Submitted to the Annals of Statistics

SOME NON-ASYMPTOTIC RESULTS ON RESAMPLING IN HIGH DIMENSION, II: MULTIPLE TESTS

By Sylvain Arlot*,†, Gilles Blanchard*,‡ and Etienne Roquain*,§

CNRS ENS, Weierstrass Institute Berlin and University of Paris 6

In the context of correlated multiple tests, we aim at controlling non-asymptotically the family-wise error rate (FWER) using resampling-type procedures. We observe repeated realizations of a Gaussian random vector in possibly high dimension and with an unknown covariance matrix, and consider the one and two-sided multiple testing problem for the mean values of its coordinates. We address this problem by using the confidence regions developed in the companion paper [1], which lead directly to single-step procedures; these can then be improved using step-down algorithms, following an established general methodology laid down by Romano and Wolf [16]. We compare the performance of the different obtained thresholds on simulated data.

1. Introduction.

1.1. *Framework and motivations.* We consider a sample $\mathbf{Y} := (\mathbf{Y}^1, \dots, \mathbf{Y}^n)$ of $n \geq 2$ i.i.d. observations of a Gaussian vector with dimensionality K possibly much larger than n. The common covariance matrix of the $\mathbf{Y}^i$ is not assumed to be known in advance. We investigate the two following multiple testing problems for the common mean $\mu \in \mathbb{R}^K$ of the $\mathbf{Y}^i$:

- *One-sided*: test simultaneously H_k : "$\mu_k \leq 0$" against A_k : "$\mu_k > 0$", for $1 \leq k \leq K$.
- *Two-sided*: test simultaneously H_k : "$\mu_k = 0$" against A_k : "$\mu_k \neq 0$", for $1 \leq k \leq K$.

For simplicity, we introduce the following single notation to cover both cases:

$$\text{test simultaneously } H_k : \text{“}[\![\mu_k]\!] = 0\text{” against } A_k : \text{“}[\![\mu_k]\!] \neq 0\text{”, for } 1 \leq k \leq K \ , \tag{1}$$

*This work was supported in part by the IST and ICT Programmes of the European Community, successively under the PASCAL (IST-2002-506778) and PASCAL2 (ICT-216886) networks of excellence.

†Research mostly carried out at Univ Paris-Sud (Laboratoire de Mathematiques, CNRS - UMR 8628).

‡This research was in part carried out while the second author held an invited position at the statistics department of the University of Chicago, which is warmly acknowledged.

§Research mostly carried out at the French institute INRA-Jouy and at the Free University of Amsterdam.

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

Keywords and phrases: family-wise error, multiple testing, high dimensional data, non-asymptotic error control, resampling, resampled quantile

where for $x \in \mathbb{R}$, $[\![x]\!]$ denotes either $\max\{x, 0\} = x_+$ in the one-sided context or $|x|$ in the two-sided context.

In this paper, we tackle the problem (1) by building multiple testing procedures which control the family-wise error rate (FWER). We emphasize that:

- we aim at obtaining a *non-asymptotical* control, valid for any fixed K and n, in particular with K possibly much larger than the number of observations n ,
- we do not want to make any particular prior assumption on the structure of covariance matrix of the $\mathbf{Y}^i$.

As explained in [1], this point of view is motivated by some practical applications, especially neuroimaging [11, 5, 10]. Multiple testing problems in this field typically have parameters $10^4 \leq K \leq 10^7$, $n \leq 100$, with strong and complex dependencies between the coordinates of $\mathbf{Y}^i$. Another motivating example is microarray data analysis (see for instance [8]).

1.2. *Goals.* We consider in this work thresholding-based procedures which reject the null hypotheses H_k for indices $k \in R_\alpha(\mathbf{Y}) \subset \mathcal{H} := \{1, ..., K\}$ corresponding to large values of $[\![\overline{\mathbf{Y}}_k]\!]$, where $\overline{\mathbf{Y}}_k = n^{-1} \sum_{i=1}^n \mathbf{Y}_k^i$ denote the vector of empirical means, that is,

$$R_\alpha(\mathbf{Y}) = \left\{ 1 \leq k \leq K \mid [\![\overline{\mathbf{Y}}_k]\!] > t_\alpha(\mathbf{Y}) \right\} \quad , \tag{2}$$

where $t_\alpha(\mathbf{Y})$ is a possibly data-dependent threshold.

The type I error of such a multiple testing procedure is measured here by the family-wise error rate (FWER), defined as the probability that at least one hypothesis is wrongly rejected:

$$\mathrm{FWER}(R_\alpha) := \mathbb{P}\left(R_\alpha\left(\mathbf{Y}\right) \cap \mathcal{H}_0 \neq \emptyset\right) \quad ,$$

where $\mathcal{H}_0 := \{k \mid [\![\mu_k]\!] = 0\}$ is the set of coordinates corresponding to the true null hypotheses. The choice of this error rate is discussed in Section 5.1. Given a level $\alpha \in (0, 1)$, the goal is now to build a multiple testing procedure R_α such that $\mathrm{FWER}(R_\alpha) \leq \alpha$ is valid for all distributions in the considered family (that is, Gaussian with arbitrary mean vector and covariance matrix); furthermore, as many false hypotheses as possible should be rejected.

To this end, we use the family of $(1-\alpha)$-resampling-based confidence regions for μ introduced in the companion paper [1]. Of interest here are regions taking the following form: for some subset $\mathcal{C} \subset \mathcal{H}$,

$$\mathcal{G}(\mathbf{Y}, 1-\alpha, \mathcal{C}) := \left\{ x \in \mathbb{R}^K \;\middle|\; \sup_{k \in \mathcal{C}} [\![\overline{\mathbf{Y}}_k - x_k]\!] \leq t_\alpha(\mathbf{Y}, \mathcal{C}) \right\} \quad , \tag{3}$$

where t_α is a data-dependent threshold built using a resampling principle. Several possible choices for this threshold were proposed; the main results of [1], as well as the link between confidence regions (3) and (single-step) multiple tests for (1), are briefly recalled in Section 2.

1.3. *Contribution in relation to previous work.* Most of the existing resampling-type multiple testing procedures have been developed in an asymptotic framework (see for instance [18, 19, 8, 13, 17]), while our present goal is to study procedures that have non-asymptotic theoretical validity (for any K and n) . The main classical alternative approach to asymptotic validity is to use an invariance of the null distribution under a group of transformations – that is, exact randomized tests [14, 15, 16] (the underlying idea can be traced back to Fisher's permutation test [7]). Additionally, and as explained in [16], exact tests can be combined with a step-down algorithm to build less conservative procedures while preserving the same non-asymptotic control on their FWER (see also [17] for a generalization to the k-FWER).

In the case considered here, Gaussian vectors $\mathbf{Y}^i$ have a symmetric distribution around their mean, so that the action of mirroring any subset of the vectors in the data sample with respect to their mean constitutes such a group of distribution preserving transformations. In the two-sided case, this group is known under the global null hypothesis $\mu = 0$ and just corresponds to arbitrary sign flipping of each data vector. A similar idea was used for instance in [6] to build an adaptive (single) test for zero mean under the assumption of symmetric and independent errors. The setting studied here is different, since we consider multiple testing with possibly dependent errors.

Consequently, it is possible to derive directly from [16] a step-down procedure whose FWER is controlled in an non-asymptotic setting (see Section 3). This approach will be referred to as *uncentered* in this paper, because the sign-flipping is applied on the $(\mathbf{Y}^i)_{1\leq i\leq n}$ themselves, without prior centering.

Compared to the latter uncentered approach, most of the procedures proposed in this paper consist in applying the sign-flipping to the *empirically centered data* $(\mathbf{Y}^i - \overline{\mathbf{Y}})_{1\leq i\leq n}$. It was proved in [1] that such an intuitive idea is theoretically valid despite the dependencies between the $\mathbf{Y}^i - \overline{\mathbf{Y}}$, $1 \leq i \leq n$, at the price of adding some second-order remainder term. We argue in the present paper that in some interesting situations, the prior centering operation leads to a noticeable decrease of the computation time of the step-down algorithm, up to some small loss in accuracy (due to the remainder term) with respect to the uncentered step-down. Additionally, the centered approach can be used both in the one-sided and two-sided contexts, while the uncentered approach is, up to our knowledge, only proved to be valid in the two-sided case.

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

- $\mathbf{Y}$ denotes the $K \times n$ data matrix $(\mathbf{Y}^i_k)_{1\leq k\leq K, 1\leq i\leq n}$. A superscript index such as $\mathbf{Y}^i$ indicates the i-th column of a matrix. The empirical mean vector is $\overline{\mathbf{Y}} := \frac{1}{n}\sum_{i=1}^n \mathbf{Y}^i$. If $\mu \in \mathbb{R}^K$, $\mathbf{Y} - \mu$ is the matrix obtained by subtracting μ from each (column) vector of $\mathbf{Y}$.
- The vector $\sigma := (\sigma_k)_{1\leq k\leq K}$ is the vector of the standard deviations of the data: $\forall k, 1 \leq k \leq K$, $\sigma_k := \mathrm{Var}^{1/2}(\mathbf{Y}^1_k)$. For $\mathcal{C} \subset \mathcal{H}$, we also denote $\|\sigma\|_{\mathcal{C}} := \sup_{k\in\mathcal{C}} \sigma_k$.
- $\overline{\Phi}$ is the standard Gaussian upper tail function: if $X \sim \mathcal{N}(0, 1)$, $\forall x \in \mathbb{R}$, $\overline{\Phi}(x) = \mathbb{P}(X \geq x)$.
- If $W \in \mathbb{R}^n$, we define the mean of $W \in \mathbb{R}^n$ as $\overline{W} := \frac{1}{n}\sum_{i=1}^n W_i$, and for every $c \in \mathbb{R}$, $W - c := (W_i - c)_{1\leq i\leq n} \in \mathbb{R}^n$.

- For a subset $\mathcal{C} \subset \mathcal{H}$, $|\mathcal{C}|$ denotes the cardinality of $\mathcal{C}$.

2. Single-step procedures using resampling-based thresholds.

2.1. *Connection between confidence regions and FWER control.* We start with recalling a simple device linking confidence regions to FWER control in multiple testing. In a nutshell, the idea is that a confidence region of the form (3) directly gives a multiple testing procedure R with controlled FWER, when taking $\mathcal{C} = \mathcal{H}_0$. Since $\mathcal{H}_0$ is not known in advance, we actually need a confidence region (3) defined for every $\mathcal{C} \subset \mathcal{H}$ and satisfying some properties.

More formally, let $\alpha \in (0,1)$ be fixed and $\mathcal{T}_\alpha = (t_\alpha(\mathbf{Y}, \mathcal{C}), \mathcal{C} \subset \mathcal{H}, \mathbf{Y} \in \mathbb{R}^{K \times n})$ be a family of thresholds indexed by subsets $\mathcal{C} \subset \mathcal{H}$. We consider threshold families satisfying the two following key properties. First, $t_\alpha(\mathbf{Y}, \mathcal{H}_0)$ is a $(1-\alpha)$ confidence bound on the deviations of $\sup_{k \in \mathcal{H}_0} [\![\overline{\mathbf{Y}}_k]\!]$:

$$(\mathbf{CB}_\alpha) \qquad \mathbb{P}\left(\sup_{k \in \mathcal{H}_0} [\![\overline{\mathbf{Y}}_k]\!] < t_\alpha(\mathbf{Y}, \mathcal{H}_0) \right) \geq 1 - \alpha \ .$$

Second, $\mathcal{T}_\alpha$ is non-decreasing w.r.t. $\mathcal{C}$, that is,

$$(\mathbf{ND}) \qquad \forall \mathbf{Y} \in \mathbb{R}^K, \forall \mathcal{C}, \mathcal{C}' \subset \mathcal{H}, \quad \mathcal{C} \subset \mathcal{C}' \Rightarrow t_\alpha(\mathbf{Y}, \mathcal{C}) \leq t_\alpha(\mathbf{Y}, \mathcal{C}') \ .$$

We now define a single-step multiple testing procedure and establish its FWER control:

PROPOSITION 2.1. *Define the* single-step multiple testing procedure *associated to* $\mathcal{T}_\alpha$ *as the procedure rejecting the set of hypotheses given by*

$$(4) \qquad \left\{ k \in \mathcal{H} \mid [\![\overline{\mathbf{Y}}_k]\!] > t_\alpha(\mathbf{Y}, \mathcal{H}) \right\} \ .$$

If the threshold family satisfies $(\mathbf{CB}_\alpha)$ *and* $(\mathbf{ND})$, *the FWER of the associated single-step procedure is controlled at level* α.

Proof : We first use $(\mathbf{ND})$, then $(\mathbf{CB}_\alpha)$:

$$\begin{aligned} \mathbb{P}\left(\exists k \mid [\![\overline{\mathbf{Y}}_k]\!] > t_\alpha(\mathbf{Y}, \mathcal{H}) \text{ and } [\![\mu_k]\!] = 0 \right) &= \mathbb{P}\left(\sup_{k \in \mathcal{H}_0} [\![\overline{\mathbf{Y}}_k]\!] > t_\alpha(\mathbf{Y}, \mathcal{H}) \right) \\ &\leq \mathbb{P}\left(\sup_{k \in \mathcal{H}_0} [\![\overline{\mathbf{Y}}_k]\!] > t_\alpha(\mathbf{Y}, \mathcal{H}_0) \right) \leq \alpha \ . \quad \square \end{aligned}$$

Note that the single-step procedure only uses the value of the largest threshold among the $t_\alpha(\mathbf{Y}, \mathcal{C}), \mathcal{C} \subset \mathcal{H}$. In Section 3, we use the iterative step-down principle to improve the procedure by making use of the thresholds $t_\alpha(\mathbf{Y}, \mathcal{C})$ for some smaller $\mathcal{C} \subset \mathcal{H}$.

The condition $(\mathbf{CB}_\alpha)$ is in particular satisfied whenever for any $\mathcal{C} \subset \mathcal{H}$, $t(\mathbf{Y}, \mathcal{C})$ provides a $(1-\alpha)$ confidence region of the form (3) for $(\mu_k)_{k \in \mathcal{C}}$. We use this idea next to derive testing thresholds from the confidence regions constructed in [1].

2.2. *Resampling thresholds.* We first give a compact recapitulation of resampling-based thresholds introduced in [1], and used to build confidence regions for the mean of a high-dimensional, correlated vector. This is intended as a single overall reference for all the thresholds we use in the present paper. Here and in the following, $W \in \mathbb{R}^n$ denotes a random vector independent from the data $\mathbf{Y}$, called the *resampling weight vector*. Moreover, in order to simplify the results of [1], we assume specifically that the W_i are i.i.d. Rademacher random variables, that is, satisfy $\mathbb{P}(W_i = 1) = \mathbb{P}(W_i = -1) = 1/2$. As first building blocks, define the two following resampling quantities, the (scaled) resampled expectation and quantile:

$$\mathcal{E}(\mathbf{Y}, \mathcal{C}) := B_W^{-1} \mathbb{E}_W \left[\sup_{k \in \mathcal{C}} \left[\!\!\left[\frac{1}{n} \sum_{i=1}^n W_i \mathbf{Y}_k^i \right]\!\!\right] \right] ; \tag{5}$$

$$q_\alpha(\mathbf{Y}, \mathcal{C}) := \inf \left\{ x \in \mathbb{R} \,\middle|\, \mathbb{P}_W \left(\sup_{k \in \mathcal{C}} \left[\!\!\left[\frac{1}{n} \sum_{i=1}^n W_i \mathbf{Y}_k^i \right]\!\!\right] > x \right) \leq \alpha \right\} , \tag{6}$$

where $B_W := \mathbb{E}_W \left[\left(\frac{1}{n} \sum_{i=1}^n \left(W_i - \overline{W} \right) \right)^{1/2} \right]$, and $\mathbb{E}_W[\cdot]$ (resp. $\mathbb{P}_W(\cdot)$) denotes the expectation (resp. probability) operator over the distribution of W only. We also define the following function which is the upper quantile function of a binomial $(n, \frac{1}{2})$ variable:

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

and the factor

$$\gamma_n(\eta) := \frac{2\overline{\mathcal{B}}\left(n, \frac{\eta}{2}\right) - n}{n} \leq \left(\frac{2 \log\left(\frac{2}{\eta}\right)}{n} \right)^{1/2} ,$$

where the last inequality, intended as a more explicit formula, is obtained via Hoeffding's inequality.

Table 1 gives a reference of the different rejection thresholds considered in this paper, depending on a target type I error level α, subset of coordinates $\mathcal{C}$, and possibly on two arbitrary parameters $\alpha_0 \in (0, \alpha)$ and $\delta \in (0, 1)$. The threshold (7) is Bonferroni's for Gaussian variables. Thresholds (8), (9), (11) and (12) were introduced in [1]. More precisely, threshold (8) is based on a Gaussian concentration result. Threshold (9) is a compound threshold which is very close to the minimum of (7) and (8). Threshold (10) is a raw resampled quantile for the empirically centered data; it is not proved theoretically that this threshold achieves the correct level (this is signalled by the star symbol). The thresholds (11) and (12) are based on the latter with an additional term which was introduced in [1] to compensate (from a theoretical point of view) for the optimism in centering the data empirically rather than using the (unknown) true mean. The thresholds (7), (8) and (9) (and thus (11) and (12)) depend on the quantity $\|\sigma\|_{\mathcal{C}}$; if it is unknown, a confidence upper bound on $\|\sigma\|_{\mathcal{C}}$ can be built (see Section 4.1 of [1]). Finally, notice

$$t_{\alpha,\mathrm{Bonf}}(\mathbf{Y},\mathcal{C}) := \frac{1}{\sqrt{n}} \|\sigma\|_{\mathcal{C}} \overline{\Phi}^{-1}\left(\frac{\alpha}{c\,|\mathcal{C}|}\right) \text{ , with } \begin{cases} c=1 & \text{(one-sided case)} \\ c=2 & \text{(two-sided case)} \end{cases} ; \tag{7}$$

$$t_{\alpha,\mathrm{conc}}(\mathbf{Y},\mathcal{C}) := \mathcal{E}(\mathbf{Y}-\overline{\mathbf{Y}},\mathcal{C}) + \|\sigma\|_{\mathcal{C}} \overline{\Phi}^{-1}(\alpha/2)\left[\frac{1}{nB_W} + \frac{1}{\sqrt{n}}\right] ; \tag{8}$$

$$t_{\alpha,\mathrm{conc}\wedge\mathrm{Bonf}}(\mathbf{Y},\mathcal{C}) := \min\Big(t_{\alpha(1-\delta),\mathrm{Bonf}}(\mathbf{Y},\mathcal{C}), \mathcal{E}(\mathbf{Y}-\overline{\mathbf{Y}},\mathcal{C}) + \frac{\|\sigma\|_{\mathcal{C}}}{\sqrt{n}}\overline{\Phi}^{-1}\left(\frac{\alpha(1-\delta)}{2}\right) + \frac{\|\sigma\|_{\mathcal{C}}}{nB_W}\overline{\Phi}^{-1}\left(\frac{\alpha\delta}{2}\right)\Big) ; \tag{9}$$

$$t^*_{\alpha,\mathrm{quant}}(\mathbf{Y},\mathcal{C}) := q_\alpha(\mathbf{Y}-\overline{\mathbf{Y}},\mathcal{C}) ; \tag{10}$$

$$t_{\alpha,\mathrm{quant+Bonf}}(\mathbf{Y},\mathcal{C}) := t^*_{\alpha_0(1-\delta),\mathrm{quant}}(\mathbf{Y},\mathcal{C}) + \gamma_n(\alpha_0\delta)\, t_{\alpha-\alpha_0,\mathrm{Bonf}}(\mathbf{Y},\mathcal{C}) ; \tag{11}$$

$$t_{\alpha,\mathrm{quant+conc}}(\mathbf{Y},\mathcal{C}) := t^*_{\alpha_0(1-\delta),\mathrm{quant}}(\mathbf{Y},\mathcal{C}) + \gamma_n(\alpha_0\delta)\, t_{\alpha-\alpha_0,\mathrm{conc}}(\mathbf{Y},\mathcal{C}) ; \tag{12}$$

$$t_{\alpha,\mathrm{quant.uncent}}(\mathbf{Y},\mathcal{C}) := q_\alpha(\mathbf{Y},\mathcal{C}) . \tag{13}$$

TABLE 1
Reference table of the different rejection thresholds.

that all these thresholds are non-decreasing w.r.t. $\mathcal{C}$, that is, they satisfy assumption (**ND**). The non-asymptotic theoretical results obtained in [1] in the Gaussian case can be summed up in the following theorem:

THEOREM 2.2. *If $t_\alpha(\mathbf{Y},\mathcal{C})$ is one of the thresholds defined either by* (7), (8), (9), (11) *or* (12), *it holds for any $\mathcal{C} \subset \mathcal{H}$, in the one-sided as well as two-sided setting, that*

$$\mathbb{P}\left(\sup_{k\in\mathcal{C}} \left[\!\left[\overline{\mathbf{Y}}_k - \mu_k\right]\!\right] < t_\alpha(\mathbf{Y},\mathcal{C})\right) \geq 1-\alpha \ . \tag{14}$$

In particular, all these thresholds satisfy ($\mathbf{CB}_\alpha$), *both in the one-sided and two-sided cases.*

Note that the results obtained in [1] have more generality: in particular, variations of the above thresholds were proposed for non-Gaussian, but bounded, data; and weight families different from Rademacher weights can be used in (8) and (9) . For the purposes of the present work we restrict ourselves to Gaussian data and Rademacher weights for simplicity. It is straightforward to show that (14) implies ($\mathbf{CB}_\alpha$): the two-sided case is obvious since $\mu_k = 0$ for $k \in \mathcal{H}_0$; the one-sided case is an easy consequence of the fact that the positive part is a non-decreasing function. Therefore, by an application of Proposition 2.1, the corresponding thresholds $t_\alpha(\mathbf{Y},\mathcal{H})$ for the full set of hypotheses can be used for multiple testing in the one-sided as well as two-sided setting with a non-asymptotic control of the FWER.

We mentioned above that the thresholds (11) and (12), based on a resampled quantile for the *empirically centered* data $(\mathbf{Y}-\overline{\mathbf{Y}})$, include an additional term in order to upper bound the variations introduced by the centering operation. In the context of testing however, it is important to notice that the quantile for the *uncentered* data defined in (13) is (without modification) a valid threshold in the two-sided setting:

THEOREM 2.3. *Assume only that* $\mathbf{Y}$ *has a symmetric distribution around its mean* μ *, that is,* $(\mathbf{Y}^1 - \mu) \sim (\mu - \mathbf{Y}^1)$ *. If* $\mu_k = 0$ *for all* $k \in \mathcal{C}$ *, then the threshold* $t_{\alpha,\text{quant.uncent}}(\mathbf{Y}, \mathcal{C})$ *defined by* (13) *satisfies* (14) *. In particular, the threshold* $t_{\alpha,\text{quant.uncent}}(\mathbf{Y}, \mathcal{C})$ *satisfies* $(\mathbf{CB}_\alpha)$ *in the two-sided setting.*

This result can probably be considered well-known, and corresponds for example to Lemma 3.1 in [1] . Again by Proposition 2.1, the threshold defined by (13) can be therefore used for multiple testing (though for the two-sided setting only).

It is useful at this point to have a brief qualitative comparison of the uncentered quantile threshold (13) versus the centered quantile thresholds (11) and (12) (in the two-sided setting). The obvious differences between the two types of thresholds are:

- the data vectors are not centered around the empirical mean $\overline{\mathbf{Y}}$ prior to computing the threshold (13) ;
- the centered thresholds (11) and (12) have an additional additive term with respect to the main resampled quantile; furthermore, the main centered quantile is computed at a shrunk error level $\alpha_0(1 - \delta) < \alpha$.

The second point is a net drawback of the "centered" family compared to the "uncentered" one. On the other hand, empirical centering of the data has the advantage of making the corresponding threshold $t_\alpha(\mathbf{Y}, \mathcal{C})$ *translation invariant* , that is, for every $\mathbf{Y} \in \mathbb{R}^{K \times n}$ and $x \in \mathbb{R}^K$, the following property holds:

$$(\mathbf{TI}) \qquad \forall \mathcal{C} \subset \mathcal{H}\,, \qquad t_\alpha(\mathbf{Y} + x, \mathcal{C}) = t_\alpha(\mathbf{Y}, \mathcal{C}) \ .$$

This property is also shared by the concentration-based thresholds (8) and (9). Therefore, large values of non-zero means μ_k do not affect these thresholds. To understand the practical consequences of this point, let us consider the following informal and qualitative argumentation. If some coordinates of $(\mathbf{Y}^1_k)_{k \in \mathcal{C}}$ have a large mean relative to the noise (that is, a large signal-to-noise ratio, SNR), then the corresponding coordinates of $\overline{\mathbf{Y}}$ will have on average a large absolute value relative to the coordinates with zero mean, and the contribution of the former to the threshold will make the uncentered quantile significantly larger. By contrast, the centered quantile threshold is translation invariant and thus unaffected by the signal itself. Hence, in this situation, it is likely that the centered quantile threshold will be smaller. This effect is illustrated in the simulations presented in Section 4.

3. Step-down procedures. Single-step procedures can often be improved by iteration based on the *step-down* principle. Roughly, the idea is to repeat the multiple testing procedure with $\mathcal{H}$ replaced by $\mathcal{H} \backslash R_\alpha(\mathbf{Y})$, and to iterate this process as long as new coordinates are rejected. Again, consider a threshold family $\mathcal{T}_\alpha = (t_\alpha(\mathbf{Y}, \mathcal{C}), \mathcal{C} \subset \mathcal{H}, \mathbf{Y} \in \mathbb{R}^{K \times n})$ satisfying $(\mathbf{CB}_\alpha)$ and (**ND**).

DEFINITION 3.1. *Consider the nonincreasing sequence* $(\mathcal{C}_j, j \geq 0)$ *of subsets of* $\mathcal{H}$ *defined by*

$$\mathcal{C}_0 := \mathcal{H} \qquad \text{and} \qquad \forall j \geq 1, \mathcal{C}_j := \left\{ k \in \mathcal{C}_{j-1} \mid \left[\!\left[\overline{\mathbf{Y}}_k \right]\!\right] \leq t_\alpha(\mathbf{Y}, \mathcal{C}_{j-1}) \right\}$$

and let $\hat{\ell}$ *be the stopping rule* $\hat{\ell} = \min\{j \geq 1 \mid \mathcal{C}_j = \mathcal{C}_{j-1}\}$. *Then the* step-down multiple testing procedure *associated to* $\mathcal{T}_\alpha$ *rejects the hypotheses of the set* $\mathcal{H} \backslash \mathcal{C}_{\hat{\ell}}$, *that is,*

$$\left\{ k \in \mathcal{H} \mid \left[\!\left[\overline{\mathbf{Y}}_k \right]\!\right] > t_\alpha(\mathbf{Y}, \mathcal{C}_{\hat{\ell}}) \right\} \quad . \tag{15}$$

A very general result on step-down procedures was established in [16, Theorem 3], which we reproduce here with our notation:

THEOREM 3.2 (Romano and Wolf, 2005). *Let* $\mathcal{T}_\alpha$ *be a threshold family satisfying* (**ND**). *Then the FWER of the step-down procedure* (15) *is controlled by*

$$\mathbb{P}\left(\sup_{k \in \mathcal{H}_0} \left[\!\left[\overline{\mathbf{Y}}_k \right]\!\right] > t_\alpha\left(\mathbf{Y}, \mathcal{H}_0\right) \right) \quad .$$

Therefore, if $\mathcal{T}_\alpha$ *additionally satisfies* (**CB**$_\alpha$), *the FWER of the associated step-down procedure is upper bounded by* α.

A sketch of the proof can be given as follows: assume $\mathbf{Y}$ is such that $\sup_{k \in \mathcal{H}_0} \left[\!\left[\overline{\mathbf{Y}}_k \right]\!\right] \leq t_\alpha\left(\mathbf{Y}, \mathcal{H}_0\right)$. Then $\mathcal{H}_0 \subset \mathcal{C}_{j-1}$ implies $t_\alpha(\mathbf{Y}, \mathcal{C}_{j-1}) \geq t_\alpha(\mathbf{Y}, \mathcal{H}_0) \geq \sup_{k \in \mathcal{H}_0} \left[\!\left[\overline{\mathbf{Y}}_k \right]\!\right]$, and in turn $\mathcal{H}_0 \subset \mathcal{C}_j$ by definition of $\mathcal{C}_j$. By recursion, $\mathcal{H}_0$ is contained in $\mathcal{C}_j$ for every j and the step-down procedure therefore makes no type I error on the event $\left\{ \sup_{k \in \mathcal{H}_0} \left[\!\left[\overline{\mathbf{Y}}_k \right]\!\right] \leq t_\alpha\left(\mathbf{Y}, \mathcal{H}_0\right) \right\}$.

A direct consequence of Theorem 3.2 is that the step-down procedures based on any of the thresholds considered in the previous section (defined by either (7), (8), (9), (11), (12) or (13)) have a FWER controlled at level α (for (13), only in the two-sided setting). Note that the step-down procedure based on Bonferroni's threshold (7) is exactly Holm's procedure [9].

Parallel to the discussion at the end of Section 2.2, we can make a short qualitative comparison of the step-down procedure based on the uncentered quantile threshold (13) versus the step-down procedures based on the centered quantile thresholds (11) and (12) (in the two-sided setting). Again, if some coordinates have a large SNR, they certainly contribute to making the the uncentered quantile threshold significantly larger at the first step of the step-down procedure. This time, however, even if this first threshold is relatively large, it will still be able to rule out at the first step precisely those coordinates having the largest means. This will result in turn in an important improvement of the uncentered threshold at the second iteration, and so on, until all coordinates with a large SNR have been weeded out. Thus, the initial disadvantage of the uncentered threshold will be automatically corrected along the step-down iterations, and the final threshold will be close to the ideal resampled quantile $q_\alpha(\mathbf{Y}, \mathcal{H}_0)$ in the last iterations. By contrast, the centered thresholds (11) and (12) still suffer from the loss due to the remainder

term and level shrinkage along the step-down. In conclusion, contrary to the single-step case, we expect the uncentered procedure to eventually outperform the centered ones after some step-down iterations. This is in accordance with the simulations of Section 4.

At this point, it could seem that the uncentered step-down procedure is both simpler and more effective than the centered step-down ones, and thus should always be preferred. However, the above discussion gives us another insight: the step-down procedure based on the uncentered quantile should require more iterations to converge because the first iterations return inaccurate thresholds. In order to fix this drawback, we propose to use the leverage of the centered quantile thresholds for the first step—weeding 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 obtain the following new algorithm:

ALGORITHM 3.3 (Hybrid approach).

1. *Compute the threshold* $t_{\alpha,\text{quant+Bonf}}(\mathbf{Y},\mathcal{H})$ *defined by* (11) *with a given* $\delta \in (0,1)$ *and* $\alpha_0 \in (0,\alpha)$*, and consider* R_0 *the corresponding single-step procedure* (4)*.*
2. *If* $R_0 = \mathcal{H}$ *then stop and reject all the null hypotheses. Otherwise, consider the set of the remaining coordinates* $\mathcal{C}_0 = \mathcal{H}\backslash R_0$ *and apply on it the step-down procedure associated to the threshold* $t_{\alpha_0,\text{quant.uncent}}(\mathbf{Y},\mathcal{C})$ *defined by* (13) *(at level* α_0*).*

PROPOSITION 3.4. *Fix* $\delta \in (0,1)$ *and* $\alpha_0 \in (0,\alpha)$*. In the two-sided context, Algorithm 3.3 gives rise to a multiple testing procedure with a FWER upper bounded by* α*.*

Proposition 3.4 is proved in Section 6. What we expect is that Algorithm 3.3 essentially yields the same final result as the step-down procedure using the uncentered quantile (up to some negligible loss in the level by taking α_0 close to α), while requiring less iterations. In applications such as neuroimaging, where one iteration can take up to one day, this can result in a significant improvement.

4. Simulation study. The (matlab) code used to perform the simulations of this section is available on the first author's webpage (currently at url `http://www.di.ens.fr/~arlot/code/CRMTR.htm`).

4.1. *Framework.* 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 $F : \mathbb{T}_d^2 \to \mathbb{R}$ 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 $\mathbb{E}\left[G_t^2\right] = 1$ for all $t \in \mathbb{T}_d^2$.

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 flat Riemannian distance on the torus and C_b is a normalizing constant. We then compare the different thresholds

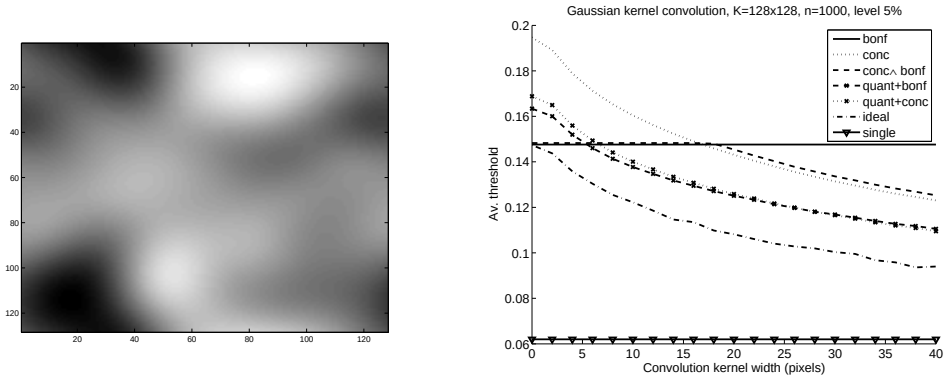

FIG 1. *Left: example of a* 128×128 *pixels image obtained by convolution of Gaussian white noise with a pseudo-Gaussian filter with width* $b = 18$ *pixels. Right: average thresholds obtained for the different approaches, see text.*

obtained by the methods proposed in this work for varying values of b. Remember the algorithms considered here have no prior knowledge on the specific form of the function F_b and would work in other more complex dependency contexts.

We consider the two-sided case only. In all of the simulations to come, we fix the following parameters: the dimension is $K = 128^2 = 16\,384$, the number of data points per sample is $n = 1\,000$, (hence significantly smaller than K), the width b takes even values in the range $[0, 40]$ ($b = 0$ is white noise; see the left-hand side of Figure 1 for an example of noise realization when $b = 18$). The target test level is $\alpha = 0.05$. We report the (empirical) expectation of each threshold over 250 draws of $\mathbf{Y}$.

For computation of the thresholds (9), (11) and (12), we have to pick some parameters $\delta \in (0, 1)$ and (for the two latter ones) $\alpha_0 < \alpha$. These parameters establish in each case a trade-off between a main term and a remainder term; generally speaking, as n grows one should choose $\delta \to 0$ and $\alpha_0 \to \alpha$ so that the level of the main resampled term tends to the target level α. In [1] it was suggested to take δ of order $1/n$ and $(1 - \frac{\alpha_0}{\alpha})$ of order $n^{-\gamma}$ for some $\gamma > 0$ to ensure that the remainder terms are indeed of lower order, but there is no exact recommendation for fixed n. In all the simulations below we decided to fix $\delta = (1 - \frac{\alpha_0}{\alpha}) = 0.1$, without particularly trying to optimize this choice. We noticed when varying these parameter values that the results were not overly sensitive to them. Finally, for all the thresholds, the resampling quantities (quantiles or expectations) are estimated by Monte-Carlo with 1 000 draws (but we disregarded the additional terms proposed in [1] to account for the Monte-Carlo random error).

4.2. *Simulations with unspecified alternative: single-step, translation invariant procedures.* We first study the performance of the multiple testing procedures which have a translation invariant threshold (**TI**), that is, the single-step procedures using thresholds (7), (8), (9), (11) and (12), denoted respectively by "bonf", "conc", "conc∧bonf", "quant+bonf" and "quant+conc". Their distribution do not depend on the true mean vector μ chosen to generate data, and we

fixed $\mu = 0$ without loss of generality. Provided the FWER constraint is satisfied, procedures with a smaller threshold are less conservative and hence more powerful.

We report on Figure 1 the (averaged) values of each threshold. 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. For comparison, we also included an estimation of the true quantile, that is, the $(1-\alpha)$ quantile of the distribution of $\sup_{k\in\mathcal{H}} |\overline{\mathbf{Y}}_k - \mu_k|$ (more precisely, an empirical quantile over 1 000 samples), denoted by "ideal". The exact threshold corresponding to $K = 1$ (test of a single coordinate Gaussian mean) is also included for comparison and is denoted by "single". In the context of this experiment, we also computed the threshold (10), that is, the raw symmetrized quantile obtained after empirical recentering of the data (for which no non-asymptotical theoretical results are available). This threshold was not reported in the plots, because it turns out to be so close to the true quantile that they are almost indistinguishable.

The overall conclusion of this first experiment is that the different thresholds proposed in this work are relevant: they improve over the Bonferroni threshold, provided the vector has strong enough correlations. As expected, the quantile approach appears to lead to tighter thresholds. (This might however not be always the case for smaller sample sizes because of the additional term.) One remaining advantage of the concentration approach is that the compound threshold (9) falls back on the Bonferroni threshold when needed, at the price of a minimal threshold increase. Finally, the remainder terms introduced by the theory in the centered quantile thresholds appears over-estimated, since the raw resampled quantile is in fact extremely close to the true quantile.

4.3. *Simulations with a specific alternative.* We consider the experiment of the previous section, with the following choice for the vector of true means:

$$\forall (i,j) \in \{0,\dots,127\}^2\,, \qquad \mu_{(i,j)} = \frac{(64-j)_+}{64} \times 20\, t_{\alpha,\mathrm{Bonf}}(\mathcal{H}) \ . \tag{16}$$

In this situation, half of the null hypotheses are true while the non-zero means are increasing linearly from $(5/16)t_{\alpha,\mathrm{Bonf}}(\mathcal{H})$ to $20t_{\alpha,\mathrm{Bonf}}(\mathcal{H})$. The thresholds obtained are drawn on Figure 2, along with the averaged power of the corresponding procedures, defined as the expected proportion of signal correctly detected (that is, averaged proportion of rejections among the false null hypotheses).

In this experiment we concentrated on the quantile-based thresholds. We picked the threshold (12) "quant+bonf" as a representative of the centered quantile approach, and its step-down counterpart denoted "s.d. quant+bonf". We compare these to the uncentered quantile threshold (13) denoted "quant. uncent." and its step-down version "s.d. quant. uncent.". Bonferroni's threshold and its step-down version "Holm" are included for comparison. The threshold denoted "ideal" is now derived from the $(1-\alpha)$ quantile of the distribution of $\sup_{k,\mu_k=0} |\overline{\mathbf{Y}}_k|$, and corresponds to the optimal threshold for FWER control.

The results of the experiment can be summarized as follows:

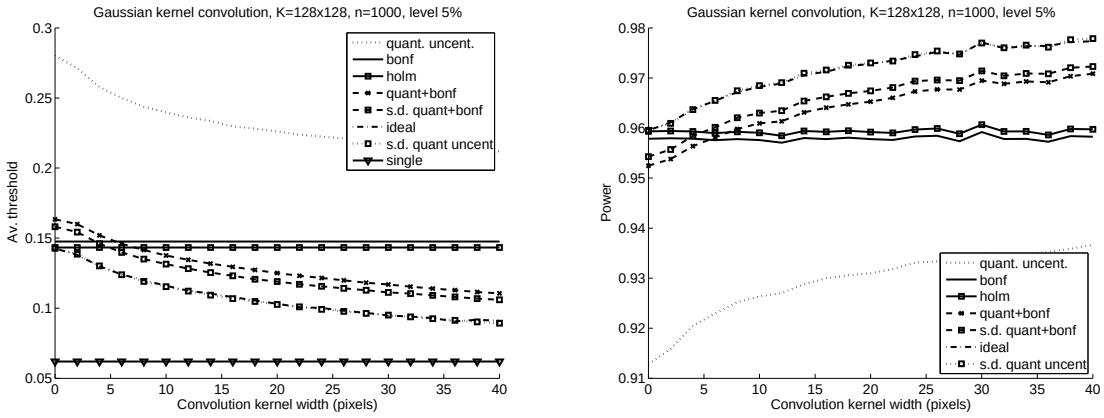

FIG 2. *Multiple testing problem with μ defined by* (16) *for different approaches, see text. Left: average thresholds. Right: average power.*

- the single-step centered quantile procedure "quant+bonf" outperforms Holm's procedure as soon as the coordinates of the vector are sufficiently correlated. Its step-down version "s.d. quant+bonf" performs even better, although the difference is not huge.
- the single-step procedure based on the uncentered quantile "quant.uncent" has the worst performance, confirming the qualitative analysis following Theorem 2.3.
- the step-down procedure based on the uncentered quantile "s.d. quant. uncent." seems on the other hand to be the most accurate among the procedures considered here, also on par with the qualitative analysis following Theorem 3.2.

The latter point must be balanced with computation time considerations. When K and n are large, the step-down algorithm for the uncentered quantile takes longer to compute because of its iterative nature, while the single-step centered quantile procedure "quant+bonf" provides a relatively good accuracy without iterating. This brings us to the next point.

4.4. *Hybrid approach.* We show here with a specific simulation study that the hybrid approach proposed in Algorithm 3.3 results in a speed-accuracy trade-off which is particularly noticeable when the mean values take on a large range.

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

$$\forall k \in \{0,\dots,128^2/2\}, \qquad f(k) = 0.5 t_{\alpha,\mathrm{Bonf}}(\mathcal{H}) \times \exp\left(\frac{128^2/2 - k}{128^2/2}\,\frac{r}{10}\,\log(10)\right), \tag{17}$$

and $f(k) = 0$ for the other values of k. In this situation, the $128^2/2$ non-zero means are increasing exponentially between $0.5\, t_{\alpha,\mathrm{Bonf}}(\mathcal{H}) 10^{r/10}$ and $0.5\, t_{\alpha,\mathrm{Bonf}}(\mathcal{H})$, where r is the dynamic range (in dB) of the signal.

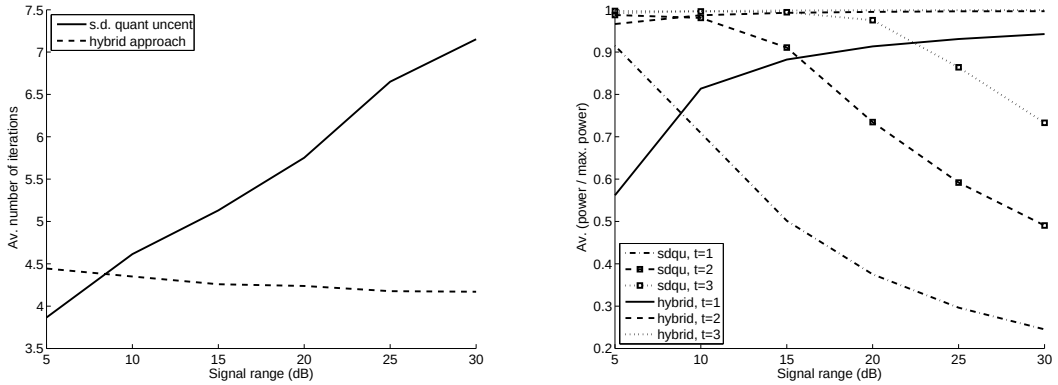

FIG 3. *Multiple testing problem with μ corresponding to* (17) *for the step-down procedure based on the uncentered quantile (sdqu) and the hybrid step-down approach. Left: average number of step-down iterations. Right: average of the ratio of power to maximum power when the step-down is stopped after at most t iterations. Here the maximum power is taken to be the power of "sdqu" without early stopping. (For the hybrid approach, the first step counts as 1 iteration.)*

We computed on Figure 3, for several values of r, the average number of iterations for the above step-down procedures, as well as their power when these procedures are stopped early after at most t iterations (such an early stopping is relevant in the case of a strict computation time constraint). We can sum up these results as follows:

- The hybrid approach needs on average significantly less iterations to converge as soon as $r \geq 10$.
- Stopping the hybrid approach procedure after only two iterations results in an average power that is virtually undistinguishable from the power obtained without early stopping, uniformly over values of r. By contrast, as r increases, more iterations are needed for the step-down quantile uncentered threshold in order to reach full power.

While these results are certainly specific to the particular simulation setup we used, they illustrate that the informal and qualitative analysis presented in Section 3 is correct when the signal (non-zero means) has a wide dynamic range. In particular, the fact that the hybrid approach gives already very satisfactory results after the two first iterations reinforces the interpretation that the first step (using the centered 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.

5. Discussion and concluding remarks.

5.1. *Discussion: FWER versus FDR in multiple testing.* It can legitimately be asked whether the FWER is an appropriate measure of type I error. The false discovery rate (FDR), introduced

in [2] 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 (for instance, 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 mean 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 [3, 4] for finding FDR controlling procedures adaptive to the proportion of true nulls and in [12] to find specific regions in random fields with application to neuroimaging data as well.

5.2. *Conclusion.* In this work, the main point was to introduce multiple testing procedures based on resampling thresholds (9), (11) and (12) coming from non-asymptotic confidence regions constructed in [1]. These confidence regions have theoretical control of the confidence level for any n, so the FWER of the corresponding multiple testing procedures is also controlled non-asymptotically. This issue is important in practice, because the sample size is often much smaller than the number of tests to perform ($K \gg n$). Nevertheless, as the simulations of Section 4 suggest, remainder terms in the thresholds—precisely introduced to deal with this non-asymptotic setting—are over-estimated by the theory and could probably be improved.

Even in the presence of these corrective terms, we showed through experiments that these thresholds are able to capture the unknown dependency structure of the data, and significantly outperform Holm's procedure when this dependency is strong enough. In comparison to exact randomization tests (based on an uncentered quantile), which also provide non-asymptotic level control, we argued that the empirical centering operation before random sign-flipping results in translation invariant thresholds. These thresholds are for this reason unaffected by the unknown signal, and thus relevant for testing already in the first iteration of the step-down algorithm. The method also applies to one-sided testing problems, where the uncentered approach is not theoretically justified as far as we know. Finally, the hybrid algorithm can approach the accuracy of the uncentered step-down threshold (which does not require corrective terms) while taking initially advantage of the centered threshold, resulting in a faster computation.

For practical purposes, it is certainly tempting to recommend using a (step-down) procedure

based on the raw, unmodified centered quantile without remainder terms (10): this would correspond to the principle of traditional resampling. To this extent, and to rephrase the discussion in [1], non-asymptotic theoretical results can also be understood in an asymptotic point of view, justifying the use of resampling (in a specific setting: Gaussian variables, test for the mean, Rademacher weights) for a regime that is not usually overed by traditional asymptotics (that is, dimension K_n increasing with n).

6. Proof of Proposition 3.4. First note that $q_{\alpha_0}(\mathbf{Y}, \mathcal{H}_0) \leq q_{\alpha_0}(\mathbf{Y}-\mu, \mathcal{H})$. From the proof of Theorem 3.2 in [1], with probability larger than $1-(\alpha-\alpha_0)$ we have

$$q_{\alpha_0}(\mathbf{Y}-\mu, \mathcal{H}) \leq t_{\alpha,\text{quant+Bonf}}(\mathbf{Y}, \mathcal{H}) \ .$$

Take $\mathbf{Y}$ in the event where the above inequality holds. If the global procedure rejects at least one true null hypothesis, let j_0 denote 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 $\sup_{k\in\mathcal{H}_0} \left|\overline{\mathbf{Y}}_k\right| \geq t_{\alpha,\text{quant+Bonf}}(\mathbf{Y}, \mathcal{H}) \geq q_{\alpha_0}(\mathbf{Y}-\mu, \mathcal{H}) \geq q_{\alpha_0}(\mathbf{Y}, \mathcal{H}_0)$,
- if $j_0 \geq 1$, $\sup_{k\in\mathcal{H}_0} \left|\overline{\mathbf{Y}}_k\right| \geq t_{\alpha_0,\text{quant.uncent}}(\mathbf{Y}, \mathcal{C}_{j_0-1})$ and $\mathcal{H}_0 \subset \mathcal{C}_{j_0-1}$ (from the definition of j_0), so that $\sup_{k\in\mathcal{H}_0} \left|\overline{\mathbf{Y}}_k\right| \geq t_{\alpha_0,\text{quant.uncent}}(\mathbf{Y}, \mathcal{H}_0) = q_{\alpha_0}(\mathbf{Y}, \mathcal{H}_0)$.

In both cases, $\sup_{k\in\mathcal{H}_0} \left|\overline{\mathbf{Y}}_k\right| \geq q_{\alpha_0}(\mathbf{Y}, \mathcal{H}_0)$, which occurs with probability smaller than α_0. ■

Acknowledgements. We would like to thank the two referees and the AE for their insight, leading in particular to a more rational organization of the paper.

REFERENCES

[1] S. Arlot, G. Blanchard, and É. Roquain. Some non-asymptotic results on resampling in high dimension, I: Confidence regions. *Ann. Statist.*, 2009. To appear.

[2] 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.

[3] G. Blanchard and É. Roquain. Two simple sufficient conditions for FDR control. *Electron. J. Stat.*, 2:963–992, 2008.

[4] G. Blanchard and É. Roquain. Adaptive FDR control under independence and dependence. *J. Mach. Learn. Res.*, 2009. To appear.

[5] F. Darvas, M. Rautiainen, D. Pantazis, S. Baillet, H. Benali, J. Mosher, L. Garnero, and R. Leahy. Investigations of dipole localization accuracy in MEG using the bootstrap. *NeuroImage*, 25:355–368, 2005.

[6] C. Durot and Y. Rozenholc. An adaptive test for zero mean. *Math. Methods Statist.*, 15(1):26–60, 2006.

[7] R. A. Fisher. *The Design of Experiments.* Oliver and Boyd, Edinburgh, 1935.

[8] Y. Ge, S. Dudoit, and T. P. Speed. Resampling-based multiple testing for microarray data analysis. *Test*, 12(1):1–77, 2003.

[9] S. Holm. A simple sequentially rejective multiple test procedure. *Scand. J. Statist.*, 6(2):65–70, 1979.

[10] K. Jerbi, J.-P. Lachaux, K. N'Diaye, D. Pantazis, R. M. Leahy, L. Garnero, and S. Baillet. Coherent neural representation of hand speed in humans revealed by MEG imaging. *PNAS*, 104(18):7676–7681, 2007.

[11] D. Pantazis, T. E. Nichols, S. Baillet, and R. M. Leahy. A comparison of random field theory and permutation methods for statistical analysis of MEG data. *NeuroImage*, 25:383–394, 2005.

[12] M. Perone Pacifico, I. Genovese, I. Verdinelli, and L. Wasserman. False discovery control for random fields. *J. Amer. Statist. Assoc.*, 99(468):1002–1014, 2004.

[13] K. S. Pollard and M. J. van der Laan. Choice of a null distribution in resampling-based multiple testing. *J. Statist. Plann. Inference*, 125(1-2):85–100, 2004.

[14] J. P. Romano. Bootstrap and randomization tests of some nonparametric hypotheses. *Ann. Statist.*, 17(1):141–159, 1989.

[15] J. P. Romano. On the behavior of randomization tests without a group invariance assumption. *J. Amer. Statist. Assoc.*, 85(411):686–692, 1990.

[16] J. P. Romano and M. Wolf. Exact and approximate stepdown methods for multiple hypothesis testing. *J. Amer. Statist. Assoc.*, 100(469):94–108, 2005.

[17] J. P. Romano and M. Wolf. Control of generalized error rates in multiple testing. *Ann. Statist.*, 35(4):1378–1408, 2007.

[18] P. H. Westfall and S. S. Young. *Resampling-Based Multiple Testing*. Wiley, 1993. Examples and Methods for *P*- Value Adjustment.

[19] D. Yekutieli and Y. 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
CNRS ; Willow Project-Team
Laboratoire d'Informatique de l'Ecole Normale Superieure
(CNRS/ENS/INRIA UMR 8548)
45, rue d'Ulm, 75230 Paris, France
E-mail: sylvain.arlot@ens.fr

Gilles Blanchard
Weierstrass institute for applied stochastics and analysis
Mohrenstrasse 39, 10117 Berlin, Germany, and
Fraunhofer FIRST.IDA, Kekuléstr. 7, 12489 Berlin,
Germany
E-mail: blanchar@wias-berlin.de

Etienne Roquain
University of Paris 6,LPMA,
4, Place Jussieu, 75252 Paris cedex 05, France
E-mail: etienne.roquain@upmc.fr