuroimaging (see [18, 6, 14]). In a magnetoencephalography (MEG) experiment, each observation \mathbf{Y}^i is a two or three dimensional brain activity map, obtained as a difference between brain activities with and without some stimulation. This map is typically composed of 15000 points, or a time series of length between 50 and 1000 of such data. The dimensionality K thus goes from 10^4 to 10^7 . Such observations are repeated $n = 15$ up to 4000 times, but this upper bound is seldom attained (see [28]). Typically, $n \leq 100 \ll K$. In such data, there are strong dependencies between locations (the 15000 points are obtained by pre-processing data of 150 sensors) which

are highly spatially non-homogeneous, as remarked by [18]. Moreover, there may be long-distance correlations, *e.g.* depending on neural connections inside the brain, so that a simple parametric model of the dependency structure is generally not adequate. Another motivating example is given by microarray data, where it is common to observe samples of limited size (*e.g.* less than 100) of a vector in high dimension (*e.g.* more than 20,000, each dimension corresponding to a specific gene), and where there the dependency structure can be quite arbitrary.

1.2. *Two approaches to our goal.* The ideal threshold t_α in (1) is obviously the $(1-\alpha)$ quantile of the distribution of $\phi(\bar{\mathbf{Y}} - \mu)$. However, this quantity depends on the unknown dependency structure of the coordinates of \mathbf{Y}^i and is therefore itself unknown.

In this work we consider using a resampling scheme in order to approach t_α . The heuristics of the resampling method (introduced by [8], generalized to exchangeable weighted bootstrap by [15] and [22]) is that the distribution of the unobservable variable $\bar{\mathbf{Y}} - \mu$ is “mimicked” by the distribution of the observable variable

$$\bar{\mathbf{Y}}^{\langle W - \bar{W} \rangle} := \frac{1}{n} \sum_{i=1}^n (W_i - \bar{W}) \mathbf{Y}^i = \frac{1}{n} \sum_{i=1}^n W_i (\mathbf{Y}^i - \bar{\mathbf{Y}}) = \overline{(\mathbf{Y} - \bar{\mathbf{Y}})}^{\langle W \rangle},$$

conditionally to \mathbf{Y} , where $(W_i)_{1 \leq i \leq n}$ are real random variables independent of \mathbf{Y} called the *resampling weights*, and $\bar{W} = n^{-1} \sum_{i=1}^n W_i$. We emphasize that the family $(W_i)_{1 \leq i \leq n}$ itself *need not be independent*.

Following this general idea, we investigate two separate approaches in order to obtain non-asymptotic confidence regions:

- Approach 1 (“concentration approach”):

The expectations of $\phi(\bar{\mathbf{Y}} - \mu)$ and $\phi(\bar{\mathbf{Y}}^{\langle W - \bar{W} \rangle})$ can be precisely compared, and the processes $\phi(\bar{\mathbf{Y}} - \mu)$ and $\mathbb{E}_W \left[\phi(\bar{\mathbf{Y}}^{\langle W - \bar{W} \rangle}) \right]$ concentrate well around their respective expectations.

- Approach 2 (“direct quantile approach”):

The $1 - \alpha$ quantile of the distribution of $\phi(\bar{\mathbf{Y}}^{\langle W - \bar{W} \rangle})$ conditionally to \mathbf{Y} is close to the one of $\phi(\bar{\mathbf{Y}} - \mu)$.

The first approach above is closely related to the notion of Rademacher complexity in learning theory, and our results in this direction are heavily inspired by the recent work of Fromont [9], who studies general resampling schemes in a learning theoretical setting. It may also be seen to some extent as a generalization of cross-validation methods. For what concerns the second approach, we will restrict ourselves specifically to Rademacher weights in our analysis, because we rely heavily on a symmetrization principle.

1.3. *Relation to previous work.* Using resampling to construct confidence regions (see *e.g.* [8, 12, 11, 7, 4, 20]) or multiple testing procedures (see *e.g.* [29, 30, 21, 10, 25]) is a vast field of study in statistics. Roughly speaking, we can mainly distinguish between two types of results:

- asymptotic results, which are based on the fact that the bootstrap process is asymptotically close to the original empirical process (see [27]).
- exact randomized tests (see *e.g.* [23, 24, 26]), which are based on an invariance of the null distribution under a given transformation ; the underlying idea can be traced back to Fisher’s permutation test (see [1]).

Because we focus on a non-asymptotic viewpoint, the asymptotic approach mentioned above is not adapted to the goals we have fixed.

Our “concentration approach” of the previous section is not directly related to either type of the above previous results, but, as already pointed out earlier, is strongly inspired by results coming from learning theory. On the other hand, what we called our “quantile approach” in the previous section is strongly related to exact randomization tests. Namely, we will only consider symmetric distributions: this is a specific instance of an invariance with respect to a transformation and will allow us to make use of distribution-preserving randomization via sign-flipping. The main difference with traditional exact randomization tests is that, because our first goal is to derive a confidence region, the vector of the means is unknown and therefore, so is the exact invariant transformation. Our contribution to this point is essentially to show that the true vector of the means can be replaced by the empirical one in the randomization, for the price of additional terms of smaller order in the threshold thus obtained. To our knowledge, this gives the first non-asymptotic approximation result on resampled quantiles with an unknown distribution mean.

1.4. *Notations.* Let us now define a few notations that will be useful throughout this paper.

- A boldface letter indicates a matrix. This will almost exclusively concern the $K \times n$ data matrix \mathbf{Y} . A superscript index such as \mathbf{Y}^i indicates the i -th column of a matrix.
- If $\mu \in \mathbb{R}^K$, $\mathbf{Y} - \mu$ is the matrix obtained by subtracting μ from each (column) vector of \mathbf{Y} . If $c \in \mathbb{R}$ and $W \in \mathbb{R}^n$, $W - c = (W_i - c)_{1 \leq i \leq n} \in \mathbb{R}^n$.
- If X is a random variable, $\mathcal{D}(X)$ is its distribution and $\text{Var}(X)$ is its variance. We use the notation $X \sim Y$ to indicate that X and Y have the same distribution.
- We denote by $\mathbb{E}_W[\dots]$, the expectation operator over the distribution of the weight vector W only, *i.e.*, conditional to \mathbf{Y} . We use a similar notation \mathbb{P}_W for the corresponding probability operator and $\mathbb{E}_{\mathbf{Y}}, \mathbb{P}_{\mathbf{Y}}$ for the same operations conditional to W . Since \mathbf{Y} and W are always assumed to be independent, the operators \mathbb{E}_W and $\mathbb{E}_{\mathbf{Y}}$ commute by Fubini’s theorem.
- The vector $\sigma = (\sigma_k)_{1 \leq k \leq K}$ is the vector of the standard deviations of the data: $\forall k, 1 \leq k \leq K, \sigma_k = \text{Var}^{1/2}(\mathbf{Y}_k^1)$.
- $\bar{\Phi}$ is the standard Gaussian upper tail function: if $X \sim \mathcal{N}(0, 1)$, $\forall x \in \mathbb{R}, \bar{\Phi}(x) = \mathbb{P}(X \geq x)$.

- We define the mean of the weight vector $\overline{W} = \frac{1}{n} \sum_{i=1}^n W_i$, the empirical mean vector $\overline{Y} = \frac{1}{n} \sum_{i=1}^n Y^i$, and the resampled empirical mean vector $\overline{Y}^{(W)} := \frac{1}{n} \sum_{i=1}^n W_i Y^i$.
- We use the operator $|\cdot|$ to denote the cardinal of a set.

Several properties may be assumed for the function $\phi : \mathbb{R}^K \rightarrow \mathbb{R}$ that will be used to define confidence regions of the form (1):

- Subadditivity: $\forall x, x' \in \mathbb{R}^K, \quad \phi(x + x') \leq \phi(x) + \phi(x')$.
- Positive-homogeneity: $\forall x \in \mathbb{R}^K, \forall \lambda \in \mathbb{R}^+, \quad \phi(\lambda x) = \lambda \phi(x)$.
- Bounded by the p -norm, $p \in [1, \infty]$: $\forall x \in \mathbb{R}^K, |\phi(x)| \leq \|x\|_p$, where $\|x\|_p$ is equal to $(\sum_{k=1}^K |x_k|^p)^{1/p}$ if $p < \infty$ and $\max_k \{|x_k|\}$ for $p = +\infty$.

Finally, we define the following possible assumptions on the generating distribution of \mathbf{Y} :

(GA) The Gaussian assumption: the \mathbf{Y}^i are Gaussian vectors.

(SA) The symmetric assumption: the \mathbf{Y}^i are symmetric with respect to μ i.e. $(\mathbf{Y}^i - \mu) \sim (\mu - \mathbf{Y}^i)$.

(BA)(p, M) The bounded assumption: $\|\mathbf{Y}^i - \mu\|_p \leq M$ a.s.

In this paper, we primarily focus on the Gaussian framework (GA), where the corresponding results will be more accurate. In addition, under (GA) we will always assume that we know some upper bound on a p -norm of σ for some $p \geq 1$ (this assumption is not restrictive, see discussion in Section 6.3).

The paper is organized as follows. We first build confidence regions following the two different techniques sketched above; Section 2 deals with the concentration method with general weights, and Section 3 with a direct quantile approach using Rademacher weights. We then focus on the multiple testing problem in Section 4, where we deduce step-down multiple testing procedures from our previous confidence regions, following the general principles laid down in [26]. Finally, Section 5 illustrates our results on both confidence regions and multiple testing with a simulation study. Section 6 gives discussions and concluding remarks. All the proofs are given in Section 7.

2. Confidence region using concentration.

2.1. *Main result.* We consider here a general *resampling weight vector* W , that is, a \mathbb{R}^n -valued random vector $W = (W_i)_{1 \leq i \leq n}$ independent of \mathbf{Y} satisfying the following properties: for all $i \in \{1, \dots, n\}$ $\mathbb{E}[W_i^2] < \infty$ and $n^{-1} \sum_{i=1}^n \mathbb{E}|W_i - \overline{W}| > 0$.

We will mainly consider in this section an *exchangeable resampling weight vector*, that is, a resampling weight vector W such that $(W_i)_{1 \leq i \leq n}$ has an exchangeable distribution (*i.e.*, invariant under any permutation of the indices). Several examples of exchangeable resampling weight vectors are given below in Section 2.4, where we also address the question of how to choose between different possible distributions of W . An extension of our results to non-exchangeable weight vectors is studied in Section 2.5.1.

Four constants that depend only on the distribution of W appear in the results below (the fourth one is defined only for a particular class of weights). They are defined as follows and

computed for classical resamplings in Table 1:

$$(3) \quad A_W := \mathbb{E} |W_1 - \bar{W}|$$

$$(4) \quad B_W := \mathbb{E} \left[\left(\frac{1}{n} \sum_{i=1}^n (W_i - \bar{W})^2 \right)^{\frac{1}{2}} \right]$$

$$(5) \quad C_W := \left(\frac{n}{n-1} \mathbb{E} \left[(W_1 - \bar{W})^2 \right] \right)^{\frac{1}{2}}$$

$$(6) \quad D_W := a + \mathbb{E} |\bar{W} - x_0| \quad \text{if } \forall i, |W_i - x_0| = a \text{ a.s. (with } a > 0, x_0 \in \mathbb{R}).$$

Note that these quantities are positive for an exchangeable resampling weight vector W :

$$0 < A_W \leq B_W \leq C_W \sqrt{1 - 1/n}.$$

Moreover, if the weights are i.i.d., we have $C_W = \text{Var}(W_1)^{\frac{1}{2}}$. We can now state the main result of this section:

Theorem 2.1 *Fix $\alpha \in (0, 1)$ and $p \in [1, \infty]$. Let $\phi : \mathbb{R}^K \rightarrow \mathbb{R}$ be any function subadditive, positive-homogeneous and bounded by the p -norm, and let W be an exchangeable resampling weight vector.*

1. *If \mathbf{Y} satisfies (GA), then*

$$(7) \quad \phi(\bar{\mathbf{Y}} - \mu) < \frac{\mathbb{E}_W \left[\phi \left(\bar{\mathbf{Y}}^{\langle W - \bar{W} \rangle} \right) \right]}{B_W} + \|\sigma\|_p \bar{\Phi}^{-1}(\alpha/2) \left[\frac{C_W}{nB_W} + \frac{1}{\sqrt{n}} \right]$$

holds with probability at least $1 - \alpha$. The same bound holds for the lower deviations, i.e. with inequality (7) reversed and the additive term replaced by its opposite.

2. *If \mathbf{Y} satisfies (BA)(p, M) and (SA), then*

$$(8) \quad \phi(\bar{\mathbf{Y}} - \mu) < \frac{\mathbb{E}_W \left[\phi \left(\bar{\mathbf{Y}}^{\langle W - \bar{W} \rangle} \right) \right]}{A_W} + \frac{2M}{\sqrt{n}} \sqrt{\log(1/\alpha)}$$

holds with probability at least $1 - \alpha$. If moreover the weight vector satisfies the assumption of (6), then

$$(9) \quad \phi(\bar{\mathbf{Y}} - \mu) > \frac{\mathbb{E}_W \left[\phi \left(\bar{\mathbf{Y}}^{\langle W - \bar{W} \rangle} \right) \right]}{D_W} - \frac{M}{\sqrt{n}} \sqrt{1 + \frac{A_W^2}{D_W^2}} \sqrt{2 \log(1/\alpha)}$$

holds with probability at least $1 - \alpha$.

Inequalities (7) and (8) give regions of the form (1) that are confidence regions of level at least $1 - \alpha$.

In specific situations, it can be the case that an alternate analysis of the problem can lead to deriving a deterministic threshold t_α such that $\mathbb{P}(\phi(\bar{\mathbf{Y}} - \mu) > t_\alpha) \leq \alpha$. In this case, we would ideally like to take the “best of two approaches” and consider the minimum of t_α and the resampling-based thresholds considered above. In the Gaussian case, the following corollary establishes that we can combine the concentration threshold corresponding to (7) with t_α to obtain a threshold that is very close to the minimum of the two.

Corollary 2.2 *Fix $\alpha, \delta \in (0, 1)$, $p \in [1, \infty]$ and take ϕ and W as in Theorem 2.1. Suppose that \mathbf{Y} satisfies (GA) and that $t_{\alpha(1-\delta)}$ is a real number such that $\mathbb{P}(\phi(\bar{\mathbf{Y}} - \mu) > t_{\alpha(1-\delta)}) \leq \alpha(1-\delta)$. Then with probability at least $1 - \alpha$, $\phi(\bar{\mathbf{Y}} - \mu)$ is less than or equal to the minimum between $t_{\alpha(1-\delta)}$ and*

$$(10) \quad \frac{\mathbb{E}_W \left[\phi \left(\bar{\mathbf{Y}}^{\langle W - \bar{W} \rangle} \right) \right]}{B_W} + \frac{\|\sigma\|_p \bar{\Phi}^{-1} \left(\frac{\alpha(1-\delta)}{2} \right)}{\sqrt{n}} + \frac{\|\sigma\|_p C_W \bar{\Phi}^{-1} \left(\frac{\alpha\delta}{2} \right)}{nB_W}.$$

Remark 2.3

1. Corollary 2.2 is more precisely a consequence of Proposition 2.8 (ii).
2. The important point to notice in Corollary 2.2 is that, since the last term of (10) becomes negligible with respect to the rest when n grows large, we can choose δ to be quite small (for instance $\delta = 1/n$), and obtain a threshold very close to the minimum between t_α and the threshold corresponding to (7). Therefore, this result is more subtle than just considering the minimum of two testing thresholds each taken at level $1 - \frac{\alpha}{2}$, as would be obtained by a direct union bound.
3. For instance, if $\phi = \sup(\cdot)$ (resp. $\sup|\cdot|$), Corollary 2.2 may be applied with t_α equal to the classical Bonferroni threshold (obtained using a simple union bound over coordinates)

$$(11) \quad t_{Bonf,\alpha} := \frac{1}{\sqrt{n}} \|\sigma\|_\infty \bar{\Phi}^{-1} \left(\frac{\alpha}{K} \right) \left(\text{resp. } t'_{Bonf,\alpha} := \frac{1}{\sqrt{n}} \|\sigma\|_\infty \bar{\Phi}^{-1} \left(\frac{\alpha}{2K} \right) \right).$$

We thus obtain a confidence region almost equal to Bonferroni’s for small correlations and better than Bonferroni’s for strong correlations (see simulations in Section 5).

The proof of Theorem 2.1 involves results which are of self interest: the comparison between the expectations of the two processes $\mathbb{E}_W \left[\phi \left(\bar{\mathbf{Y}}^{\langle W - \bar{W} \rangle} \right) \right]$ and $\phi(\bar{\mathbf{Y}} - \mu)$ and the concentration of these processes around their means. These two issues are correspondingly examined in the two next sections (2.2 and 2.3). In Section 2.4, we give some elements for an appropriate choice of resampling weight vectors among several classical examples. The last section (2.5) tackles the practical issue of computation time.

2.2. *Comparison in expectation.* In this section, we compare $\mathbb{E} \left[\phi \left(\bar{\mathbf{Y}}^{\langle W - \bar{W} \rangle} \right) \right]$ and $\mathbb{E} \left[\phi \left(\bar{\mathbf{Y}} - \mu \right) \right]$.

We note that these expectations exist in the Gaussian (GA) and the bounded (BA) cases provided that ϕ is measurable and bounded by a p -norm. Otherwise, in particular in Propositions 2.4 and 2.6, we assume that these expectations exist. In the Gaussian case, these quantities are equal up to a factor that depends only on the distribution of W :

Proposition 2.4 *Let \mathbf{Y} be a sample satisfying (GA) and let W be a resampling weight vector. Then, for any measurable positive-homogeneous function $\phi : \mathbb{R}^K \rightarrow \mathbb{R}$, we have the following equality:*

$$(12) \quad B_W \mathbb{E} \left[\phi \left(\bar{\mathbf{Y}} - \mu \right) \right] = \mathbb{E} \left[\phi \left(\bar{\mathbf{Y}}^{\langle W - \bar{W} \rangle} \right) \right] .$$

Remark 2.5

1. In general, we can compute the value of B_W by simulation. For some classical weights, we give bounds or exact expressions (see Table 1 and Section 7.4).
2. In a non-Gaussian framework, the constant B_W is still relevant, at least asymptotically: Theorem 3.6.13 in [27] uses the limit of B_W when n goes to infinity as a normalizing constant.
3. If the weights satisfy $\sum_{i=1}^n (W_i - \bar{W})^2 = n$ a.s., then (12) holds for any function ϕ (and $B_W = 1$).

When the sample is only supposed to be symmetric we obtain the following inequalities:

Proposition 2.6 *Let \mathbf{Y} be a sample satisfying (SA), W an exchangeable resampling weight vector and $\phi : \mathbb{R}^K \rightarrow \mathbb{R}$ any subadditive, positive-homogeneous function.*

(i) *We have the general following lower bound:*

$$(13) \quad A_W \mathbb{E} \left[\phi \left(\bar{\mathbf{Y}} - \mu \right) \right] \leq \mathbb{E} \left[\phi \left(\bar{\mathbf{Y}}^{\langle W - \bar{W} \rangle} \right) \right] .$$

(ii) *If the weight vector satisfies the assumption of (6), we have the following upper bound:*

$$(14) \quad D_W \mathbb{E} \left[\phi \left(\bar{\mathbf{Y}} - \mu \right) \right] \geq \mathbb{E} \left[\phi \left(\bar{\mathbf{Y}}^{\langle W - \bar{W} \rangle} \right) \right] .$$

Remark 2.7

1. The bounds (13) and (14) are tight for Rademacher and Random hold-out ($n/2$) weights, but far less optimal in some other cases like Leave-one-out (see Section 2.4 for details).
2. When \mathbf{Y} is not assumed to have a symmetric distribution and $\bar{W} = 1$ a.s., Proposition 2 of [9] shows that (13) holds with $\mathbb{E}(W_1 - \bar{W})_+$ instead of A_W . Therefore, assumption (SA) allows us to get a tighter result (for instance twice sharper with Efron or Random hold-out (q) weights). Moreover, it can be shown (see [2], Chapter 9) that this factor 2 is

unavoidable in general without (SA), although it is unnecessary when n goes to infinity. Nevertheless, we conjecture that an inequality close to (13) holds under an assumption less restrictive than (SA) (e.g. \mathbf{Y}^1 is not “too asymmetric”).

2.3. Concentration around the expectation. In this section we present concentration results for the two processes $\phi(\bar{\mathbf{Y}} - \mu)$ and $\mathbb{E}_W \left[\phi \left(\bar{\mathbf{Y}}^{\langle W - \bar{W} \rangle} \right) \right]$ in the Gaussian framework.

Proposition 2.8 *Let $p \in [1, \infty]$, \mathbf{Y} a sample satisfying (GA) and $\phi : \mathbb{R}^K \rightarrow \mathbb{R}$ be any subadditive function, bounded by the p -norm.*

(i) *For all $\alpha \in (0, 1)$, with probability at least $1 - \alpha$ the following holds:*

$$(15) \quad \phi(\bar{\mathbf{Y}} - \mu) < \mathbb{E} \left[\phi(\bar{\mathbf{Y}} - \mu) \right] + \frac{\|\sigma\|_p \bar{\Phi}^{-1}(\alpha/2)}{\sqrt{n}},$$

and the same bound holds for the corresponding lower deviations.

(ii) *Let W be an exchangeable resampling weight vector. Then, for all $\alpha \in (0, 1)$, with probability at least $1 - \alpha$ the following holds:*

$$(16) \quad \mathbb{E}_W \left[\phi \left(\bar{\mathbf{Y}}^{\langle W - \bar{W} \rangle} \right) \right] < \mathbb{E} \left[\phi \left(\bar{\mathbf{Y}}^{\langle W - \bar{W} \rangle} \right) \right] + \frac{\|\sigma\|_p C_W \bar{\Phi}^{-1}(\alpha/2)}{n},$$

and the same bound holds for the corresponding lower deviations.

The bound (15) with a remainder in $n^{-1/2}$ is classical. The bound (16) is much more interesting because it illustrates one of the key properties of resampling: the “stabilization effect”. Indeed, the resampling quantity $\mathbb{E}_W \left[\phi \left(\bar{\mathbf{Y}}^{\langle W - \bar{W} \rangle} \right) \right]$ concentrates around its expectation at the rate $C_W n^{-1} = o(n^{-1/2})$ for most of the weights (see Section 2.4 and Table 1 for more details). Thus, compared to the original process, the resampled mean is “almost deterministic” and equal to $B_W \mathbb{E} \left[\phi(\bar{\mathbf{Y}} - \mu) \right]$. In an asymptotic viewpoint, this may be understood through Edgeworth expansions. Indeed, it is well-known (see for instance [11]) that when ϕ is smooth enough, the first non-zero term in the Edgeworth expansion of

$$\mathbb{E}_W \left[\phi \left(\bar{\mathbf{Y}}^{\langle W - \bar{W} \rangle} \right) \right] - \mathbb{E} \left[\phi \left(\bar{\mathbf{Y}}^{\langle W - \bar{W} \rangle} \right) \right]$$

is at least of order n^{-1} .

Remark 2.9 *Combining expression (12) and Proposition 2.8 (ii), we derive that for a Gaussian sample \mathbf{Y} and any $p \in [1, \infty]$, the following upper bound holds with probability at least $1 - \alpha$:*

$$(17) \quad \mathbb{E} \left\| \bar{\mathbf{Y}} - \mu \right\|_p < \frac{\mathbb{E}_W \left[\left\| \bar{\mathbf{Y}}^{\langle W - \bar{W} \rangle} \right\|_p \right]}{B_W} + \frac{\|\sigma\|_p C_W \bar{\Phi}^{-1}(\alpha/2)}{n B_W},$$

Efron Efr., $n \rightarrow +\infty$	$2\left(1 - \frac{1}{n}\right)^n = A_W \leq B_W \leq \sqrt{\frac{n-1}{n}} \quad C_W = 1$ $\frac{2}{e} = A_W \leq B_W \leq 1 = C_W$
Rademacher Rad., $n \rightarrow +\infty$	$1 - \frac{1}{\sqrt{n}} \leq A_W \leq B_W \leq \sqrt{1 - \frac{1}{n}} \quad C_W = 1 \leq D_W \leq 1 + \frac{1}{\sqrt{n}}$ $A_W = B_W = C_W = D_W = 1$
rho(q) rho($n/2$)	$A_W = 2\left(1 - \frac{q}{n}\right) \quad B_W = \sqrt{\frac{n}{q} - 1}$ $C_W = \sqrt{\frac{n}{n-1}} \sqrt{\frac{n}{q} - 1} \quad D_W = \frac{n}{2q} + \left 1 - \frac{n}{2q}\right $ $A_W = B_W = D_W = 1 \quad C_W = \sqrt{\frac{n}{n-1}}$
Leave-one-out	$\frac{2}{n} = A_W \leq B_W = \frac{1}{\sqrt{n-1}} \quad C_W = \frac{\sqrt{n}}{n-1} \quad D_W = 1$
regular V -fcv	$A_W = E_W = \frac{2}{V} \leq B_W = \frac{1}{\sqrt{V-1}} \quad C_W = \sqrt{n(V-1)^{-1}} \quad D_W = 1.$

TABLE 1

Resampling constants for some classical resampling weight vectors.

and a similar lower bound holds. This gives an observable control with high probability of the L^p -risk of the estimator $\bar{\mathbf{Y}}$ of the mean $\mu \in \mathbb{R}^K$ at the rate $C_W B_W^{-1} n^{-1}$.

2.4. *Resampling weight vectors.* In this section, we consider the question of choosing some appropriate exchangeable resampling weight vector W when using Theorem 2.1 or Corollary 2.2. We define the following classical resampling weight vectors:

1. **Rademacher:** W_i i.i.d. Rademacher variables, *i.e.* $W_i \in \{-1, 1\}$ with equal probabilities.
2. **Efron** (Efron's bootstrap weights): W has a multinomial distribution with parameters $(n; n^{-1}, \dots, n^{-1})$.
3. **Random hold-out** (q) (rho(q) for short), $q \in \{1, \dots, n\}$: $W_i = \frac{n}{q} \mathbb{1}_{i \in I}$, where I is uniformly distributed on subsets of $\{1, \dots, n\}$ of cardinality q . These weights may also be called cross validation weights, or leave- $(n - q)$ -out weights. A classical choice is $q = n/2$ (when n is even). When $q = n - 1$, these weights are called **leave-one-out** weights. Note that this resampling scheme is a particular case of subsampling.

For these classical weights, exact or approximate values for the quantities A_W , B_W , C_W and D_W (defined by equations (3) to (6)) can be easily derived (see Table 1). Proofs are given in Section 7.4, where several other weights are considered. Now, to use Theorem 2.1 or Corollary 2.2, we have to choose a particular resampling weight vector. In the Gaussian case, we propose the following accuracy and complexity criteria:

- first, relation (7) suggests that the quantity $C_W B_W^{-1}$ can be proposed as *accuracy* index for W . Namely, this index enters directly in the deviation term of the corresponding upper bound and the smaller the index is, the sharper the bound.

- second, an upper bound on the computational burden to compute exactly the resampling quantity is given by the cardinality of the support of $\mathcal{D}(W)$, thus providing a *complexity* index. These two criteria are estimated in Table 2 for classical weights. For any exchangeable weight vector W , we have $C_W B_W^{-1} \geq [n/(n-1)]^{1/2}$ and the cardinality of the support of $\mathcal{D}(W)$ is larger than n . Therefore, the *leave-one-out weights* satisfy the best accuracy-complexity trade-off among exchangeable weights.

Resampling	$C_W B_W^{-1}$ (accuracy)	$ \text{supp } \mathcal{L}(W) $ (complexity)
Efron	$\leq \frac{1}{2} \left(1 - \frac{1}{n}\right)^{-n} \xrightarrow{n \rightarrow \infty} \frac{e}{2}$	$\binom{2n-1}{n-1} = \Omega(n^{-\frac{1}{2}} 4^n)$
Rademacher	$\leq \left(1 - n^{-1/2}\right)^{-1} \xrightarrow{n \rightarrow \infty} 1$	2^n
rho ($n/2$)	$= \sqrt{\frac{n}{n-1}} \xrightarrow{n \rightarrow \infty} 1$	$\binom{n}{n/2} = \Omega(n^{-1/2} 2^n)$
Leave-one-out	$= \sqrt{\frac{n}{n-1}} \xrightarrow{n \rightarrow \infty} 1$	n
regular V -fcv	$= \sqrt{\frac{n}{V-1}}$	V

TABLE 2

Choice of the resampling weight vectors: accuracy-complexity trade-off.

Remark 2.10 (Link to leave-one-out prediction risk estimation) Consider using $\bar{\mathbf{Y}}$ for predicting a new data point $\mathbf{Y}^{n+1} \sim \mathbf{Y}^1$ (independent of $\mathbf{Y} = (\mathbf{Y}^1, \dots, \mathbf{Y}^n)$). The corresponding L^p -prediction risk is given by $\mathbb{E} \left\| \bar{\mathbf{Y}} - \mathbf{Y}^{n+1} \right\|_p$. In the Gaussian setting, this prediction risk is proportional to the L^p -risk: $\mathbb{E} \left\| \bar{\mathbf{Y}} - \mu \right\|_p = (n+1)^{\frac{1}{2}} \mathbb{E} \left\| \bar{\mathbf{Y}} - \mathbf{Y}^{n+1} \right\|_p$, so that the estimator of the L^p -risk proposed in Remark 2.9 leads to an estimator of the prediction risk. In particular, using leave-one-out weights and noting $\bar{\mathbf{Y}}^{(-i)}$ the mean of the $(\mathbf{Y}^j, j \neq i, 1 \leq j \leq n)$, we have then established that the leave-one-out estimator

$$\frac{1}{n} \sum_{i=1}^n \left\| \bar{\mathbf{Y}}^{(-i)} - \mathbf{Y}^i \right\|_p$$

correctly estimates the prediction risk (up to the factor $(1 - 1/n^2)^{\frac{1}{2}} \sim 1$).

2.5. *Practical computation of the thresholds.* In practice, the exact computation of the resampling quantity $\mathbb{E}_W \left[\phi \left(\bar{\mathbf{Y}}^{(W-\bar{W})} \right) \right]$ can still be too complex for the weights defined above. In this section we consider two possible ways to address this issue. First, it is possible to use non-exchangeable weights with a lower complexity index and for which the exact computation is tractable. Alternatively, we propose to use a Monte-Carlo approximation, as is often done in practice to compute resampled quantities. In both cases, the thresholds have to be made slightly larger in order to keep the level larger than $1 - \alpha$. This is detailed in the two paragraphs below.

2.5.1. *V -fold cross-validation weights.* In order to reduce the computation complexity, we can use “piece-wise exchangeable” weights: consider a regular partition $(B_j)_{1 \leq j \leq V}$ of $\{1, \dots, n\}$ (where $V \in \{2, \dots, n\}$ and $V|n$), and define the weights $W_i = \frac{V}{V-1} \mathbb{1}_{i \notin B_J}$ with J uniformly distributed on $\{1, \dots, V\}$. These weights are called the **(regular) V -fold cross validation weights** (V -fcv for short).

By applying our results to the process $(\tilde{\mathbf{Y}}^j)_{1 \leq j \leq V}$ where $\tilde{\mathbf{Y}}^j = \frac{V}{n} \sum_{i \in B_j} \mathbf{Y}^i$ is the empirical mean of \mathbf{Y} on block B_j , we can show that Theorem 2.1 can be extended to (regular) V -fold

cross validation weights with the following resampling constants:

$$A_W = \frac{2}{V} \quad B_W = \frac{1}{\sqrt{V-1}} \quad C_W = \frac{\sqrt{n}}{V-1} \quad D_W = 1 .$$

Additionally, when V does not divide n and the blocks are no longer regular, Theorem 2.1 can also be generalized, but the constants have more complex expressions (see Section 7.5 for details). With V -fcv weights, the complexity index is only V , but we lose a factor $[(n-1)/(V-1)]^{1/2}$ in the accuracy index. With regard to the accuracy/complexity tradeoff, the most accurate cross-validation weights are leave-one-out ($V = n$), whereas the 2-fcv weights are the best from the computational viewpoint (but also the less accurate). The choice of V is thus a trade-off between these two terms and depends on the particular constraints of each problem.

However, it is worth noting that as far as the bound of inequality (7) is tight, having an accuracy index close to 1 is not necessarily useful. Namely, this will result in a corresponding deviation term of order n^{-1} , while there is additionally another unavoidable deviation term of order $n^{-\frac{1}{2}}$ in the bound. This suggests that an accuracy index of order $o(n^{\frac{1}{2}})$ would actually be sufficient. In other words, using V -fcv with $V = Cn^{\frac{1}{2}}$ and a “large” constant C would result in only a negligible loss of overall accuracy as compared to leave-one-out. (Of course, we have to point out that this discussion is specific to the form of our bound (7). We cannot exclude in principle that a different approach would lead to a different conclusion, unless it can be proved that the deviation terms in our bound cannot be significantly improved, which is an issue we don’t address here.)

2.5.2. Monte-Carlo approximation. When we use a Monte-Carlo approximation in order to evaluate $\mathbb{E}_W \left[\phi \left(\overline{\mathbf{Y}}^{\langle W - \overline{W} \rangle} \right) \right]$, we draw randomly a small number B of i.i.d. weight vectors W^1, \dots, W^B and compute

$$\frac{1}{B} \sum_{j=1}^B \phi \left(\overline{\mathbf{Y}}^{\langle W^j - \overline{W}^j \rangle} \right).$$

This method is quite standard in the bootstrap literature and can be improved in several ways (see for instance [11], appendix II). In Proposition 2.11 below, we propose an explicit correction of the concentration thresholds that takes into account B weight vectors, for bounded weights.

Proposition 2.11 *Let $B \geq 1$ and W^1, \dots, W^B be i.i.d. exchangeable resampling weight vectors such that $W_1^1 - \overline{W}^1 \in [c_1, c_2]$ a.s. Let $p \in [1, \infty]$, $\phi : \mathbb{R}^K \rightarrow \mathbb{R}$ be any subadditive function, bounded by the p -norm.*

If \mathbf{Y} is a fixed sample and for every $k \in \{1, \dots, K\}$, M_k is a median of $(\mathbf{Y}_k^i)_{1 \leq i \leq n}$, then, for

every $\beta \in (0, 1)$,

$$(18) \quad \frac{1}{B} \sum_{j=1}^B \phi \left(\overline{\mathbf{Y}}^{\langle W^j - \overline{W}^j \rangle} \right) \geq \mathbb{E}_W \left[\phi \left(\overline{\mathbf{Y}}^{\langle W - \overline{W} \rangle} \right) \right] \\ - (c_2 - c_1) \sqrt{\frac{\ln(\beta^{-1})}{2B}} \left\| \frac{1}{n} \left(\sum_{i=1}^n |\mathbf{Y}_k^i - M_k| \right) \right\|_k \Big\|_p$$

holds with probability at least $1 - \beta$.

If \mathbf{Y} is generated according to a distribution satisfying (GA), then, for every $\beta \in (0, 1)$ and any deterministic $\nu \in \mathbb{R}^K$,

$$(19) \quad \left\| \frac{1}{n} \left(\sum_{i=1}^n |\mathbf{Y}_k^i - M_k| \right) \right\|_k \Big\|_p \leq \mathbb{E} \left\| \frac{1}{n} \left(\sum_{i=1}^n |\mathbf{Y}_k^i - \nu_k| \right) \right\|_k \Big\|_p + \frac{\|\sigma\|_p \overline{\Phi}^{-1}(\beta/2)}{\sqrt{n}}$$

holds with probability at least $1 - \beta$.

For instance, with Rademacher weights, we can use (18) with $c_2 - c_1 = 2$ and $\beta = \delta\alpha$ ($\delta \in (0, 1)$). Then, in the thresholds built upon Theorem 2.1 and Corollary 2.2, one can replace $\mathbb{E}_W \left[\phi \left(\overline{\mathbf{Y}}^{\langle W - \overline{W} \rangle} \right) \right]$ by its Monte-Carlo approximation at the price of changing α into $(1 - \delta)\alpha$, and adding

$$(20) \quad \frac{2}{B_W} \sqrt{\frac{\ln((\delta\alpha)^{-1})}{2B}} \left\| \frac{1}{n} \left(\sum_{i=1}^n |\mathbf{Y}_k^i - M_k| \right) \right\|_k \Big\|_p$$

to the threshold.

Note that (18) holds conditionally to the observed sample, so that B can be chosen in function of \mathbf{Y} in (20). Therefore, we can choose B with the following strategy: first, compute a rough estimate $t_{\text{est}, \alpha}$ of the final threshold (e.g. if $\phi = \|\cdot\|_\infty$ and \mathbf{Y} is gaussian, take the Bonferroni threshold $\|\sigma\|_\infty n^{-1/2} \overline{\Phi}^{-1}(\alpha/(2K))$ or the single test threshold $\|\sigma\|_\infty n^{-1/2} \overline{\Phi}^{-1}(\alpha/2)$). Second, choose B such that (20) is much smaller than $t_{\text{est}, \alpha}$.

Remark 2.12 *In the Gaussian case, (19) gives a theoretical upper bound on the additive term (if one can bound the expectation term). This is only useful to ensure that the correction (20) is negligible for reasonable values of B .*

3. Confidence region using resampled quantiles.

3.1. *Main result.* In this section, we consider a different approach to construct confidence regions, directly based on the estimation of the quantile via resampling. Remember that our setting is non-asymptotic, so that the standard asymptotic approaches cannot be applied here.

For this reason, we base our approach on ideas coming from exact randomized tests and consider here the case where \mathbf{Y}^1 has a symmetric distribution and where W is an i.i.d Rademacher weight vector, that is, W_i i.i.d. with $W_1 \in \{-1, 1\}$ with equal probabilities.

The idea here is to approximate the quantiles of the distribution $\mathcal{D}\left(\phi\left(\overline{\mathbf{Y}} - \mu\right)\right)$ by the quantiles of the corresponding resampling-based distribution:

$$(21) \quad \mathcal{D}\left(\phi\left(\overline{\mathbf{Y}}^{\langle W - \overline{W} \rangle}\right) \middle| \mathbf{Y}\right).$$

For this, we take advantage of the symmetry of each \mathbf{Y}^i around its mean. Let us define for a function ϕ the resampled empirical quantile by:

$$q_\alpha(\phi, \mathbf{Y}) := \inf \left\{ x \in \mathbb{R} \middle| \mathbb{P}_W \left[\phi(\overline{\mathbf{Y}}^{\langle W \rangle}) > x \right] \leq \alpha \right\}.$$

The following lemma, close in spirit to exact test results, easily derives from the ‘‘symmetrization trick’’, *i.e.* from taking advantage of the distribution invariance of the data via sign-flipping.

Lemma 3.1 *Let \mathbf{Y} be a data sample satisfying assumption (SA) and $\phi : \mathbb{R}^K \rightarrow \mathbb{R}$ be a measurable function. Then the following holds:*

$$(22) \quad \mathbb{P} \left[\phi(\overline{\mathbf{Y}} - \mu) > q_\alpha(\phi, \mathbf{Y} - \mu) \right] \leq \alpha.$$

Of course, since $q_\alpha(\phi, \mathbf{Y} - \mu)$ still depends on the unknown μ , we cannot use this threshold to get a confidence region of the form (1). Therefore, following the general philosophy of resampling, we propose to replace the true mean μ by the empirical mean $\overline{\mathbf{Y}}$ in the quantile $q_\alpha(\phi, \mathbf{Y} - \mu)$. The main technical result of this section quantifies the price to pay to perform this operation:

Theorem 3.2 *Fix $\delta, \alpha_0 \in (0, 1)$. Let \mathbf{Y} be a data sample satisfying assumption (SA). Let $f : (\mathbb{R}^K)^n \rightarrow [0, \infty)$ be a nonnegative measurable function on the set of the data sample. Let $\phi : \mathbb{R}^K \rightarrow \mathbb{R}$ be a nonnegative, subadditive, positive-homogeneous function. Denote $\tilde{\phi}(x) = \max(\phi(x), \phi(-x))$. The following holds:*

$$(23) \quad \mathbb{P} \left[\phi(\overline{\mathbf{Y}} - \mu) > q_{\alpha_0(1-\delta)}(\phi, \mathbf{Y} - \overline{\mathbf{Y}}) + \gamma_1(\alpha_0\delta)f(\mathbf{Y}) \right] \leq \alpha_0 + \mathbb{P} \left[\tilde{\phi}(\overline{\mathbf{Y}} - \mu) > f(\mathbf{Y}) \right],$$

where

$$\gamma_1(\eta) = \frac{2\overline{\mathcal{B}}(n, \frac{\eta}{2}) - n}{n}$$

and

$$\overline{\mathcal{B}}(n, \eta) = \max \left\{ k \in \{0, \dots, n\} \middle| 2^{-n} \sum_{i=k}^n \binom{n}{i} \geq \eta \right\},$$

is the upper quantile function of a Binomial $(n, \frac{1}{2})$ variable.

Remark 3.3 Note that from Hoeffding's inequality, we have

$$\gamma_1(\alpha_0\delta) \leq \left(\frac{2 \ln \left(\frac{2}{\alpha_0\delta} \right)}{n} \right)^{1/2}.$$

We can use this in (23) to derive a more explicit (but slightly less accurate) inequality.

By iteration of Theorem 3.2, we obtain the following corollary:

Corollary 3.4 Fix J a positive integer, $(\alpha_i)_{i=0,\dots,J-1}$ a finite sequence in $(0, 1)$ and $\delta \in (0, 1)$. Consider \mathbf{Y} , f , ϕ and $\tilde{\phi}$ as in Theorem 3.2. Then the following holds:

$$(24) \quad \mathbb{P} \left[\phi(\bar{\mathbf{Y}} - \mu) > q_{\alpha_0(1-\delta)}(\phi, \mathbf{Y} - \bar{\mathbf{Y}}) + \sum_{i=1}^{J-1} \gamma_i q_{\alpha_i(1-\delta)}(\tilde{\phi}, \mathbf{Y} - \bar{\mathbf{Y}}) + \gamma_J f(\mathbf{Y}) \right] \\ \leq \sum_{i=0}^{J-1} \alpha_i + \mathbb{P} \left[\tilde{\phi}(\bar{\mathbf{Y}} - \mu) > f(\mathbf{Y}) \right],$$

where, for $k \geq 1$, $\gamma_k = n^{-k} \prod_{i=0}^{k-1} \left(2\bar{\mathcal{B}} \left(n, \frac{\alpha_i\delta}{2} \right) - n \right)$.

The rationale behind this result is that the sum appearing inside the probability in (24) should be interpreted as a series of corrective terms of decreasing order of magnitude, since we expect the sequence γ_k to be sharply decreasing. Looking at Hoeffding's bound, this will be the case if the levels are such that $\alpha_i \gg \exp(-n)$.

Looking at (24), we still have to deal with the trailing term on the right-hand-side to obtain a useful result. We did not succeed in obtaining a self-contained result based on the symmetry assumption (SA) alone. However, to upper-bound the trailing term, we can assume some additional regularity assumption on the distribution of the data. For example, if the data are Gaussian or bounded, we can apply the results of the previous section (or apply some other device like Bonferroni's bound (11)). Explicit formulas for the resulting thresholds are given in Section 4 and 5 (with $J = 1$). We want to emphasize that the bound used in this last step does not have to be particularly sharp: since we expect (in favorable cases) γ_J to be very small, the trailing probability term on the right-hand side as well as the contribution of $\gamma_J f(\mathbf{Y})$ to the left-hand side should be very minor. Therefore, even a coarse bound on this last term should suffice.

3.2. Practical computation of the resampled quantile. Since the above results use Rademacher weight vectors, the exact computation of the quantile q_α requires in principle 2^n iterations and thus is too complex as n becomes large. Therefore, it might be relevant to consider a block-wise Rademacher resampling scheme. For this, let $(B_j)_{1 \leq j \leq V}$ be a regular partition of $\{1, \dots, n\}$ and for all $i \in B_j$, $W_i = W_j^B$, where $(W_j^B)_{1 \leq j \leq V}$ are i.i.d. Rademacher. This is equivalent to applying

the previous method to the block-averaged sample $(\tilde{Y}_1, \dots, \tilde{Y}_V)$, where \tilde{Y}_j is the average of the $(Y_i)_{i \in B_j}$. Because the \tilde{Y}_j are i.i.d. variables, all of the previous results carry over when replacing n by V . However, this results in a loss of accuracy in Theorem 3.2 (and then in Corollary 3.4).

Another way to address this computation complexity issue is to consider Monte-Carlo quantile approximation: let \mathbf{W} denote a $n \times B$ matrix of i.i.d. Rademacher weights (independent of all other variables), and define

$$\tilde{q}_\alpha(\phi, \mathbf{Y}, \mathbf{W}) = \inf \left\{ x \in \mathbb{R} \mid \frac{1}{B} \sum_{j=1}^B \mathbb{1} \left\{ \phi \left(\overline{\mathbf{Y}}^{(\mathbf{w}^j)} \right) \geq x \right\} \leq \alpha \right\},$$

that is, \tilde{q}_α is defined just as q_α except that the true distribution \mathbb{P}_W of the Rademacher weight vector is replaced by the empirical distribution constructed from the columns of \mathbf{W} , $\tilde{\mathbb{P}}_W = B^{-1} \sum_{j=1}^B \delta_{\mathbf{w}^j}$. The following result then holds:

Proposition 3.5 *Consider the same conditions as in Theorem 3.2 except the function f can now be a function of both \mathbf{Y} and \mathbf{W} . We have:*

$$\begin{aligned} \mathbb{P}_{\mathbf{Y}, \mathbf{W}} \left[\phi(\overline{\mathbf{Y}} - \mu) > \tilde{q}_{\alpha_0(1-\delta)} \left(\phi, \mathbf{Y} - \overline{\mathbf{Y}}, \mathbf{W} \right) + \gamma(\mathbf{W}, \alpha_0 \delta) f(\mathbf{Y}, \mathbf{W}) \right] \\ \leq \tilde{\alpha}_0 + \mathbb{P}_{\mathbf{Y}, \mathbf{W}} \left[\tilde{\phi}(\overline{\mathbf{Y}} - \mu) > f(\mathbf{Y}, \mathbf{W}) \right], \end{aligned}$$

where $\tilde{\alpha}_0 = \frac{|B\alpha_0|+1}{B+1} \leq \alpha_0 + \frac{1}{B+1}$ and

$$\gamma(\mathbf{W}, \eta) := \max \left\{ y \geq 0 \mid \frac{1}{B} \sum_{j=1}^B \mathbb{1} \left\{ |\overline{W}^j| \geq y \right\} \geq \eta \right\}.$$

is the $(1 - \eta)$ -quantile of $|\overline{W}|$ under the empirical distribution $\tilde{\mathbb{P}}_W$.

Note that for practical purposes, we can choose $f(\mathbf{W}, \mathbf{Y})$ to depend on \mathbf{Y} only and use another type of bound to control the last term on the right-hand side, see discussion in the previous section. The above result tells us that if we replace in Theorem 3.2 the true quantile by an empirical quantile based on B i.i.d. weight vectors, and the factor γ_1 is similarly replaced by an empirical quantile of $|\overline{W}|$, then we lose at most $(B + 1)^{-1}$ in the corresponding covering probability. Furthermore, it can be seen easily that if α_0 is taken to be a positive multiple of $(B + 1)^{-1}$, then there is no loss in the final covering probability (*i.e.* $\tilde{\alpha}_0 = \alpha_0$).

4. Application to multiple testing. In this section, we describe how the results of Sections 2 and 3 can be used to derive multiple testing procedures. We focus on the two following multiple testing problems:

- *One-sided problem:* test simultaneously the null hypotheses $H_k : “\mu_k \leq 0”$ against $A_k : “\mu_k > 0”$, for $1 \leq k \leq K$.

- *Two-sided problem*: test simultaneously the null hypotheses $H_k : “\mu_k = 0”$ against $A_k : “\mu_k \neq 0”$, for $1 \leq k \leq K$.

In this context, we detail the link between confidence regions and multiple testing, and explain how to use the general principle of step-down methods (as exposed in [26]) to obtain sharper thresholds in this context.

We first introduce a few more notations:

- Put $\mathcal{H} := \{1, \dots, K\}$, $\mathcal{H}_0 := \{1 \leq k \leq K \mid H_k \text{ is true}\}$ and \mathcal{H}_1 its complementary in \mathcal{H} . Note that $\mathcal{H}_0, \mathcal{H}_1$ are of course unknown since the goal of multiple testing is in fact precisely to estimate these sets.
- For any $x \in \mathbb{R}$, the bracket $[x]$ denotes either x in the one-sided context or $|x|$ in the two-sided context.
- Reordering the coordinates of $\bar{\mathbf{Y}}$ in decreasing order

$$\left[\bar{\mathbf{Y}}_{\sigma(1)} \right] \geq \left[\bar{\mathbf{Y}}_{\sigma(2)} \right] \geq \dots \geq \left[\bar{\mathbf{Y}}_{\sigma(K)} \right]$$

with a permutation σ of $\{1, \dots, K\}$, we define for every $i \in \{1, \dots, K\}$, $\mathcal{C}_i(\mathbf{Y}) := \{\sigma(j) \mid j \geq i\}$ the set which contains the $K - i + 1$ smaller coordinates of $\left[\bar{\mathbf{Y}} \right]$ (in the sequel, we simply write \mathcal{C}_i instead of $\mathcal{C}_i(\mathbf{Y})$). In particular, note that $\mathcal{C}_1 = \mathcal{H}$.

- For any $\mathcal{C} \subset \mathcal{H}$, define

$$T(\mathcal{C}) := \sup_{k \in \mathcal{C}} \left[\bar{\mathbf{Y}}_k - \mu_k \right] \quad \text{and} \quad T'(\mathcal{C}) := \sup_{k \in \mathcal{C}} \left[\bar{\mathbf{Y}}_k \right] .$$

Note that $T(\mathcal{H}) \geq T(\mathcal{H}_0) \geq T'(\mathcal{H}_0)$ in general and $T(\mathcal{H}_0) = T'(\mathcal{H}_0)$ in the two-sided context.

4.1. *Multiple testing and connection with confidence regions.* A multiple testing procedure is a (measurable) function

$$R(\mathbf{Y}) \subset \mathcal{H} ,$$

that rejects the null hypotheses H_k for all $k \in R(\mathbf{Y})$. Considering such a multiple testing procedure R , a type I error arises for the null hypothesis H_k as soon as R rejects H_k although it was true, *i.e.* $k \in R(\mathbf{Y}) \cap \mathcal{H}_0$. The family-wise error rate of R is then the probability that at least one type I error occurs:

$$\text{FWER}(R) := \mathbb{P}(|R(\mathbf{Y}) \cap \mathcal{H}_0| > 0) .$$

Given a level $\alpha \in (0, 1)$, our goal is to build a multiple testing procedure R with

$$(25) \quad \text{FWER}(R) \leq \alpha .$$

Of course, choosing the procedure $R = \emptyset$ (*i.e.* the procedure which rejects no null hypothesis) satisfies trivially this property. Therefore, under the constraint (25), we want the average number of rejected false null hypotheses, that is,

$$(26) \quad \mathbb{E}|R(\mathbf{Y}) \cap \mathcal{H}_1| ,$$

to be as large as possible.

A common way to build a multiple testing procedure is to reject the null hypotheses H_k corresponding to

$$(27) \quad R(\mathbf{Y}) = \left\{ 1 \leq k \leq K \mid \left[\overline{\mathbf{Y}}_k \right] > t \right\} ,$$

where t is a (possibly data-dependent) threshold. From now on, we will restrict our attention to multiple testing procedures of the previous form. In this case, the deterministic threshold that maximizes (26) under the constraint (25) is obviously the $(1 - \alpha)$ quantile of the distribution of $T'(\mathcal{H}_0)$. However, the latter quantile cannot be directly accessed, because it depends both on the unknown dependency structure between the coordinates of \mathbf{Y}^i and on the unknown set \mathcal{H}_0 . The aim of the following sections (4.2, 4.3, 4.4) will be to approach this quantity.

This should be compared to the confidence region context, where the smallest deterministic threshold for which (1) holds with $\phi(x) = \sup_k [x_k]$ is the $(1 - \alpha)$ quantile of the distribution of $T(\mathcal{H})$. Since $T(\mathcal{H}) \geq T'(\mathcal{H}_0)$, we observe the following:

1. The thresholds that give confidence regions of the form (1) with $\phi(x) = \sup_k [x_k]$ also give multiple testing procedures with a FWER smaller than α (following the thresholding procedure (27)). Therefore, we can directly derive from Sections 2 and 3 resampling-based multiple testing procedures with controlled FWER.
2. One might expect to be able to find better (*i.e.* smaller) thresholds in the multiple testing framework than in the confidence region framework. Namely, when \mathcal{H}_1 is “large”, we expect $T(\mathcal{H})$ to be “significantly larger” than $T'(\mathcal{H}_0)$; therefore procedures based on upper bounding $T(\mathcal{H})$ are likely to be too conservative. However, if we try to follow the same approach as above for bounding directly $T'(\mathcal{H}_0)$, we run into the problem that the function $\phi(x) = \sup_{k \in \mathcal{H}_0} [x_k]$ is not observable since \mathcal{H}_0 is unknown. A method commonly used to address this issue is to consider step-down procedures. This is examined in the following section.

4.2. *Background on step-down procedures.* We review in this section known facts on step-down procedures (see [26]). We consider here thresholds \mathbf{t} of the following general form:

$$\mathbf{t} : \mathcal{C} \subset \mathcal{H} \mapsto \mathbf{t}(\mathcal{C}) \in \mathbb{R} .$$

We call such a threshold a *subset-based threshold* since it gives a value to each subset of \mathcal{H} . Note that these thresholds can also be possibly data-dependent, but we omit the dependence on \mathbf{Y}

here to lighten notation. A subset-based threshold is said to be *non-decreasing* if for all subsets \mathcal{C} and \mathcal{C}' , we have

$$\mathcal{C} \subset \mathcal{C}' \quad \Rightarrow \quad \mathbf{t}(\mathcal{C}) \leq \mathbf{t}(\mathcal{C}') .$$

In our setting, a non-decreasing subset-based threshold is easily obtained by taking a supremum over a subset \mathcal{C} of coordinates. In particular, the thresholds derived from Section 2 (resp. Section 3) define non-decreasing subset-based thresholds, by taking $\phi_{\mathcal{C}}(x) = \sup_{k \in \mathcal{C}} [x_k]$ (resp. $\phi_{\mathcal{C}}(x) = 0 \vee \sup_{k \in \mathcal{C}} [x_k]$).

Definition 4.1 (Step-down procedure with subset-based threshold) *Let \mathbf{t} be a non-decreasing subset-based threshold and note for all i , $t_i = \mathbf{t}(\mathcal{C}_i)$. The step-down procedure with threshold \mathbf{t} rejects*

$$\{1 \leq k \leq K \mid [\bar{\mathbf{Y}}_k] \geq t_{\hat{\ell}}\}$$

where $\hat{\ell} = \max \{1 \leq i \leq K \mid \forall j \leq i, [\bar{\mathbf{Y}}_{\sigma(j)}] \geq t_j\}$ when the latter maximum exists, and the procedure rejects no null hypothesis otherwise.

A step-down procedure of the above form can be computed using the following iterative algorithm:

Algorithm 4.2

1. *Init:* define $R_0 := \emptyset$, $\mathcal{E}_0 := \mathcal{H}$.
2. *Iteration $i \geq 1$:* put $\mathcal{E}_i := \mathcal{E}_{i-1} \setminus R_{i-1}$ and $R_i = \{k \in \mathcal{E}_i \mid [\bar{\mathbf{Y}}_k] \geq \mathbf{t}(\mathcal{E}_i)\}$
If $R_i = \emptyset$, stop and reject the null hypotheses corresponding to:

$$R(\mathbf{Y}) := \{\sigma(k), k \in \cup_{j \leq i-1} R_j\} .$$

Otherwise, go to iteration $i + 1$.

We recall here Theorem 1 of [26], adapted to our setting:

Theorem 4.3 (Romano and Wolf, 2005) *Let \mathbf{t} be a non-decreasing subset-based threshold. Then the step-down procedure R of threshold \mathbf{t} satisfies,*

$$(28) \quad FWER(R) \leq \mathbb{P}(T(\mathcal{H}_0) \geq \mathbf{t}(\mathcal{H}_0)) .$$

As a consequence, Algorithm 4.2 with thresholds derived from Section 2 (resp. Section 3) with $\phi_{\mathcal{C}}(x) = \sup_{k \in \mathcal{C}} [x_k]$ (resp. $\phi_{\mathcal{C}}(x) = 0 \vee \sup_{k \in \mathcal{C}} [x_k]$) gives a multiple testing procedure with control of the FWER. We detail this in the following section.

4.3. *Using our confidence regions to build step-down procedures.* Using Theorem 4.3 and Corollary 2.2 (wherein we use the Bonferroni threshold), we derive:

Corollary 4.4 Fix $\alpha, \delta \in (0, 1)$. Let W be an exchangeable resampling weight vector and suppose that \mathbf{Y} satisfies (GA). Then, in the one-sided context, the step-down procedure with the following subset-based threshold controls the FWER at level α :

$$\mathcal{C} \mapsto \min \left(\frac{\|\sigma\|_\infty}{\sqrt{n}} \overline{\Phi}^{-1} \left(\frac{\alpha(1-\delta)}{|\mathcal{C}|} \right), \frac{\mathbb{E}_W \left[\sup_{k \in \mathcal{C}} \left\{ \left(\overline{\mathbf{Y}}^{(W-\overline{W})} \right)_k \right\} \right]}{B_W} + \varepsilon(\alpha, \delta, n) \right)$$

where $\varepsilon(\alpha, \delta, n) = \frac{\|\sigma\|_\infty}{\sqrt{n}} \overline{\Phi}^{-1} \left(\frac{\alpha(1-\delta)}{2} \right) + \frac{\|\sigma\|_\infty C_W}{n B_W} \overline{\Phi}^{-1} \left(\frac{\alpha \delta}{2} \right)$.

Using Theorem 4.3 and Theorem 3.2 (with $\alpha_0 = \alpha(1-\gamma)$ and f equal to the Bonferroni threshold at level $\alpha\gamma/2$), we derive:

Corollary 4.5 Fix $\alpha, \gamma, \delta \in (0, 1)$. Let W be a Rademacher weight vector and suppose that \mathbf{Y} satisfies (GA). Then, in the one-sided context, the step-down procedure with the following subset-based threshold controls the FWER at level α :

$$\mathcal{C} \mapsto q_{\alpha(1-\gamma)(1-\delta)} \left(0 \vee \phi_{\mathcal{C}}, \mathbf{Y} - \overline{\mathbf{Y}} \right) + \varepsilon'(\alpha, \delta, \gamma, n, |\mathcal{C}|)$$

where $\varepsilon'(\alpha, \delta, \gamma, n, k) = \frac{2\overline{B}(n, \alpha(1-\gamma)\delta/2) - n}{n} \frac{\|\sigma\|_\infty}{\sqrt{n}} \overline{\Phi}^{-1} \left(\frac{\alpha\gamma}{2k} \right)$ and $\phi_{\mathcal{C}}(x) = \sup_{k \in \mathcal{C}} \{x_k\}$.

Of course, analogues of Corollaries 4.4 and 4.5 can also be derived for the two-sided problem.

Remark 4.6

1. Note that the above (data-dependent) subset-based thresholds are translation-invariant because $\mathbf{Y} - \overline{\mathbf{Y}}$ is. Therefore, large values of non-zero means μ_k will not make these thresholds larger.
2. Both subset-based thresholds of Corollary 4.4 and 4.5 are built in order to improve the ‘‘Bonferroni’s subset-based threshold’’

$$\mathcal{C} \mapsto \frac{\|\sigma\|_\infty}{\sqrt{n}} \overline{\Phi}^{-1} \left(\frac{\alpha}{|\mathcal{C}|} \right) .$$

Therefore, the step-down procedures proposed here are expected to perform better than Holm’s procedure (i.e. the step-down version of Bonferroni’s procedure, see [13]).

4.4. *Uncentered quantile approach for two-sided testing.* We now focus specifically on the two-sided multiple testing problem. A fundamental consequence of Theorem 4.3 is that only a weak control (i.e., when $\mathcal{C} = \mathcal{H}_0$) of $T'(\mathcal{C}) = \sup_{k \in \mathcal{C}} |\overline{\mathbf{Y}}_k|$ is needed to obtain a step-down procedure with a strong control (i.e., for arbitrary mean $\mu \in \mathbb{R}^K$) of the FWER. In this situation, the main problem dealt with in Section 3 disappears: namely, under the hypothesis that $\mathcal{H}_0 = \mathcal{C}$, by definition all the coordinates contributing to the supremum in $T'(\mathcal{C})$ are assumed to have zero mean, and therefore, following the reasoning in Lemma 3.1, a direct exact quantile approach is possible.

Corollary 4.7 *Let W be a Rademacher weight vector and suppose that \mathbf{Y} satisfies (SA). Denote $\phi_{\mathcal{C}}(x) = \sup_{k \in \mathcal{C}} |x_k|$. Then for two-sided testing, the step down procedure with the subset-based threshold*

$$\mathcal{C} \mapsto q_{\alpha}(\phi_{\mathcal{C}}, \mathbf{Y})$$

controls the FWER at level α .

Note the differences of this result with our main approach (*i.e.*, the analogue of Corollary 4.5 in the two-sided setting):

- there is no additional trailing term ε' and no “shrinking” in the level of the computed empirical quantile.
- the data is not recentered around the empirical expectation to compute the quantile.

In the following, we will call the threshold $q_{\alpha}(\phi_{\mathcal{C}}, \mathbf{Y})$ the “uncentered quantile threshold”, while the threshold built using our main approach (including the additional term) will be called “recentered quantile threshold” for brevity.

To understand the practical consequences of these differences, let us consider an informal and qualitative argumentation.

- if $\mathcal{C} = \mathcal{H}_0$, then the empirical mean $\bar{\mathbf{Y}}$ should be close to 0. Hence, if we assume that replacing $\bar{\mathbf{Y}}$ by 0 does not change the centered quantile significantly, we conclude that the uncentered quantile threshold will be smaller (hence better) than the recentered quantile threshold, since the latter has the same form but at a slightly shrunk level and has an additional term ε' . In this situation the uncentered quantile will actually achieve the exact level (up to 2^{-n}).
- if on the other hand there are some coordinates with a large non-zero mean in the set \mathcal{C} (by which we mean having a large signal-to-noise (SNR) ratio), then these coordinates will on average have a large absolute value and hence make the uncentered quantile significantly larger; in this case the signal will contribute to the uncentered quantile more than the noise. By contrast, and as remarked earlier, the recentered quantile threshold is translation invariant and thus not affected by the relative strength of the signal. Hence, in this situation, it is likely that the recentered quantile threshold will be smaller.

While the second situation above appears to be detrimental to the uncentered quantile, this disadvantage will in some sense be “automatically corrected” by the step-down procedure. Namely, if some coordinates have a large SNR, they will certainly contribute significantly to the uncentered quantile threshold at the first step of the step-down procedure; however even if this threshold is relatively large, it will still allow to eliminate at the first step precisely those coordinates having a very large mean. This will result in an important improvement of the threshold at the second iteration, and so on, until all coordinates with a large SNR have been weeded out, so that in the end we actually end up very close to the situation described in the first point.

Hence, the conclusion from this qualitative discussion is that, in a situation where some of the coordinates have a large SNR, we expect that the uncentered quantile will be less accurate (*i.e.*, larger) than the centered quantile threshold in the first iteration(s) of the step-down procedure, but that it will then improve along the iterations and eventually prevail in the race. (This behavior will be confirmed by our simulations in the next section.)

At this point, it seems that the step-down using the uncentered quantile is both simpler and more effective than our main approach and thus should always be preferred. However, this qualitative discussion also gives us another insight: the step-down procedure based on the uncentered quantile may need more iterations to converge since the first steps result in an inaccurate threshold. In order to fix this drawback, we propose to use the leverage of the recentered quantile for the first step in order to weed out in one single step most of coordinates having a large SNR, and then continue subsequently with the uncentered threshold in the next steps for more accuracy. We thus obtain the following algorithm:

Algorithm 4.8

1. *Reject the null hypotheses corresponding to:*

$$R_0 := \left\{ k \mid \left| \bar{\mathbf{Y}}_k \right| \geq q_{\alpha(1-\delta)(1-\gamma)}(\|\cdot\|_\infty, \mathbf{Y} - \bar{\mathbf{Y}}) + \varepsilon'(\alpha, \delta, \gamma, n, K) \right\} .$$

2. *If $R_0 = \mathcal{H}$ then stop.*

Otherwise, consider the set of the remaining coordinates $\mathcal{H} \setminus R_0$ and apply on it the step-down algorithm 4.2 with the subset-based threshold

$$\mathcal{C} \mapsto q_{\alpha(1-\gamma)}(\phi_{\mathcal{C}}, \mathbf{Y}) ,$$

where $\phi_{\mathcal{C}}(x) = \sup_{k \in \mathcal{C}} |x_k|$.

Proposition 4.9 *Fix $\alpha, \gamma, \delta \in (0, 1)$. Let W be a Rademacher weight vector and suppose that \mathbf{Y} satisfies (GA). In the two-sided context, the algorithm 4.8 gives a multiple testing procedure with a FWER smaller than α .*

What we expect is that the above algorithm will yield essentially the same final result as the one of Corollary 4.7 (up to some small loss in the level), while requiring less iterations. In numerical applications such as neuroimaging with a large number of images, where one iteration can take up to one day, this can result in a significant improvement.

5. Simulations. For simulations, we consider data of the form $\mathbf{Y}_t = \mu_t + G_t$, where t belongs to a $d \times d$ discretized 2D torus of $K = d^2$ “pixels”, identified with $\mathbb{T}_d^2 = (\mathbb{Z}/d\mathbb{Z})^2$, and G is a centered Gaussian vector obtained by 2D discrete convolution of an i.i.d. standard Gaussian field (“white noise”) on \mathbb{T}_d^2 with a function $F : \mathbb{T}_d^2 \rightarrow \mathbb{R}$ such that $\sum_{t \in \mathbb{T}_d^2} F^2(t) = 1$. This ensures that G is a stationary Gaussian process on the discrete torus, it is in particular isotropic with $\mathbb{E}[G_t^2] = 1$ for all $t \in \mathbb{T}_d^2$.

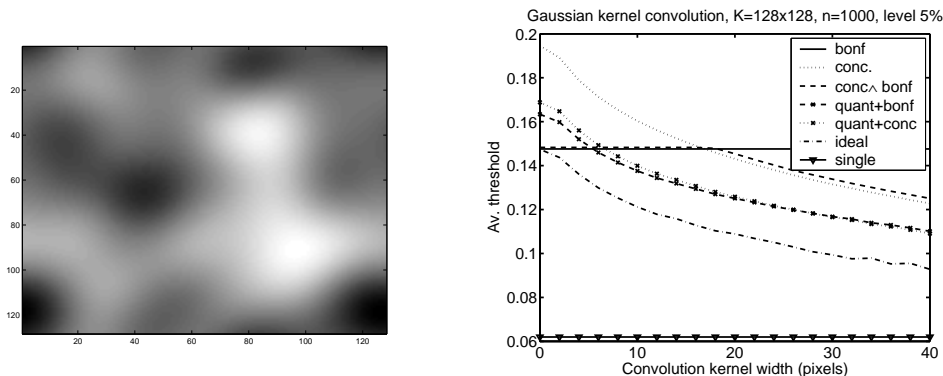


FIG 1. *Left: example of a 128x128 pixel image obtained by convolution of Gaussian white noise with a (toroidal) Gaussian filter with width $b = 18$ pixels. Right: average thresholds obtained for the different approaches, see text.*

In the simulations below we consider for the function F a “pseudo Gaussian” convolution filter of bandwidth b on the torus:

$$F_b(t) = C_b \exp\left(-d(0, t)^2/b^2\right) ,$$

where $d(t, t')$ is the standard distance on the torus and C_b is a normalizing constant. Note that for actual simulations it is more convenient to work in the Fourier domain and to apply the inverse DFT which can be computed efficiently. We then compare the different thresholds obtained by the methods proposed in this work for varying values of b . Remember that the only information available to our algorithms is the bound on the marginal variance; the form of the function F_b itself is of course unknown.

5.1. *Confidence balls.* On Fig 1 we compare the thresholds obtained when $\phi = \|\cdot\|_\infty$, which corresponds to L^∞ confidence balls. Remember that these thresholds can be also directly used in the two-sided multiple testing situation (see Section 4). We use the different approaches proposed in this work, with the following parameters: the dimension is $K = 128^2 = 16384$, the number of data points per sample is $n = 1000$ (much smaller than K , so that we really are in a non-asymptotic framework), the width b takes even values in the range $[0, 40]$, the overall level is $\alpha = 0.05$.

Recall that the Bonferroni threshold is

$$t'_{\text{Bonf}, \alpha} := \frac{1}{\sqrt{n}} \|\sigma\|_\infty \overline{\Phi}^{-1}\left(\frac{\alpha}{2K}\right) .$$

For the concentration threshold (7)

$$t_{\text{conc}, \alpha}(\mathbf{Y}) := \frac{\mathbb{E} \left[\left\| \overline{\mathbf{Y}}^{\langle W - \overline{W} \rangle} \right\|_\infty \mid \mathbf{Y} \right]}{B_W} + \|\sigma\|_\infty \overline{\Phi}^{-1}(\alpha/2) \left[\frac{C_W}{nB_W} + \frac{1}{\sqrt{n}} \right] ,$$

we used Rademacher weights. For the ‘‘compound’’ threshold of Corollary 2.2 (with the Bonferroni threshold as deterministic reference threshold)

$$t_{\text{conc} \wedge \text{Bonf}, \alpha}(\mathbf{Y}) := \min \left\{ t'_{\text{Bonf}, \alpha}, \frac{\mathbb{E} \left[\left\| \overline{\mathbf{Y}}^{\langle W - \overline{W} \rangle} \right\|_{\infty} \mid \mathbf{Y} \right]}{B_W} + \frac{\|\sigma\|_{\infty}}{\sqrt{n}} \overline{\Phi}^{-1} \left(\frac{\alpha(1-\delta)}{2} \right) + \frac{\|\sigma\|_{\infty} C_W}{n B_W} \overline{\Phi}^{-1} \left(\frac{\alpha\delta}{2} \right) \right\},$$

we used $\delta = 0.1$. For the quantile approach of Theorem 3.2, we considered the two variants

$$t_{\text{quant} + \text{Bonf}, \alpha}(\mathbf{Y}) := q_{\alpha_0(1-\delta)} \left(\|\cdot\|_{\infty}, \mathbf{Y} - \overline{\mathbf{Y}} \right) + \frac{2\overline{\mathcal{B}} \left(n, \frac{\alpha_0\delta}{2} \right) - n}{n} t'_{\text{Bonf}, \alpha - \alpha_0}$$

$$t_{\text{quant} + \text{conc}, \alpha}(\mathbf{Y}) := q_{\alpha_0(1-\delta)} \left(\|\cdot\|_{\infty}, \mathbf{Y} - \overline{\mathbf{Y}} \right) + \frac{2\overline{\mathcal{B}} \left(n, \frac{\alpha_0\delta}{2} \right) - n}{n} t_{\text{conc}, \alpha - \alpha_0}(\mathbf{Y}),$$

where we used $\alpha_0 = 0.9\alpha$ ($= (1 - \gamma)\alpha$, with $\gamma = 0.1$), $\delta = 0.1$ and took f either equal to the Bonferroni or the concentration threshold, respectively (these values of $\alpha_0, \alpha, \gamma, \delta$ will stay unchanged for all the experiments presented here, including in the next section). Finally, for comparison purposes, we included in the figure the threshold corresponding to $K = 1$ (estimation of a single coordinate mean)

$$t_{\text{single}, \alpha} := \frac{1}{\sqrt{n}} \|\sigma\|_{\infty} \overline{\Phi}^{-1} \left(\frac{\alpha}{2} \right).$$

We also included an estimation of the true quantile (actually, an empirical quantile over 1 000 samples), *i.e.* $t_{\text{ideal}, \alpha}$ the $1 - \alpha$ quantile of the distribution of $\left\| \overline{\mathbf{Y}} - \mu \right\|_{\infty}$.

Each point represents an average over 50 experiments (except of course for $t'_{\text{Bonf}, \alpha}$ and $t_{\text{single}, \alpha}$). The quantiles or expectations with respect to Rademacher weights were estimated by Monte-Carlo with 1 000 draws (without the additional terms introduced in Section 2.5.2 and Section 3.2). On the figure, we did not include standard deviations. They are quite low, of the order of 10^{-3} , although it is worth noting that the quantile threshold has a standard deviation roughly twice as large as the concentration threshold (we did not investigate at this point what part of this variation is due to the MC approximation).

We also computed the quantile threshold $q_{\alpha}(\|\cdot\|_{\infty}, \mathbf{Y} - \overline{\mathbf{Y}})$ without second-order term: it is so close to $t_{\text{ideal}, \alpha}$ that they would be almost indistinguishable on Fig 1.

The overall conclusion of this first experiment is that the different thresholds proposed in this work are relevant in the sense that they are smaller than the Bonferroni threshold provided the vector has strong enough correlations. As expected, the quantile approach appears to lead to tighter thresholds. (However, this might not be always the case for smaller sample sizes because of the additional term ε' .) One advantage of the concentration approach is that the ‘compound’ threshold can ‘‘fall back’’ on the Bonferroni threshold when needed, at the price of a minimal threshold increase.

5.2. *Multiple testing.* We now focus on the multiple testing problem. We present here only the two-sided case because the one-sided case gives similar results, except that we can not use the “uncentered quantile” method of Corollary 4.7.

We consider the experiment of the previous section, with the following choice for the vector of means:

$$(29) \quad \forall (i, j) \in \{0, \dots, 127\}^2, \quad \mu_{(i,j)} = \frac{(64 - j)_+}{64} \times 20t'_{\text{Bonf},\alpha} .$$

In this situation, note that the half of the null hypotheses are true while the non-zero means are increasing linearly from $(5/16)t'_{\text{Bonf},\alpha}$ to $20t'_{\text{Bonf},\alpha}$. The thresholds obtained are given on Fig 2 (100 simulations). The ideal threshold $t_{\text{ideal},\alpha}$ is now derived from the $1 - \alpha$ quantile of the distribution of $T'(\mathcal{H}_0) = \sup_{\mathcal{H}_0} |\bar{\mathbf{Y}}|$. We did not report $t_{\text{conc},\alpha}$ and $t_{\text{conc} \wedge \text{Bonf},\alpha}$ in order to simplify Fig 2. In addition to the previous thresholds, we considered:

- the uncentered quantile defined by:

$$t_{\text{quant.uncent.},\alpha}(\mathbf{Y}) := q_\alpha(\|\cdot\|_\infty, \mathbf{Y}),$$

and its step down version $t_{\text{s.d.quant.uncent.},\alpha}(\mathbf{Y})$ (see Corollary 4.7).

- the step down version $t_{\text{s.d.quant+Bonf},\alpha}(\mathbf{Y})$ of $t_{\text{quant+Bonf},\alpha}(\mathbf{Y})$.
- Holm’s threshold $t_{\text{Holm},\alpha}(\mathbf{Y})$ (*i.e.* the step-down version of Bonferroni’s procedure).

On the right-hand-side of Fig 2, we evaluated the power of the different thresholds $t_\alpha(\mathbf{Y})$, defined as the expected proportion of signal correctly detected (*i.e.* expected proportion of rejections among the false null hypotheses):

$$(30) \quad \text{Power}(t_\alpha(\mathbf{Y})) := \mathbb{E} \left(\frac{|\{1 \leq k \leq K \mid \mu_k \neq 0 \text{ and } |\mathbf{Y}_k| > t_\alpha(\mathbf{Y})\}|}{|\{1 \leq k \leq K \mid \mu_k \neq 0\}|} \right) .$$

For single-step resampling-based procedures, the results of the experiment lead us to conclude the following:

- the single-step procedure based on our quantile approach (“quant+Bonf”) can outperform Holm’s procedure as soon as the the coordinates of the vector are sufficiently correlated.
- the single-step procedure based on the uncentered quantile (“quant. uncent”) has bad performance.

For step-down resampling-based procedures, we draw the following conclusions:

- the step-down procedure based on our quantile approach (“s.d. quant+Bonf”) can outperform Holm’s procedure as soon as the coordinates of the vector are sufficiently correlated (obvious from the point 1).

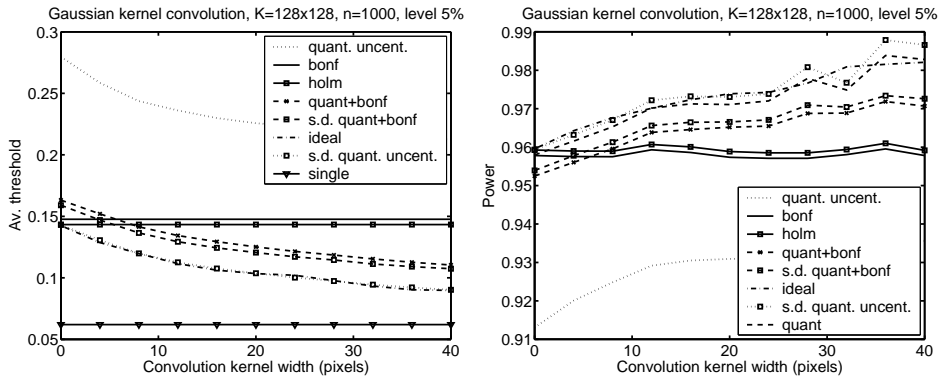


FIG 2. Multiple testing problem with μ defined by (29) for different approaches, see text. Left: average thresholds. Right: power, defined by (30).

- the step-down procedure based on the uncentered quantile (“s.d. quant uncent.”) seems to be the most accurate threshold of the step-down procedures considered here.

However, when K and n are large, each iteration of the step-down algorithm for the uncentered quantiles may be quite long to compute while our quantile approach (“quant+Bonf”) provides in only one step a quite good accuracy. Following Section 4.4, these two methods can be combined (see Algorithm 4.8, called here “mixed approach”), resulting in a speed-accuracy trade-off.

We illustrate this with a specific simulation study. Consider the same simulation framework as above except that the bandwidth b is now fixed at 30, the size of the sample is $n = 100$, and the means are given by: $\forall (i, j) \in \{0, \dots, 127\}^2$, $\mu_{(i,j)} = f(i + 128j)$, where

$$(31) \quad \forall k \in \{0, \dots, 8192\}, \quad f(k) = 50t'_{\text{Bonf},\alpha} \times \exp\left(-\frac{(8192 - k)_+}{8192} \log(100)\right),$$

and $f(k) = 0$ for the other values of k . In this situation, the non-zero means are increasing log-linearly from $0.5 t'_{\text{Bonf},\alpha}$ to $50 t'_{\text{Bonf},\alpha}$. With 100 simulations, we computed in Table 3 the average number of iterations for the above step-down procedures. Additionally, on Fig 3, the power is given as a function of the number of iterations. We can read the following results:

- The “mixed approach” needs on average significantly less iterations to converge.
- In the case of a very strict computation time constraint, it is possible to stop the step-down procedures early after a fixed number of iterations. Stopping the mixed approach procedure after only 2 iterations results in an average power that is virtually undistinguishable from the power obtained without early stopping. By contrast 3 iterations are needed for the step-down with uncentered quantile threshold.

While these results are certainly specific to the particular simulation setup we used, they illustrate that the informal and qualitative analysis we presented in Section 4.4 appears to be

correct. In particular, the fact that the mixed approach appears to give already very satisfactory results after the two first iterations reinforces the interpretation that the first step (using the recentered quantile threshold with remainder term) rules out at once all coordinates with a large SNR while the second step (using the exact, uncentered quantile) improves the precision once these high-SNR coordinates have been eliminated.

Therefore, this mixed approach can be an interesting alternative to the uncentered quantile approach when several long iterations in the step-down algorithm are expected. This situation arises typically when the signal (non-zero means) has a wide dynamic range (in our above simulation, the signal-to-noise ratio for non-true null hypotheses had a dynamic range of 100 or 20dB).

Holm's procedure	"s.d. quant+Bonf"	"s.d. quant. uncent."	"mixed approach"
3.25	3.13	4.92	3.94

TABLE 3

Multiple testing problem with μ corresponding to (31) for different step-down approaches. Average number of iterations in the step-down algorithm.

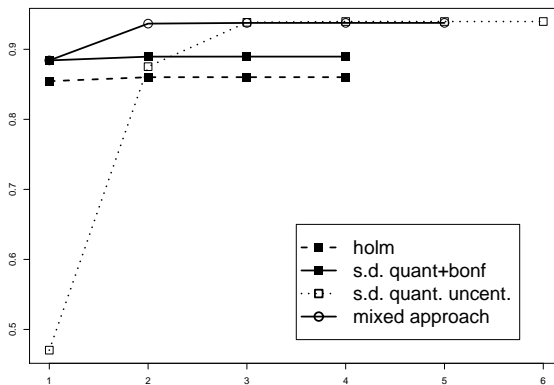


FIG 3. Multiple testing problem with μ corresponding to (31) for different step-down approaches. Power in function of the number of iterations in the step-down algorithm.

6. Discussions and concluding remarks.

6.1. *Discussion: confidence regions and tests.* In this paper we have first constructed confidence regions of the form (1) and presented an application of this result to (multiple) testing. Because of the duality between confidence regions and tests, a natural question is whether conversely, one could construct tests first and deduce confidence regions. In particular, testing the (single) null hypothesis $H_{\mu_0} : \mu = \mu_0$ is very simple using an exact symmetrization test: using directly Lemma 3.1 we know that the test $T_{\mu_0, \phi}$ rejecting H_{μ_0} if $\phi(\mathbf{Y} - \mu_0) > q_\alpha(\phi, \mathbf{Y} - \mu_0)$ has

significance level bounded by α . We can construct from this the confidence region

$$\mathcal{F}_\phi(\mathbf{Y}, 1 - \alpha) = \left\{ \mu_0 \in \mathbb{R}^K : T_{\mu_0, \phi} \text{ does not reject } H_{\mu_0} \right\}.$$

This method avoids completely the problems linked to the direct construction of a confidence region that we faced in Section 3; furthermore, the above confidence region is almost exactly of level $1 - \alpha$ (up to 2^{-n}), while the region constructed in Section 3 is certainly more conservative. Nevertheless, argue that the approach developed in Section 3 is much more practically relevant:

- the region $\mathcal{F}_\phi(\mathbf{Y}, 1 - \alpha)$ constructed above by test inversion is not of the form (1), that is, it is not a “ ϕ -ball” around the empirical mean. However, it might be required by external constraints, for example for further analysis, that the confidence region should be of this form.
- more generally, it does not seem clear at all what shape the above region would take or even if would enjoy some desirable properties such as convexity. This seems very impractical, particularly in high dimension, where regions which cannot be described under a simple form seem very difficult to handle. In fact, it seems actually very difficult to obtain any explicit description of this region short of calculating $T_{\mu_0, \phi}$ for every point μ_0 on a discretized grid of \mathbb{R}^K , which becomes intractable for both computational burden and memory usage as soon as K is large.

6.2. Discussion: FWER versus FDR in multiple testing. It can legitimately be asked if the FWER is in fact an appropriate measure of type I error. Namely, the false discovery rate (FDR), introduced in [3] and defined as the average proportion of wrongly rejected hypotheses among all the rejected hypotheses, appears to have recently become a *de facto* standard, in particular in the setting of a large number of hypotheses to test as we consider here. One reason for the popularity of FDR is that it is a less strict measure of error as the FWER and to this extent, FDR-controlled procedures reject more hypotheses than FWER-controlled ones. We give two reasons why the FWER is still a quantity of interest to investigate. First, the FDR is not always relevant, in particular for neuroimaging data. Indeed, in this context the signal is often strong over some well-known large areas of the brain (*e.g.* the motor and visual cortex). Therefore, if for instance 95 percent of the detected locations belong to these well-known areas, FDR control (at level 5%) does not provide evidence for any new true discovery. On the contrary, FWER control is more conservative, but each detected location outside these well-known areas is a new true discovery with high probability. Secondly, assuming the FDR or a related quantity is nevertheless the endgoal, it can be very useful to consider a two-step procedure, where the first step consists in a FWER-controlled multiple test. Namely, this first step can be used as a means to estimate the FDR or the FDP (false discovery proportion) of another procedure used in the second step and thus fine-tune the parameters of this second step for the desired goal. This approach has been for example advocated in [19] with application to neuroimaging data as well.

6.3. *Discussion: about the variances of the coordinates.* In the concentration approach and in the Gaussian case, the derived thresholds depend explicitly on the p -norm of the vector of standard deviations $\sigma = (\sigma_k)_k$ (an upper bound on this quantity can be used as well). While we have left aside the problem of determining this parameter if no prior information is available, there is at least a simple solution available: build (using standard techniques) an individual upper confidence bound for each σ_k , then combine these different confidence bounds with the Bonferroni method. While this naive method will not take into account the possible dependence between the coordinates for the estimation of σ itself, it will generally only contribute a lower order term in the final threshold defined by (7).

A second and potentially more crucial problem is that, since the confidence regions proposed in this paper are balls rather than ellipsoids, these regions will — inevitably — be conservative when the variances of the coordinates are very different. The standard way to address this issue is to consider studentized data. While this would solve this heteroscedasticity issue, it also voids the assumption of independent datapoints — a crucial assumption in all of our proofs. Therefore, generalizing our approach to studentized observations is an important (and probably challenging) direction for future research.

6.4. *Conclusion.* In this paper, we proposed two approaches to build non-asymptotic resampling-based confidence regions for a correlated random vector:

- The first one is strongly inspired by results coming from learning theory and is based on a concentration argument. An advantage of this method is that it allows to use a very large class of resampling weights. However, these concentration-based thresholds have relatively conservative deviation terms and they are better than the Bonferroni threshold only if there are very strong correlations in the data. Therefore, using this method when we do not have any prior knowledge on the correlations can be too risky. To address this issue, we propose (under the Gaussian assumption) to combine the corresponding concentration threshold with the Bonferroni threshold to obtain a threshold very close to the minimum of the two (using the so-called “stabilization property” of the resampling).
- The second method is closer to the idea of randomization tests: it estimates directly the quantile of $\phi(\bar{\mathbf{Y}} - \mu)$ using a symmetrization argument (it is therefore restricted to Rademacher weights). The point is that an exact approach is not possible because we have to replace the unknown parameter μ by the empirical mean $\bar{\mathbf{Y}}$. Therefore, the derived thresholds have a remainder term, but it is quite small when n is sufficiently large (typically $n \geq 1000$).

Our simulations have shown that for confidence regions in supremum norm, the confidence balls obtained with the second method are better than the regions based on the Bonferroni threshold, when there are important correlations between the coordinates. Moreover, it seems that the quantile threshold without the remainder term is very close to the ideal quantile, so that we may conjecture that the additional term is unnecessary (or at least too large).

Finally, we have used the two previous methods to derive step-down multiple testing procedures that control the FWER when testing simultaneously the means of a (Gaussian) random vector (in the one-sided or two-sided context). Because these procedures use translation-invariant thresholds, the number of iterations in the step-down algorithm is generally small. Moreover, they can outperform Holm's procedure when the coordinates of the observed vector has strong enough correlations. However, these procedures are somewhat too conservative because of the remainder terms (in the quantile approach, the remainder terms arise as a consequence of empirically recentering the data).

In the two-sided context, an exact step-down procedure based on the resampled quantiles of the *uncentered* data is valid and turns out to be more accurate than the above methods (because no remainder term is then necessary). However, this exact method needs generally more iterations in the step-down algorithm. Therefore, we propose to combine our quantile approach with the latter exact method to get a faster procedure with (almost) the same accuracy.

Again, we may conjecture that the step-down procedure using the recentered quantile without the additional term (or at least with a smaller term) still controls the FWER for a fixed n . This would give an accurate procedure in both two-sided and one-sided contexts, and the latter would be faster than the exact step-down procedure in the two-sided context. This is certainly an interesting direction for future work.

7. Proofs.

7.1. *Confidence regions using concentration.* In this section, we prove all the statements of Section 2 except computations of resampling weight constants (made in Section 7.4) and statements with non-exchangeable resampling weights (made in Section 7.5).

7.1.1. *Comparison in expectation.*

Proof of Proposition 2.4. Denoting by Σ the common covariance matrix of the \mathbf{Y}^i , we have $\mathcal{D}\left(\bar{\mathbf{Y}}^{\langle W-\bar{W} \rangle} \mid W\right) = \mathcal{N}\left(0, (n^{-1} \sum_{i=1}^n (W_i - \bar{W})^2) n^{-1} \Sigma\right)$, and the result follows because $\mathcal{D}(\bar{\mathbf{Y}} - \mu) = \mathcal{N}(0, n^{-1} \Sigma)$ and ϕ is positive-homogeneous. ■

Proof of Proposition 2.6. By independence between W and \mathbf{Y} , exchangeability of W and the positive homogeneity of ϕ , for every realization of \mathbf{Y} we have:

$$A_W \phi(\bar{\mathbf{Y}} - \mu) = \phi\left(\mathbb{E}\left[\frac{1}{n} \sum_{i=1}^n |W_i - \bar{W}| (\mathbf{Y}^i - \mu) \mid \mathbf{Y}\right]\right).$$

Then, by convexity of ϕ ,

$$A_W \phi(\bar{\mathbf{Y}} - \mu) \leq \mathbb{E} \left[\phi \left(\frac{1}{n} \sum_{i=1}^n |W_i - \bar{W}| (\mathbf{Y}^i - \mu) \right) \middle| \mathbf{Y} \right].$$

We integrate with respect to \mathbf{Y} , and use the symmetry of the \mathbf{Y}^i with respect to μ and again the independence between W and \mathbf{Y} to show finally that

$$\begin{aligned} A_W \mathbb{E} \left[\phi(\bar{\mathbf{Y}} - \mu) \right] &\leq \mathbb{E} \left[\phi \left(\frac{1}{n} \sum_{i=1}^n |W_i - \bar{W}| (\mathbf{Y}^i - \mu) \right) \right] \\ &= \mathbb{E} \left[\phi \left(\frac{1}{n} \sum_{i=1}^n (W_i - \bar{W}) (\mathbf{Y}^i - \mu) \right) \right] = \mathbb{E} \left[\phi \left(\bar{\mathbf{Y}}^{\langle W - \bar{W} \rangle} \right) \right]. \end{aligned}$$

The point (ii) comes from :

$$\begin{aligned} \mathbb{E} \left[\phi \left(\bar{\mathbf{Y}}^{\langle W - \bar{W} \rangle} \right) \right] &= \mathbb{E} \left[\phi \left(\frac{1}{n} \sum_{i=1}^n (W_i - \bar{W}) (\mathbf{Y}^i - \mu) \right) \right] \\ &\leq \mathbb{E} \left[\phi \left(\frac{1}{n} \sum_{i=1}^n (W_i - x_0) (\mathbf{Y}^i - \mu) \right) \right] + \mathbb{E} \left[\phi \left(\frac{1}{n} \sum_{i=1}^n (x_0 - \bar{W}) (\mathbf{Y}^i - \mu) \right) \right]. \end{aligned}$$

Then, by symmetry of the \mathbf{Y}^i with respect to μ and independence between W and \mathbf{Y} , we get

$$\begin{aligned} \mathbb{E} \left[\phi \left(\bar{\mathbf{Y}}^{\langle W - \bar{W} \rangle} \right) \right] &\leq \mathbb{E} \left[\phi \left(\frac{1}{n} \sum_{i=1}^n |W_i - x_0| (\mathbf{Y}^i - \mu) \right) \right] + \mathbb{E} \left[\phi \left(\frac{1}{n} \sum_{i=1}^n |x_0 - \bar{W}| (\mathbf{Y}^i - \mu) \right) \right] \\ &\leq (a + \mathbb{E}|\bar{W} - x_0|) \mathbb{E} \left[\phi(\bar{\mathbf{Y}} - \mu) \right]. \end{aligned}$$

■

7.1.2. Concentration inequalities.

Proof of Proposition 2.8. We use here concentration principles applied to a supremum of Gaussian random vectors, following closely the approach in [16], Section 3.2.4. The essential ingredient is the Gaussian concentration theorem of Cirel'son, Ibragimov and Sudakov ([5] and recalled in [16], Theorem 3.8), stating that if F is a Lipschitz function on \mathbb{R}^N with constant L , then for the standard Gaussian measure on \mathbb{R}^N we have $\mathbb{P}[F \geq \mathbb{E}[F] + t] \leq 2\bar{\Phi}(t/L)$.

Let us denote by \mathbf{A} a square root of the common covariance matrix of the \mathbf{Y}^i . If \mathbf{G} is a $K \times n$ matrix with standard centered i.i.d. Gaussian entries, then $\mathbf{A}\mathbf{G}$ has the same distribution as $\mathbf{Y} - \mu$. We let for all $\zeta \in (\mathbb{R}^K)^n$, $T_1(\zeta) := \phi \left(\frac{1}{n} \sum_{i=1}^n \mathbf{A}\zeta_i \right)$ and $T_2(\zeta) := \mathbb{E} \left[\phi \left(\frac{1}{n} \sum_{i=1}^n (W_i - \bar{W}) \mathbf{A}\zeta_i \right) \right]$.

From the Gaussian concentration theorem recalled above, to reach the conclusion we just need to prove that T_1 (resp. T_2) is a Lipschitz function with constant $\|\sigma\|_p / \sqrt{n}$ (resp. $\|\sigma\|_p C_W/n$) with

respect to the Euclidean norm $\|\cdot\|_{2,Kn}$ on $(\mathbb{R}^K)^n$. Let $\zeta, \zeta' \in (\mathbb{R}^K)^n$ and denote by $(a_k)_{1 \leq k \leq K}$ the rows of \mathbf{A} . Using that ϕ is 1-Lipschitz with respect to the p -norm (because it is subadditive and bounded by the p -norm), we get

$$\begin{aligned} |T_1(\zeta) - T_1(\zeta')| &\leq \left\| \frac{1}{n} \sum_{i=1}^n \mathbf{A}(\zeta_i - \zeta'_i) \right\|_p \\ &\leq \left\| \left(\left\langle a_k, \frac{1}{n} \sum_{i=1}^n (\zeta_i - \zeta'_i) \right\rangle \right)_k \right\|_p. \end{aligned}$$

For each coordinate k , by Cauchy-Schwartz's inequality and since $\|a_k\|_2 = \sigma_k$, we deduce

$$\left| \left\langle a_k, \frac{1}{n} \sum_{i=1}^n (\zeta_i - \zeta'_i) \right\rangle \right| \leq \sigma_k \left\| \frac{1}{n} \sum_{i=1}^n (\zeta_i - \zeta'_i) \right\|_2.$$

Therefore, we get

$$\begin{aligned} |T_1(\zeta) - T_1(\zeta')| &\leq \|\sigma\|_p \left\| \frac{1}{n} \sum_{i=1}^n (\zeta_i - \zeta'_i) \right\|_2 \\ &\leq \frac{\|\sigma\|_p}{\sqrt{n}} \|\zeta - \zeta'\|_{2,Kn}, \end{aligned}$$

using the convexity of $x \in \mathbb{R}^K \mapsto \|x\|_2^2$, and we obtain (i). For T_2 , we use the same method as for T_1 :

$$\begin{aligned} |T_2(\zeta) - T_2(\zeta')| &\leq \|\sigma\|_p \mathbb{E} \left\| \frac{1}{n} \sum_{i=1}^n (W_i - \overline{W})(\zeta_i - \zeta'_i) \right\|_2 \\ (32) \qquad \qquad \qquad &\leq \frac{\|\sigma\|_p}{n} \sqrt{\mathbb{E} \left\| \sum_{i=1}^n (W_i - \overline{W})(\zeta_i - \zeta'_i) \right\|_2^2}. \end{aligned}$$

Note that since $\left(\sum_{i=1}^n (W_i - \overline{W}) \right)^2 = 0$, we have $\mathbb{E}(W_1 - \overline{W})(W_2 - \overline{W}) = -C_W^2/n$. We now develop $\left\| \sum_{i=1}^n (W_i - \overline{W})(\zeta_i - \zeta'_i) \right\|_2^2$ in the Euclidean space \mathbb{R}^K :

$$\begin{aligned} \mathbb{E} \left\| \sum_{i=1}^n (W_i - \overline{W})(\zeta_i - \zeta'_i) \right\|_2^2 &= C_W^2 (1 - n^{-1}) \sum_{i=1}^n \|\zeta_i - \zeta'_i\|_2^2 - \frac{C_W^2}{n} \sum_{i \neq j} \langle \zeta_i - \zeta'_i, \zeta_j - \zeta'_j \rangle \\ &= C_W^2 \sum_{i=1}^n \|\zeta_i - \zeta'_i\|_2^2 - \frac{C_W^2}{n} \left\| \sum_{i=1}^n (\zeta_i - \zeta'_i) \right\|_2^2. \end{aligned}$$

Consequently,

$$(33) \quad \mathbb{E} \left\| \sum_{i=1}^n (W_i - \overline{W}) (\zeta_i - \zeta'_i) \right\|_2^2 \leq C_W^2 \sum_{i=1}^n \|\zeta_i - \zeta'_i\|_2^2 = C_W^2 \|\zeta - \zeta'\|_{2,Kn}^2.$$

Combining expression (32) and (33), we find that T_2 is $\|\sigma\|_p C_W/n$ -Lipschitz. \blacksquare

Remark 7.1 *The proof of Proposition 2.8 is still valid under the weaker assumption (instead of exchangeability of W) that $\mathbb{E}[(W_i - \overline{W})(W_j - \overline{W})]$ can only take two possible values depending on whether or not $i = j$.*

7.1.3. Main results.

Proof of Theorem 2.1. The case (BA)(p, M) and (SA) is obtained by combining Proposition 2.6 and McDiarmid's inequality (see for instance [9]). The (GA) case is a straightforward consequence of Proposition 2.4 and the proof of Proposition 2.8 (considering the Lipschitz function $T_1 - T_2$). \blacksquare

Proof of Corollary 2.2. From Proposition 2.8 (i), with probability at least $1 - \alpha(1 - \delta)$, $\phi(\overline{\mathbf{Y}} - \mu)$ is less than or equal to the minimum between $t_{\alpha(1-\delta)}$ and $\mathbb{E}[\phi(\overline{\mathbf{Y}} - \mu)] + \frac{\|\sigma\|_p \overline{\Phi}^{-1}(\alpha(1-\delta)/2)}{\sqrt{n}}$ (because these thresholds are deterministic). In addition, Proposition 2.4 and Proposition 2.8 (ii) give that with probability at least $1 - \alpha\delta$, $\mathbb{E}[\phi(\overline{\mathbf{Y}} - \mu)] \leq \frac{\mathbb{E}_W[\phi(\overline{\mathbf{Y}}^{\langle W - \overline{W} \rangle})]}{B_W} + \frac{\|\sigma\|_p C_W}{B_W n} \overline{\Phi}^{-1}(\alpha\delta/2)$. The result follows by combining the two last expressions. \blacksquare

7.1.4. Monte-Carlo approximation.

Proof of Proposition 2.11. The idea of the proof is to apply McDiarmid's inequality conditionally to \mathbf{Y} (see [17]). For any realizations W and W' of the resampling weight vector and any $\nu \in \mathbb{R}^k$,

$$\begin{aligned} \left| \phi\left(\overline{\mathbf{Y}}^{\langle W - \overline{W} \rangle}\right) - \phi\left(\overline{\mathbf{Y}}^{\langle W' - \overline{W}' \rangle}\right) \right| &\leq \phi\left(\overline{\mathbf{Y}}^{\langle W - \overline{W} \rangle} - \overline{\mathbf{Y}}^{\langle W' - \overline{W}' \rangle}\right) \\ &\leq \frac{c_2 - c_1}{n} \left\| \left(\sum_{i=1}^n |\mathbf{Y}_k^i - \nu_k| \right)_k \right\|_p \end{aligned}$$

since ϕ is sub-additive, bounded by the p -norm and $W_i - \overline{W} \in [c_1, c_2]$ a.s.

The sample \mathbf{Y} being deterministic, we can take ν equal to the median M of the sample, which realizes the infimum. Since W^1, \dots, W^B are independent, McDiarmid's inequality gives (18).

When \mathbf{Y} satisfies (GA), a proof very similar to the one of (15) in Proposition 2.8 can be applied to the remainder term with any deterministic ν . We then obtain (19). \blacksquare

7.2. *Quantiles.* Remember the following inequality coming from the definition of the quantile q_α : for any fixed \mathbf{Y}

$$(34) \quad \mathbb{P}_W \left[\phi \left(\overline{\mathbf{Y}}^{(W)} \right) > q_\alpha(\phi, \mathbf{Y}) \right] \leq \alpha \leq \mathbb{P}_W \left[\phi \left(\overline{\mathbf{Y}}^{(W)} \right) \geq q_\alpha(\phi, \mathbf{Y}) \right].$$

Proof of Lemma 3.1. We introduce the notation $\mathbf{Y} \bullet W = \mathbf{Y} \cdot \text{diag}(W)$ for the matrix obtained by multiplying the i -th column of \mathbf{Y} by W_i , $i = 1, \dots, n$.

We have

$$(35) \quad \begin{aligned} \mathbb{P}_{\mathbf{Y}} \left[\phi(\overline{\mathbf{Y}} - \mu) > q_\alpha(\phi, \mathbf{Y} - \mu) \right] &= \mathbb{E}_W \left[\mathbb{P}_{\mathbf{Y}} \left[\phi \left(\overline{(\mathbf{Y} - \mu)}^{(W)} \right) > q_\alpha(\phi, (\mathbf{Y} - \mu) \bullet W) \right] \right] \\ &= \mathbb{E}_{\mathbf{Y}} \left[\mathbb{P}_W \left[\phi \left(\overline{(\mathbf{Y} - \mu)}^{(W)} \right) > q_\alpha(\phi, \mathbf{Y} - \mu) \right] \right] \\ &\leq \alpha. \end{aligned}$$

The first equality is due to the fact that the distribution of \mathbf{Y} satisfies assumption (SA), hence the distribution of $(\mathbf{Y} - \mu)$ invariant by reweighting by (arbitrary) signs $W \in \{-1, 1\}^n$. In the second equality we used Fubini's theorem and the fact that for any arbitrary signs W as above $q_\alpha(\phi, (\mathbf{Y} - \mu) \bullet W) = q_\alpha(\phi, \mathbf{Y} - \mu)$; finally the last inequality comes from (34). \blacksquare

Proof of Theorem 3.2. Put $\gamma_1 = \gamma_1(\alpha_0 \delta)$ for short and define the event

$$\Omega = \left\{ \mathbf{Y} \mid q_{\alpha_0}(\phi, \mathbf{Y} - \mu) \leq q_{\alpha_0(1-\delta)}(\phi, \mathbf{Y} - \overline{\mathbf{Y}}) + \gamma_1 f(\mathbf{Y}) \right\};$$

then we have using (35) :

$$(36) \quad \begin{aligned} \mathbb{P} \left[\phi(\overline{\mathbf{Y}} - \mu) > q_{\alpha_0(1-\delta)}(\phi, \mathbf{Y} - \overline{\mathbf{Y}}) + \gamma_1 f(\mathbf{Y}) \right] \\ \leq \mathbb{P} \left[\phi(\overline{\mathbf{Y}} - \mu) > q_{\alpha_0}(\phi, \mathbf{Y} - \mu) \right] + \mathbb{P}[\mathbf{Y} \in \Omega^c] \\ \leq \alpha_0 + \mathbb{P}[\mathbf{Y} \in \Omega^c]. \end{aligned}$$

We now concentrate on the event Ω^c . Using the subadditivity of ϕ , and the fact that $\overline{(\mathbf{Y} - \mu)}^{(W)} = \overline{(\mathbf{Y} - \overline{\mathbf{Y}})}^{(W)} + \overline{W}(\overline{\mathbf{Y}} - \mu)$, we have for any fixed $\mathbf{Y} \in \Omega^c$:

$$\begin{aligned} \alpha_0 &\leq \mathbb{P}_W \left[\phi \left(\overline{(\mathbf{Y} - \mu)}^{(W)} \right) \geq q_{\alpha_0}(\phi, \mathbf{Y} - \mu) \right] \\ &\leq \mathbb{P}_W \left[\phi \left(\overline{(\mathbf{Y} - \mu)}^{(W)} \right) > q_{\alpha_0(1-\delta)}(\phi, \mathbf{Y} - \overline{\mathbf{Y}}) + \gamma_1 f(\mathbf{Y}) \right] \\ &\leq \mathbb{P}_W \left[\phi \left(\overline{(\mathbf{Y} - \overline{\mathbf{Y}})}^{(W)} \right) > q_{\alpha_0(1-\delta)}(\phi, \mathbf{Y} - \overline{\mathbf{Y}}) \right] + \mathbb{P}_W \left[\phi(\overline{W}(\overline{\mathbf{Y}} - \mu)) > \gamma_1 f(\mathbf{Y}) \right] \\ &\leq \alpha_0(1 - \delta) + \mathbb{P}_W \left[\phi(\overline{W}(\overline{\mathbf{Y}} - \mu)) > \gamma_1 f(\mathbf{Y}) \right]. \end{aligned}$$

For the first and last inequalities we have used (34), and for the second inequality the definition of Ω^c . From this we deduce that

$$\Omega^c \subset \left\{ \mathbf{Y} \mid \mathbb{P}_W \left[\phi(\overline{W}(\overline{\mathbf{Y}} - \mu)) > \gamma_1 f(\mathbf{Y}) \right] \geq \alpha_0 \delta \right\} .$$

Now using the homogeneity of ϕ , and the fact that both ϕ and f are nonnegative:

$$\begin{aligned} \mathbb{P}_W \left[\phi(\overline{W}(\overline{\mathbf{Y}} - \mu)) > \gamma_1 f(\mathbf{Y}) \right] &= \mathbb{P}_W \left[\left| \overline{W} \right| > \frac{\gamma_1 f(\mathbf{Y})}{\phi(\text{sign}(\overline{W})(\overline{\mathbf{Y}} - \mu))} \right] \\ &\leq \mathbb{P}_W \left[\left| \overline{W} \right| > \frac{\gamma_1 f(\mathbf{Y})}{\tilde{\phi}(\overline{\mathbf{Y}} - \mu)} \right] \\ &= 2\mathbb{P}_W \left[\frac{1}{n} (2B_{n, \frac{1}{2}} - n) > \frac{\gamma_1 f(\mathbf{Y})}{\tilde{\phi}(\overline{\mathbf{Y}} - \mu)} \right] , \end{aligned}$$

where $B_{n, \frac{1}{2}}$ denotes a binomial $(n, \frac{1}{2})$ variable (independent of \mathbf{Y}). From the two last displays and the definition of γ_1 , we conclude

$$\Omega^c \subset \left\{ \mathbf{Y} \mid \tilde{\phi}(\overline{\mathbf{Y}} - \mu) > f(\mathbf{Y}) \right\} ,$$

which, put back in (36), leads to the desired conclusion. ■

Proof of Corollary 3.4. Define the function

$$g_0(\mathbf{Y}) = q_{(1-\delta)\alpha_0}(\phi, \mathbf{Y} - \overline{\mathbf{Y}}) + \left(\sum_{i=1}^{J-1} \gamma_i q_{(1-\delta)\alpha_i}(\tilde{\phi}, \mathbf{Y} - \overline{\mathbf{Y}}) + \gamma_J f(\mathbf{Y}) \right) ,$$

and for $k = 1, \dots, J$,

$$g_k(\mathbf{Y}) = \gamma_k^{-1} \left(\sum_{i=k}^{J-1} \gamma_i q_{(1-\delta)\alpha_i}(\tilde{\phi}, \mathbf{Y} - \overline{\mathbf{Y}}) + \gamma_J f(\mathbf{Y}) \right) ,$$

with the convention $g_J = f$. For $0 \leq k \leq J-1$, applying Theorem 3.2 with the function g_{k+1} yields the relation

$$\mathbb{P}_W \left[\phi(\overline{\mathbf{Y}} - \mu) > g_k(\mathbf{Y}) \right] \leq \alpha_k + \mathbb{P}_W \left[\phi(\overline{\mathbf{Y}} - \mu) > g_{k+1}(\mathbf{Y}) \right] .$$

Therefore,

$$\mathbb{P}_W \left[\phi(\overline{\mathbf{Y}} - \mu) > g_0(\mathbf{Y}) \right] \leq \sum_{i=0}^{J-1} \alpha_i + \mathbb{P} \left[\tilde{\phi}(\overline{\mathbf{Y}} - \mu) > f(\mathbf{Y}) \right] ,$$

as announced. ■

Proof of Proposition 3.5. Let us first prove that an analogue of Lemma 3.1 holds with q_{α_0} replaced by \tilde{q}_{α_0} . First, we have

$$\begin{aligned} \mathbb{E}_{\mathbf{W}} \mathbb{P}_{\mathbf{Y}} \left[\phi(\overline{\mathbf{Y}} - \mu) > \tilde{q}_{\alpha_0}(\phi, \mathbf{Y} - \mu, \mathbf{W}) \right] \\ &= \mathbb{E}_{W'} \mathbb{E}_{\mathbf{W}} \mathbb{P}_{\mathbf{Y}} \left[\phi(\overline{(\mathbf{Y} - \mu)^{\langle W' \rangle}}) > \tilde{q}_{\alpha_0}(\phi, (\mathbf{Y} - \mu) \bullet W', \mathbf{W}) \right] \\ &= \mathbb{E}_{\mathbf{Y}} \mathbb{P}_{\mathbf{W}, W'} \left[\phi(\overline{(\mathbf{Y} - \mu)^{\langle W' \rangle}}) > \tilde{q}_{\alpha_0}(\phi, \mathbf{Y} - \mu, W' \bullet \mathbf{W}) \right], \end{aligned}$$

where W' denotes a Rademacher vector independent of all other random variables and $W' \bullet \mathbf{W} = \text{diag}(W') \cdot \mathbf{W}$ denotes the matrix obtained by multiplying the i -th row of \mathbf{W} by W'_i , $i = 1, \dots, n$. Note that $(W', W' \bullet \mathbf{W}) \sim (W', \mathbf{W})$. Therefore, by definition of the quantile \tilde{q}_{α_0} , the latter quantity is equal to

$$\mathbb{E}_{\mathbf{Y}} \mathbb{P}_{\mathbf{W}, W'} \left[\frac{1}{B} \sum_{j=1}^B \mathbb{1} \left\{ \phi(\overline{(\mathbf{Y} - \mu)^{\langle \mathbf{W}^j \rangle}}) \geq \phi(\overline{(\mathbf{Y} - \mu)^{\langle W' \rangle}}) \right\} \leq \alpha_0 \right] \leq \frac{\lfloor B\alpha_0 \rfloor + 1}{B + 1},$$

where the last step comes from Lemma 7.2 taken from [26] (see below).

The rest of the proof is similar to the one of Theorem 3.2, where \mathbb{P}_W is replaced by the empirical distribution based on \mathbf{W} , $\tilde{\mathbb{P}}_W = \frac{1}{B} \sum_{j=1}^B \delta_{\mathbf{W}^j}$. Thus, (34) becomes for any fixed \mathbf{Y}, \mathbf{W} :

$$\tilde{\mathbb{P}}_W \left[\phi(\overline{\mathbf{Y}^{\langle W \rangle}}) > \tilde{q}_{\alpha_0}(\phi, \mathbf{Y}, \mathbf{W}) \right] \leq \alpha_0 \leq \tilde{\mathbb{P}}_W \left[\phi(\overline{\mathbf{Y}^{\langle W \rangle}}) \geq \tilde{q}_{\alpha_0}(\phi, \mathbf{Y}, \mathbf{W}) \right].$$

Then, the role of Ω is taken by

$$\tilde{\Omega} := \left\{ \mathbf{Y}, \mathbf{W} \mid \tilde{q}_{\alpha_0}(\phi, \mathbf{Y} - \mu, \mathbf{W}) \leq \tilde{q}_{\alpha_0(1-\delta)}(\phi, \mathbf{Y} - \overline{\mathbf{Y}}, \mathbf{W}) + \gamma f(\mathbf{Y}, \mathbf{W}) \right\},$$

where we put $\gamma = \gamma(\mathbf{W}, \alpha_0 \delta)$ for short. We then have similarly to (36):

$$\mathbb{P}_{\mathbf{Y}, \mathbf{W}} \left[\phi(\overline{\mathbf{Y}} - \mu) > \tilde{q}_{\alpha_0(1-\delta)}(\phi, \mathbf{Y} - \overline{\mathbf{Y}}) + \gamma f(\mathbf{Y}, \mathbf{W}) \right] \leq \frac{\lfloor B\alpha_0 \rfloor + 1}{B + 1} + \mathbb{P}_{\mathbf{Y}, \mathbf{W}} \left[\tilde{\Omega}^c \right],$$

and following further the proof of Theorem 3.2, we obtain

$$\tilde{\Omega}^c \subset \left\{ \mathbf{Y}, \mathbf{W} \mid \tilde{\mathbb{P}}_W \left[|\overline{W}| > \frac{\gamma f(\mathbf{Y}, \mathbf{W})}{\tilde{\phi}(\overline{\mathbf{Y}} - \mu)} \right] \geq \alpha_0 \delta \right\},$$

which gives the result. ■

We have used the following Lemma:

Lemma 7.2 (Essentially Lemma 1 of [26]) *Let Z_0, Z_1, \dots, Z_B be exchangeable real-valued random variables. Then for all $\alpha \in (0, 1)$,*

$$\mathbb{P} \left[\frac{1}{B} \sum_{j=1}^B \mathbf{1} \{Z_j \geq Z_0\} \leq \alpha \right] \leq \frac{\lfloor B\alpha \rfloor + 1}{B + 1} \leq \alpha + \frac{1}{B + 1}.$$

The first inequality becomes an equality if $Z_i \neq Z_j$ a.s. For example, it is the case if the Z_i s are i.i.d. variables from a distribution without atoms.

We provide a proof for completeness.

Proof of Lemma 7.2. Let U denote a random variable uniformly distributed in $\{0, \dots, B\}$ and independent of the Z_i s. We then have

$$\begin{aligned} \mathbb{P} \left[\frac{1}{B} \sum_{j=1}^B \mathbf{1} \{Z_j \geq Z_0\} \leq \alpha \right] &= \mathbb{P} \left[\sum_{j=0}^B \mathbf{1} \{Z_j \geq Z_0\} \leq B\alpha + 1 \right] \\ &= \mathbb{P}_U \mathbb{P}_{(Z_i)} \left[\sum_{j=0}^B \mathbf{1} \{Z_j \geq Z_U\} \leq B\alpha + 1 \right] \\ &= \mathbb{P}_{(Z_i)} \mathbb{P}_U \left[\sum_{j=0}^B \mathbf{1} \{Z_j \geq Z_U\} \leq \lfloor B\alpha \rfloor + 1 \right] \\ &\leq \frac{\lfloor B\alpha \rfloor + 1}{B + 1}. \end{aligned}$$

Note that the last inequality is an equality if the Z_i s are a.s. distinct. ■

7.3. Multiple testing.

Proof of Theorem 4.3. (from [26]) We use the notations of Definition 4.1. If the procedure rejects at least one true null hypothesis, we may consider $j_0 = \min\{j \leq \hat{\ell} \mid H_{\sigma(j)} \text{ is true}\}$. By definition of a step-down procedure, we have $\overline{\mathbf{Y}}_{\sigma(j_0)} \geq t_{j_0}$. By definition of j_0 , we have $\mathcal{H}_0 \subset \mathcal{C}_{j_0}$ so that, since \mathbf{t} is non-decreasing, $\mathbf{t}(\mathcal{C}_{j_0}) \geq \mathbf{t}(\mathcal{H}_0)$. Finally, we can obtain (28) as follows:

$$\begin{aligned} \text{FWER}(R) &\leq \mathbb{P} \left(\exists j_0 \mid H_{\sigma(j_0)} \text{ is true and } \overline{\mathbf{Y}}_{\sigma(j_0)} \geq \mathbf{t}(\mathcal{H}_0) \right) \\ &\leq \mathbb{P} (T'(H_0) \geq \mathbf{t}(\mathcal{H}_0)) \\ &\leq \mathbb{P} (T(H_0) \geq \mathbf{t}(\mathcal{H}_0)) . \end{aligned}$$
■

Proof of Proposition 4.9. First note that

$$q_{\alpha(1-\gamma)} \left(\sup_{\mathcal{H}_0} |\cdot|, \mathbf{Y} \right) \leq q_{\alpha(1-\gamma)}(\|\cdot\|_\infty, \mathbf{Y} - \mu) .$$

Recall that from the proof of Theorem 3.2, with probability larger than $1 - \alpha\gamma$ we have

$$q_{\alpha(1-\gamma)}(\|\cdot\|_\infty, \mathbf{Y} - \mu) \leq q_{\alpha(1-\delta)(1-\gamma)} \left(\|\cdot\|_\infty, \mathbf{Y} - \bar{\mathbf{Y}} \right) + \varepsilon'(\alpha, \delta, \gamma, n, K) .$$

Take \mathbf{Y} in the event where the above inequality holds. If the global procedure rejects at least one true null hypothesis, we denote j_0 the first time that this occurs ($j_0 = 0$ if it is in the first step). There are two cases:

- if $j_0 = 0$ then we have

$$T(\mathcal{H}_0) \geq q_{\alpha(1-\delta)(1-\gamma)} \left(\|\cdot\|_\infty, \mathbf{Y} - \bar{\mathbf{Y}} \right) + \varepsilon'(\alpha, \delta, \gamma, n, K) \geq q_{\alpha(1-\gamma)} \left(\sup_{\mathcal{H}_0} |\cdot|, \mathbf{Y} \right)$$

- if $j_0 \geq 1$, following the proof of Theorem 4.3, $T(\mathcal{H}_0) \geq q_{\alpha(1-\gamma)} \left(\sup_{\mathcal{H}_0} |\cdot|, \mathbf{Y} \right)$.

In both cases, $T(\mathcal{H}_0) \geq q_{\alpha(1-\gamma)} \left(\sup_{\mathcal{H}_0} |\cdot|, \mathbf{Y} \right)$, which occurs with probability smaller than $\alpha(1 - \gamma)$. \blacksquare

7.4. Exchangeable resampling computations. In this section, we compute constants A_W, B_W, C_W and D_W (defined by (3) to (6)) for some exchangeable resamplings. This implies all the statements in Tab. 1. We first define several additional exchangeable resampling weights:

- **Bernoulli** (p), $p \in (0, 1)$: pW_i i.i.d. with a Bernoulli distribution of parameter p . A classical choice is $p = \frac{1}{2}$.
- **Efron** (q), $q \in \{1 \dots, n\}$: $qn^{-1}W$ has a multinomial distribution with parameters $(q; n^{-1}, \dots, n^{-1})$. A classical choice is $q = n$.
- **Poisson** (μ), $\mu \in (0, +\infty)$: μW_i i.i.d. with a Poisson distribution of parameter μ . A classical choice is $\mu = 1$.

Notice that $\bar{\mathbf{Y}}^{(W-\bar{W})}$ and all the resampling constants are invariant under translation of the weights, so that Bernoulli (1/2) weights are completely equivalent to Rademacher weights in this paper.

Lemma 7.3 1. Let W be Bernoulli (p) weights with $p \in (0, 1)$. Then,

$$2(1-p) - \sqrt{\frac{1-p}{pn}} \leq A_W \leq B_W \leq \sqrt{\frac{1}{p} - 1} \sqrt{1 - \frac{1}{n}}$$

$$C_W = \sqrt{\frac{1}{p} - 1} \quad \text{and} \quad D_W \leq \frac{1}{2p} + \left| \frac{1}{2p} - 1 \right| + \sqrt{\frac{1-p}{np}} .$$

2. Let W be Efron (q) weights with $q \in \{1, \dots, n\}$. Then,

$$A_W \leq B_W \leq \sqrt{\frac{n-1}{n}} \quad \text{and} \quad C_W = 1 .$$

Moreover, if $q \leq n$,

$$A_W = 2 \left(1 - \frac{1}{n}\right)^q .$$

3. Let W be Poisson (μ) weights with $\mu > 0$. Then,

$$A_W \leq B_W \leq \frac{1}{\sqrt{\mu}} \sqrt{1 - \frac{1}{n}} \quad \text{and} \quad C_W = \frac{1}{\sqrt{\mu}} .$$

Moreover, if $\mu = 1$,

$$\frac{2}{e} - \frac{1}{\sqrt{n}} \leq A_W .$$

4. Let W be Random hold-out (q) weights with $q \in \{1, \dots, n\}$. Then,

$$\begin{aligned} A_W &= 2 \left(1 - \frac{q}{n}\right) & B_W &= \sqrt{\frac{n}{q} - 1} \\ C_W &= \sqrt{\frac{n}{n-1}} \sqrt{\frac{n}{q} - 1} & \text{and} & \quad D_W = \frac{n}{2q} + \left|1 - \frac{n}{2q}\right| . \end{aligned}$$

Proof of Lemma 7.3. We consider the following cases:

General case. We first only assume that W is exchangeable. Then, from the concavity of $\sqrt{\cdot}$ and the triangular inequality, we have

$$\begin{aligned} \mathbb{E} |W_1 - \mathbb{E}[W_1]| - \sqrt{\mathbb{E} (\overline{W} - \mathbb{E}[W_1])^2} &\leq \mathbb{E} |W_1 - \mathbb{E}[W_1]| - \mathbb{E} |\overline{W} - \mathbb{E}[W_1]| \\ (37) \quad &\leq A_W \leq B_W \leq \sqrt{\frac{n-1}{n}} C_W . \end{aligned}$$

Independent weights. When we suppose that the W_i are i.i.d.,

$$(38) \quad \mathbb{E} |W_1 - \mathbb{E}[W_1]| - \frac{\sqrt{\text{Var}(W_1)}}{\sqrt{n}} \leq A_W \quad \text{and} \quad C_W = \sqrt{\text{Var}(W_1)} .$$

Bernoulli. These weights are i.i.d. with $\text{Var}(W_1) = p^{-1} - 1$, $\mathbb{E}[W_1] = 1$ and

$$\mathbb{E} |W_1 - 1| = p(p^{-1} - 1) + (1 - p) = 2(1 - p) .$$

With (37) and (38), we obtain the bounds for A_W , B_W and C_W . Moreover, Bernoulli (p) weights satisfy the assumption of (6) with $x_0 = a = (2p)^{-1}$. Then,

$$D_W = \frac{1}{2p} + \mathbb{E} \left| \overline{W} - \frac{1}{2p} \right| \leq \frac{1}{2p} + \left| 1 - \frac{1}{2p} \right| + \mathbb{E} |\overline{W} - 1| \leq \frac{1}{2p} + \frac{1}{p} \left| \frac{1}{2} - p \right| + \sqrt{\frac{1-p}{np}} .$$

Efron. We have $\overline{W} = 1$ a.s. so that

$$C_W = \sqrt{\frac{n}{n-1}} \text{Var}(W_1) = 1 .$$

If moreover $q \leq n$, then $W_i < 1$ implies $W_i = 0$ and

$$\begin{aligned} A_W &= \mathbb{E} |W_1 - 1| = \mathbb{E} [W_1 - 1 + 2\mathbf{1}\{W_1 = 0\}] \\ &= 2\mathbb{P}(W_1 = 0) = 2 \left(1 - \frac{1}{n}\right)^q . \end{aligned}$$

The result follows from (37).

Poisson. These weights are i.i.d. with $\text{Var}(W_1) = \mu^{-1}$, $\mathbb{E}[W_1] = 1$. Moreover, if $\mu \leq 1$, $W_i < 1$ implies $W_i = 0$ and

$$\mathbb{E} |W_1 - 1| = 2\mathbb{P}(W_1 = 0) = 2e^{-\mu} .$$

With (37) and (38), the result follows.

Random hold-out. These weights are such that $\{W_i\}_{1 \leq i \leq n}$ takes only two values, with $\overline{W} = 1$. Then, A_W , B_W and C_W can be directly computed. Moreover, they satisfy the assumption of (6) with $x_0 = a = n/(2q)$. The computation of D_W is straightforward. ■

7.5. Non-exchangeable weights. In Section 2.5.1, we considered non-exchangeable weights in order to reduce the complexity of computation of expectations w.r.t. the resampling randomness. Then, we are mainly interested in non-exchangeable weights with small support. This is why we focus on the two following cases:

1. deterministic weights
2. V -fold weights ($V \in \{2, \dots, n\}$): let $(B_j)_{1 \leq j \leq V}$ be a partition of $\{1, \dots, n\}$ and $W^B \in \mathbb{R}^V$ an exchangeable resampling weight vector of size V . Then, for any $i \in \{1, \dots, n\}$ with $i \in B_j$, define $W_i = W_j^B$.

We will often assume that the partition $(B_j)_{1 \leq j \leq V}$ is “regular”, *i.e.* that V divides n and $|B_j| = n/V$ for every $j \in \{1, \dots, V\}$. When V does not divide n , the B_j can be chosen approximatively of the same size.

In the following, we make use of five constants that depend only on the resampling scheme: B_W and D_W stay unchanged (see definitions (4) and (6)), we modify the definitions of A_W and C_W (notice that we stay consistent with (3) and (5) when W is exchangeable), and we introduce

a fifth constant E_W (which is equal to A_W in the exchangeable case):

$$(39) \quad A_W := \frac{1}{n} \sum_{i=1}^n \mathbb{E} |W_i - \bar{W}|$$

$$(40) \quad C_W := \sqrt{n} B_W \quad \text{if } W \text{ is deterministic}$$

$$(41) \quad C_W := \sqrt{\max_j |B_j| C_{W^B} + \sqrt{n} \mathbb{E} |\bar{W}^B - \bar{W}|} \quad \text{if } W \text{ is } V\text{-fold}$$

$$(42) \quad E_W := \sqrt{\frac{1}{n} \sum_{i=1}^n (\mathbb{E} |W_i - \bar{W}|)^2} .$$

We can now state the main theorem of this section.

Theorem 7.4 *Let W be either a deterministic or V -fold resampling weight vector, and define the constants A_W , B_W , C_W , D_W and E_W by (39), (4), (40), (41), (6) and (42). Then, all the results of Theorem 2.1 and Corollary 2.2 hold, with only a slight modification in (8):*

$$\phi(\bar{\mathbf{Y}} - \mu) < \frac{\mathbb{E}_W \left[\phi(\bar{\mathbf{Y}}^{(W-\bar{W})}) \right]}{A_W} + \frac{M}{\sqrt{n}} \sqrt{1 + \frac{A_W^2}{E_W^2}} \sqrt{2 \log(1/\alpha)} .$$

Proof of Theorem 7.4. In the Gaussian case, we use the same proof as Theorem 2.1 and Corollary 2.2, but we replace the concentration result (16) by the one of Proposition 7.6.

In the bounded case, the proof is identical (it relies on Mc Diarmid inequality), but we no longer have $A_W = E_W$ because the weights are non-exchangeable. \blacksquare

Remark 7.5 *When V divides n , the constants of a (regular) V -fold weight vector are derived from those of the associated exchangeable weight vector W^B in the following way:*

$$A_W = E_W = A_{W^B} \quad B_W = B_{W^B} \quad C_W = \sqrt{\frac{n}{V}} C_{W^B} .$$

We now give two natural examples of non-exchangeable weights:

1. **Hold-out** (q): $W_i = \frac{n}{q} \mathbb{1}\{i \in I\}$ for some deterministic subset $I \subset \{1, \dots, n\}$ of cardinality q . A classical choice is $q = \lfloor n/2 \rfloor$.
2. **V -fold cross validation** (possibly non-regular), $V \in \{2, \dots, n\}$: V -fold weights with W^B leave-one-out (which is often called cross-validation). More precisely, $W_i = \frac{V}{V-1} \mathbb{1}\{i \notin B_J\}$, J uniform on $\{1, \dots, V\}$, $(B_j)_{1 \leq j \leq V}$ partition of $\{1, \dots, n\}$.

The terms ‘‘hold-out’’, ‘‘cross-validation’’ and ‘‘ V -fold cross-validation’’ refer to slightly different procedures which inspired these weights. In those two cases, we can compute the resampling constants :

1. **Hold-out** (q) :

$$A_W = 2 \left(1 - \frac{q}{n}\right) \quad B_W = E_W = \sqrt{\frac{n}{q} - 1}$$

$$C_W = \sqrt{n \left(\frac{n}{q} - 1\right)} \quad \text{and} \quad D_W = \frac{n}{2q} + \left|1 - \frac{n}{2q}\right| .$$

2. **V-fold cross validation** (possibly non-regular):

$$A_W = \frac{2}{V-1} \sum_{j=1}^V \frac{|B_j|}{n} \left(1 - \frac{|B_j|}{n}\right)$$

$$B_W = \frac{1}{V-1} \sum_{j=1}^V \sqrt{\frac{|B_j|}{n} \left(1 - \frac{|B_j|}{n}\right)}$$

$$C_W = \sqrt{\frac{\max_j |B_j|}{V-1}} \frac{\sqrt{V}}{V-1} + \frac{\sqrt{n}}{V-1} \sum_{j=1}^V \left| \frac{|B_j|}{n} - \frac{1}{V} \right|$$

$$D_W = \frac{1}{V-1} \sum_{j=1}^V \left(\frac{1}{2} + \left| \frac{1}{2} - \frac{|B_j|}{n} \right| \right)$$

$$E_W = \frac{2}{V-1} \sqrt{\sum_{j=1}^V \frac{|B_j|}{n} \left(1 - \frac{|B_j|}{n}\right)^2} .$$

When the partition $(B_j)_{1 \leq j \leq V}$ is almost regular, *i.e.* $\max_j ||B_j| - nV^{-1}| \leq 1$ and $n \gg V \geq 3$, then $C_W B_W^{-1} \leq \sqrt{n/(V-1)} (1 + o(1))$ which is close to its value in the “regular” case. This means that the concentration thresholds behave as in the regular case provided that n is large enough.

The proofs of these results are given at the end of this section. Before this, we give analogues of the results of Section 2.2 and 2.3 in the non-exchangeable case.

7.5.1. *Expectations.* The Proposition 2.4 is valid with non-exchangeable weights. The proof of Proposition 2.6 remains unchanged with non-exchangeable weights, with A_W defined by (39).

7.5.2. *Concentration inequalities.* Whereas Proposition 2.8 deals only with exchangeable weights, we can derive a similar result for deterministic and V -fold exchangeable weights. This is the object of the following result.

Proposition 7.6 *Let $p \in [1, +\infty]$, \mathbf{Y} a sample satisfying (GA) and $\phi : \mathbb{R}^K \rightarrow \mathbb{R}$ any subadditive function, bounded by the p -norm. Let W be some resampling weight vector among*

- (i) *Deterministic weights.*
- (ii) *V -fold exchangeable resampling weight for some $V \in \{2, \dots, n\}$.*

Then, for all $\alpha \in (0, 1)$, (16) and the corresponding lower bound hold with C_W defined by (40) (deterministic case) (41) (V -fold case).

Proof of Proposition 7.6. Deterministic weights (i): we can use (15) and the corresponding lower bound with $B_W\sigma$ instead of σ because $\mathcal{D}(\bar{\mathbf{Y}}^{\langle W - \bar{W} \rangle}) = \mathcal{D}(B_W(\bar{\mathbf{Y}} - \mu))$. The result follows with $C_W = \sqrt{n}B_W$.

V -fold weights (ii): the proof is widely inspired from the one of Proposition 2.8. We have to compute the Lipschitz constant of T_2 defined by

$$T_2(\zeta) = \mathbb{E} \phi \left(\frac{1}{n} \sum_{i=1}^n (W_i - \bar{W}) A \zeta_i \right) .$$

For all $\zeta, \zeta' \in \mathbb{R}^K$, we use the triangular inequality and the same arguments as in the proof of Proposition 2.8:

$$\begin{aligned} |T_2(\zeta) - T_2(\zeta')| &\leq \mathbb{E} \left\| \frac{1}{n} \sum_{i=1}^n (W_i - \bar{W}) A(\zeta_i - \zeta'_i) \right\|_p \\ &\leq \mathbb{E} \left\| \frac{1}{n} \sum_{i=1}^n (W_i - \bar{W}^B) A(\zeta_i - \zeta'_i) \right\|_p + \mathbb{E} |\bar{W}^B - \bar{W}| \left\| \frac{1}{n} \sum_{i=1}^n A(\zeta_i - \zeta'_i) \right\|_p \\ &\leq \frac{\|\sigma\|_p}{n} \sqrt{\mathbb{E} \left\| \sum_{i=1}^n (W_i - \bar{W}^B)(\zeta_i - \zeta'_i) \right\|_2^2} + \mathbb{E} |\bar{W}^B - \bar{W}| \frac{\|\sigma\|_p}{\sqrt{n}} \|\zeta - \zeta'\|_{2,Kn} \end{aligned}$$

Using the exchangeability of the W^B , we show that

$$\begin{aligned} \mathbb{E} \left\| \sum_{i=1}^n (W_i - \bar{W}^B)(\zeta_i - \zeta'_i) \right\|_2^2 &= \mathbb{E} \left\| \sum_{j=1}^V (W_j^B - \bar{W}^B) \sum_{i \in B_j} (\zeta_i - \zeta'_i) \right\|_2^2 \\ &\leq C_{W^B}^2 \sum_{j=1}^V \left\| \sum_{i \in B_j} (\zeta_i - \zeta'_i) \right\|_2^2 \\ &\leq C_{W^B}^2 \sum_{j=1}^V |B_j| \sum_{i \in B_j} \|\zeta_i - \zeta'_i\|_2^2 \end{aligned}$$

by convexity of $\|\cdot\|_2^2$. Finally, this implies that T_2 is Lipschitz of parameter

$$\frac{\|\sigma\|_p}{n} \sqrt{\max_j |B_j|} C_{W^B} + \frac{\|\sigma\|_p}{\sqrt{n}} \mathbb{E} |\bar{W}^B - \bar{W}| .$$

■

7.5.3. *Computation of the constants.* We first remark that the following statements are straightforward:

- if W is deterministic, $B_W = E_W$.
- if W is regular V -fold exchangeable,

$$A_W = E_W = A_{WB} \quad B_W = B_{WB} \quad C_W = \sqrt{\frac{n}{V}} C_{WB}.$$

In the hold-out (q) case, we compute A_W , B_W and D_W exactly as in the Random hold-out (q) case.

In the general V -fold cross-validation case, we use the following trick : conditionally to the index J of the removed block, W is a deterministic hold-out ($n - |B_J|$) weight multiplied by a factor $c(J) = \frac{V(n-|B_J|)}{(V-1)n}$. This allows to compute A_W , B_W and D_W from the hold-out case: for instance,

$$\begin{aligned} A_W &= \frac{1}{V} \sum_{j=1}^V \left[2c(J) \left(1 - \frac{q}{n} \right) \right] \\ &= \frac{2}{V-1} \sum_{j=1}^V \frac{|B_j|}{n} \left(1 - \frac{|B_j|}{n} \right) . \end{aligned}$$

This also shows

$$\mathbb{E} \left| \overline{W^B} - \overline{W} \right| = \frac{1}{V} \sum_{j=1}^V \left| \frac{V}{V-1} \frac{n - |B_j|}{n} - 1 \right|$$

from which we obtain C_W . The computation of E_W is done directly by noting that

$$\mathbb{E} \left| W_j^B - \overline{W} \right| = \frac{V}{V-1} \mathbb{E} \left| \mathbf{1}_{\{j \neq J\}} - 1 + \frac{|B_j|}{n} \right| = \frac{2}{V-1} \left(1 - \frac{|B_j|}{n} \right) ,$$

$$\begin{aligned} E_W^2 &= \frac{1}{n} \sum_{i=1}^n (\mathbb{E} |W_i - \overline{W}|)^2 \\ &= \sum_{j=1}^V \frac{|B_j|}{n} (\mathbb{E} |W_j^B - \overline{W}|)^2 \\ &= \left(\frac{2}{V-1} \right)^2 \sum_{j=1}^V \frac{|B_j|}{n} \left(1 - \frac{|B_j|}{n} \right)^2 . \end{aligned}$$

We now prove the last statement about “almost regular” V -fold cross-validation: when $\max_j |B_j| \leq nV^{-1} + 1$,

$$\begin{aligned} C_W &\leq \sqrt{\frac{n}{V} + 1} \frac{\sqrt{V}}{V-1} + \frac{V\sqrt{n}}{n(V-1)} \\ &\leq \frac{\sqrt{n}}{V-1} \left(1 + \sqrt{\frac{V}{n} + \frac{V}{n}} \right). \end{aligned}$$

If moreover $V^{-1} + n^{-1} \leq 1/2$, we have:

$$\begin{aligned} B_W &\geq \frac{V}{V-1} \sqrt{\left(\frac{1}{V} - \frac{1}{n}\right) \left(1 - \frac{1}{V} + \frac{1}{n}\right)} \\ &= \frac{1}{\sqrt{V-1}} \sqrt{1 + \frac{V^2}{(V-1)n} \left(\frac{2}{V} - 1 - \frac{1}{n}\right)} \\ &\geq \frac{1}{\sqrt{V-1}} - \frac{V}{(V-1)\sqrt{n}} \sqrt{\left(1 + \frac{1}{n} - \frac{2}{V}\right)_+}. \end{aligned}$$

Acknowledgements. We want to thank Pascal Massart for his particularly relevant suggestions.

REFERENCES

- [1] Fisher R. A. *The Design of Experiments*. Oliver and Boyd, Edinburgh, 1935.
- [2] Sylvain Arlot. *Resampling and Model Selection*. PhD thesis, University Paris XI, December 2007.
- [3] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B*, 57(1):289–300, 1995.
- [4] Rudolf Beran. The impact of the bootstrap on statistical algorithms and theory. *Statist. Sci.*, 18(2):175–184, 2003. Silver anniversary of the bootstrap.
- [5] B. R. Cirel’son, I. A. Ibragimov, and V. N. Sudakov. Norms of Gaussian sample functions. In *Proceedings of the Third Japan–USSR Symposium on Probability Theory*, volume 550 of *Lecture notes in mathematics*, pages 20–41. Springer, 1976.
- [6] F. Darvas, M. Rautiainen, D. Pantazis, S. Baillet, H. Benali, J.C. Mosher, L. Garnero, and R.M. Leahy. Investigations of dipole localization accuracy in meg using the bootstrap. *NeuroImage*, 25:355–368, 2005.
- [7] Thomas J. DiCiccio and Bradley Efron. Bootstrap confidence intervals. *Statist. Sci.*, 11(3):189–228, 1996. With comments and a rejoinder by the authors.
- [8] B. Efron. Bootstrap methods: another look at the jackknife. *Ann. Statist.*, 7(1):1–26, 1979.
- [9] M. Fromont. Model selection by bootstrap penalization for classification. *Machine Learning*, 66(2-3):165–207, 2006.
- [10] Yongchao Ge, Sandrine Dudoit, and Terence P. Speed. Resampling-based multiple testing for microarray data analysis. *Test*, 12(1):1–77, 2003. With comments and a rejoinder by the authors.
- [11] Peter Hall. *The bootstrap and Edgeworth expansion*. Springer Series in Statistics. Springer-Verlag, New York, 1992.

- [12] Peter Hall and Enno Mammen. On general resampling algorithms and their performance in distribution estimation. *Ann. Statist.*, 22(4):2011–2030, 1994.
- [13] Sture Holm. A simple sequentially rejective multiple test procedure. *Scand. J. Statist.*, 6(2):65–70, 1979.
- [14] Karim Jerbi, Jean-Philippe Lachaux, Karim N’Diaye, Dimitrios Pantazis, Richard M. Leahy, Line Garnero, and Sylvain Baillet. Coherent neural representation of hand speed in humans revealed by meg imaging. *PNAS*, 104(18):7676–7681, 2007.
- [15] David M. Mason and Michael A. Newton. A rank statistics approach to the consistency of a general bootstrap. *Ann. Statist.*, 20(3):1611–1624, 1992.
- [16] Pascal Massart. *Concentration Inequalities and Model Selection (Lecture notes of the St-Flour probability summer school 2003)*, volume 1896 of *Lecture notes in Mathematics*. Springer, 2007.
- [17] C. McDiarmid. On the method of bounded differences. In *Surveys in combinatorics*, volume 141 of *London Mathematical Society Lecture Notes*, pages 148–188. Cambridge University Press, 1989.
- [18] Dimitrios Pantazis, Thomas E. Nichols, Sylvain Baillet, and Richard M. Leahy. A comparison of random field theory and permutation methods for statistical analysis of meg data. *NeuroImage*, 25:383–394, 2005.
- [19] M. Perone Pacifico, I. Genovese, I. Verdinelli, and L. Wasserman. False discovery control for random fields. *Journal of the American Statistical Association*, 99(468):1002–1014, 2004.
- [20] Dimitris N. Politis, Joseph P. Romano, and Michael Wolf. *Subsampling*. Springer Series in Statistics. Springer-Verlag, New York, 1999.
- [21] Katherine S. Pollard and Mark J. van der Laan. Resampling-based multiple testing: Asymptotic control of type i error and applications to gene expression data. Working Paper Series Working Paper 121, U.C. Berkeley Division of Biostatistics, 2003. available at <http://www.bepress.com/ucbbiostat/paper121>.
- [22] Jens Præstgaard and Jon A. Wellner. Exchangeably weighted bootstraps of the general empirical process. *Ann. Probab.*, 21(4):2053–2086, 1993.
- [23] Joseph P. Romano. Bootstrap and randomization tests of some nonparametric hypotheses. *Ann. Statist.*, 17(1):141–159, 1989.
- [24] Joseph P. Romano. On the behavior of randomization tests without a group invariance assumption. *J. Amer. Statist. Assoc.*, 85(411):686–692, 1990.
- [25] Joseph P. Romano and Michael Wolf. Control of generalized error rates in multiple testing. IEW - Working Papers iewwp245, Institute for Empirical Research in Economics - IEW, 2005. available at <http://ideas.repec.org/p/zur/iewwp/245.html>.
- [26] Joseph P. Romano and Michael Wolf. Exact and approximate stepdown methods for multiple hypothesis testing. *J. Amer. Statist. Assoc.*, 100(469):94–108, 2005.
- [27] Aad W. van der Vaart and Jon A. Wellner. *Weak convergence and empirical processes*. Springer Series in Statistics. Springer-Verlag, New York, 1996.
- [28] T. Waberski, R. Gobbele, W. Kawohl, C. Cordes, and H. Buchner. Immediate cortical reorganization after local anesthetic block of the thumb: source localization of somatosensory evoked potentials in human subjects. *Neurosci. Lett.*, 347:151–154, 2003.
- [29] P. H. Westfall and S. S. Young. *Resampling-Based Multiple Testing*. Wiley, 1993. Examples and Methods for P -Value Adjustment.
- [30] Daniel Yekutieli and Yoav Benjamini. Resampling-based false discovery rate controlling multiple test procedures for correlated test statistics. *J. Statist. Plann. Inference*, 82(1-2):171–196, 1999. Multiple comparisons (Tel Aviv, 1996).

SYLVAIN ARLOT
UNIV PARIS-SUD, UMR 8628,
LABORATOIRE DE MATHEMATIQUES,
ORSAY, F-91405 ; CNRS, ORSAY, F-91405 ;
INRIA-FUTURS, PROJET SELECT
E-MAIL: sylvain.arlot@math.u-psud.fr

GILLES BLANCHARD
FRAUNHOFER FIRST.IDA,
BERLIN,
GERMANY
E-MAIL: blanchar@first.fraunhofer.de

ETIENNE ROQUAIN
DEPARTMENT OF MATHEMATICS, VRIJE UNIVERSITEIT,
DE BOELELAAN 1081A, 1081 HV AMSTERDAM,
THE NETHERLANDS
E-MAIL: eroquain@few.vu.nl