Resampling-based confidence regions and multiple tests for a correlated random vector

Sylvain Arlot, Gilles Blanchard, Etienne Roquain

Resampling-based confidence regions and multiple tests for a correlated random vector

Sylvain Arlot
Univ Paris-Sud, Laboratoire de Mathématiques d'Orsay,
Orsay Cedex, F-91405; CNRS, Orsay cedex, F-91405
sylvain.arlot@math.u-psud.fr
INRIA Futurs, Projet Select

Gilles Blanchard
Fraunhofer FIRST.IDA, Berlin, Germany,
blanchar@first.fraunhofer.de

Étienne Roquain
INRA Jouy-en-Josas, unité MIG,
78 352 Jouy-en-Josas Cedex, France,
etienne.roquain@jouy.inra.fr

22nd January 2007

Abstract

We derive non-asymptotic confidence regions for the mean of a random vector whose coordinates have an unknown dependence structure. The random vector is supposed to be either Gaussian or to have a symmetric bounded distribution, and we observe n i.i.d copies of it. The confidence regions are built using a data-dependent threshold based on a weighted bootstrap procedure. We consider two approaches, the first based on a concentration approach and the second on a direct boostrapped quantile approach. The first one allows to deal with a very large class of resampling weights while our results for the second are restricted to Rademacher weights. However, the second method seems more accurate in practice. Our results are motivated by multiple testing problems, and we show on simulations that our procedures are better than the Bonferroni procedure (union bound) as soon as the observed vector has sufficiently correlated coordinates.

1 Introduction

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 a dimension K possibly much greater than n. 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 for μ, of the form:

$$\left\{x \in \mathbb{R}^K \text{ s.t. } \phi\left(\overline{\mathbf{Y}} - x\right) \leq t_\alpha(\mathbf{Y})\right\}, \tag{1}$$

where $\phi : \mathbb{R}^K \to \mathbb{R}$ is a measurable function fixed in advance by the user (measuring a kind of distance), $\alpha \in (0,1)$, $t_\alpha : \left(\mathbb{R}^K\right)^n \to \mathbb{R}$ is a measurable data-dependent threshold, and $\overline{\mathbf{Y}} = \frac{1}{n}\sum_{i=1}^n \mathbf{Y}^i$ is the empirical mean of the sample $\mathbf{Y}$.

The form of the confidence region (1) is motivated by the following multiple testing problem: if we want to test simultaneously for all $1 \leq k \leq K$ the hypotheses $H_{0,k} = \{\mu_k \leq 0\}$ against $H_{1,k} = \{\mu_k > 0\}$, we propose to reject the $H_{0,k}$ corresponding to

$$\left\{1 \leq k \leq K \text{ s.t. } \overline{\mathbf{Y}}_k > t_\alpha(\mathbf{Y})\right\}.$$

The error of this multiple testing procedure can be measured by the family-wise error rate defined by the probability that at least one hypothesis is wrongly rejected. Here, this error will be strongly (i.e. for any value of μ) controlled by α as soon as the confidence region (1) for μ with $\phi = \sup(\cdot)$ is of level at least $1-\alpha$. Indeed, for all μ,

$$\begin{aligned}\mathbb{P}\left(\exists k \text{ s.t. } \overline{\mathbf{Y}}_k > t_\alpha(\mathbf{Y}) \text{ and } \mu_k \leq 0\right) &\leq \mathbb{P}\left(\exists k \text{ s.t. } \overline{\mathbf{Y}}_k - \mu_k > t_\alpha(\mathbf{Y})\right) \\ &= \mathbb{P}\left(\sup_k \left\{\overline{\mathbf{Y}}_k - \mu_k\right\} > t_\alpha(\mathbf{Y})\right).\end{aligned}$$

The same reasoning with $\phi = \sup|\cdot|$ allows us to test $H_{0,k} = \{\mu_k = 0\}$ against $H_{1,k} = \{\mu_k \neq 0\}$, by choosing the rejection set $\{1 \leq k \leq K \text{ s.t. } \left|\overline{\mathbf{Y}}_k\right| > t_\alpha(\mathbf{Y})\}$.

While this goal is statistical in motivation, to tackle it we want to follow a point of view inspired from learning theory, in the following sense: first, we want a non-asymptotical result valid for any fixed K and n, and secondly, we want to make no assumptions 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).

The ideal threshold t_α in (1) is obviously the $1-\alpha$ quantile of the distribution of $\phi\left(\overline{\mathbf{Y}} - \mu\right)$. However, this quantity depends on the unknown

dependency structure of the coordinates of $\mathbf{Y}^i$ and is therefore itself unknown.

We propose here to approach t_α by some resampling scheme: the heuristics of the resampling method (introduced by Efron [Efr79]) is that the distribution of $\overline{\mathbf{Y}} - \mu$ is "close" to the one of

$$\overline{\mathbf{Y}}_{[W-\overline{W}]} := \frac{1}{n}\sum_{i=1}^{n}(W_i - \overline{W})\mathbf{Y}^i = \frac{1}{n}\sum_{i=1}^{n} W_i(\mathbf{Y}^i - \overline{\mathbf{Y}}) = \overline{\left(\mathbf{Y} - \overline{\mathbf{Y}}\right)}_{[W]},$$

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

Following this idea, we propose two different approaches to obtain non-asymptotic confidence regions in this paper:

1. The expectations of $\phi\left(\overline{\mathbf{Y}} - \mu\right)$ and $\phi\left(\overline{\mathbf{Y}}_{[W-\overline{W}]}\right)$ can be precisely compared, and the processes $\phi\left(\overline{\mathbf{Y}} - \mu\right)$ and $\mathbb{E}\left[\phi\left(\overline{\mathbf{Y}}_{[W-\overline{W}]}\right)|\mathbf{Y}\right]$ concentrate well around their expectations.

2. The $1-\alpha$ quantile of the distribution of $\phi\left(\overline{\mathbf{Y}}_{[W-\overline{W}]}\right)$ conditionally to $\mathbf{Y}$ is close to the one of $\phi\left(\overline{\mathbf{Y}} - \mu\right)$.

Method 1 above is closely related to the Rademacher complexity approach in learning theory, and our results in this direction are heavily inspired by the work of Fromont [Fro04], who studies general resampling schemes in a learning theoretical setting. It may also be seen as a generalization of cross-validation methods. For method 2, we will restrict ourselves specifically to Rademacher weights in our analysis, because we use a symmetrization trick. Although this kind of method is not new in the resampling literature, to our knowledge our result is the first to provide a non-asymptotic analysis based on empirical resampled quantiles.

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

- Vectors, such as data vectors $\mathbf{Y}^i = (\mathbf{Y}^i_k)_{1\le k\le K}$, will always be column vectors. Thus, $\mathbf{Y}$ is a $K \times n$ data matrix.

- If $\mu \in \mathbb{R}^K$, $\mathbf{Y} - \mu$ is the matrix obtained by subtracting μ to each (column) vector of $\mathbf{Y}$. If $c \in \mathbb{R}$ and $W \in \mathbb{R}^n$, $W - c = (W_i - c)_{1\le i\le n} \in \mathbb{R}^n$.

- $\overline{\Phi}$ is the standard Gaussian upper tail function.

Several properties may be assumed for the function $\phi : \mathbb{R}^K \to \mathbb{R}$:

- 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 $\max_k\{|x_k|\}$ otherwise.

Finally, different assumptions on the generating distribution of $\mathbf{Y}$ can be made:

(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, our primary focus is on the Gaussian framework (GA), because the corresponding results will be more accurate.

The paper is organized as follows: Section 2 deals with the concentration method with general weights. In Section 3, we propose an approach based on resampling quantiles, with Rademacher weights. We illustrate our methods in Section 4 with a simulation study. The proofs of our results are given in Section 5.

2 Confidence region using concentration

In this section, we consider a general $\mathbb{R}^n$-valued *resampling weight vector* W, satisfying the following properties: W is independent of $\mathbf{Y}$, for all $i \in \{1, \ldots, n\}$ $\mathbb{E}[W_i^2] < \infty$, the $(W_i)_{1 \leq i \leq n}$ have an exchangeable distribution (*i.e.* invariant under every permutation of the indices) and the coordinates of W are not a.s. equal, *i.e.* $\mathbb{E}\left|W_1 - \overline{W}\right| > 0$. Several examples of resampling weight vectors are given in Section 2.3, where we also tackle the question of choosing a resampling.

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 Tab. 1:

$$A_W := \mathbb{E}\left|W_1 - \overline{W}\right| \tag{2}$$

$$B_W := \mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^{n}\left(W_i - \overline{W}\right)^2\right)^{\frac{1}{2}}\right] \tag{3}$$

$$C_W := \left(\frac{n}{n-1}\mathbb{E}\left[\left(W_1 - \overline{W}\right)^2\right]\right)^{\frac{1}{2}} \tag{4}$$

$$D_W := a + \mathbb{E}\left|\overline{W} - x_0\right| \quad \text{if } \forall i,\ |W_i - x_0| = a \text{ a.s. (with } a > 0, x_0 \in \mathbb{R}). \tag{5}$$

Note that under our assumptions, these quantities are positive. Moerover, if the weights are i.i.d., $C_W = \mathrm{Var}(W_1)^{\frac{1}{2}}$. We can now state the main result of this section:

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

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

$$\phi\left(\overline{\mathbf{Y}} - \mu\right) < \frac{\mathbb{E}\left[\phi\left(\overline{\mathbf{Y}}_{[W-\overline{W}]}\right)\left|\mathbf{Y}\right.\right]}{B_W} + \|\sigma\|_p \,\overline{\Phi}^{-1}(\alpha/2)\left[\frac{C_W}{nB_W} + \frac{1}{\sqrt{n}}\right] \tag{6}$$

holds with probability at least $1-\alpha$, where σ is the vector $[\mathrm{Var}^{1/2}(\mathbf{Y}_k^1)]_k$. The same bound holds for the lower deviations, i.e. with inequality (6) reversed and the additive term replaced by its opposite.

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

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

holds with probability at least $1-\alpha$. If moreover the weights satisfy the assumption of (5), then

$$\phi\left(\overline{\mathbf{Y}} - \mu\right) > \frac{\mathbb{E}\left[\phi\left(\overline{\mathbf{Y}}_{[W-\overline{W}]}\right)\left|\mathbf{Y}\right.\right]}{D_W} - \frac{M}{\sqrt{n}}\sqrt{1 + \frac{A_W^2}{D_W^2}}\sqrt{2\log(1/\alpha)}$$

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

If there exists a deterministic threshold t_α such that $\mathbb{P}(\phi\left(\overline{\mathbf{Y}}-\mu\right) > t_\alpha) \leq \alpha$, the following corollary establishes that we can combine the above concentration threshold with t_α to get a new threshold almost better than both.

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 $\mathbb{P}\left(\phi\left(\overline{\mathbf{Y}}-\mu\right) > t_{\alpha(1-\delta)}\right) \leq \alpha(1-\delta)$. Then with probability at least $1-\alpha$, $\phi\left(\overline{\mathbf{Y}}-\mu\right)$ is upper bounded by the minimum between $t_{\alpha(1-\delta)}$ and*

$$\frac{\mathbb{E}\left[\phi\left(\overline{\mathbf{Y}}_{[W-\overline{W}]}\right)\left|\mathbf{Y}\right.\right]}{B_W} + \frac{\|\sigma\|_p}{\sqrt{n}}\overline{\Phi}^{-1}\left(\frac{\alpha(1-\delta)}{2}\right) + \frac{\|\sigma\|_p C_W}{nB_W}\overline{\Phi}^{-1}\left(\frac{\alpha\delta}{2}\right). \quad (7)$$

Remark 2.3.

1. *Corollary 2.2 is a consequence of the proof of Theorem 2.1, rather than of the theorem itself. The point here is that $\mathbb{E}\left[\phi\left(\overline{\mathbf{Y}}_{[W-\overline{W}]}\right)\left|\mathbf{Y}\right.\right]$ is almost deterministic, because it concentrates at the rate n^{-1} $(= o(n^{-1/2}))$.*

2. *For instance, if $\phi = \sup(\cdot)$ (resp. $\sup|\cdot|$), Corollary 2.2 may be applied with t_α equal to the classical Bonferroni threshold for multiple testing (obtained using a simple union bound over coordinates)*

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

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 4).

The proof of Theorem 2.1 involves results which are of self interest: the comparison between the expectations of the two processes $\mathbb{E}\left[\phi\left(\overline{\mathbf{Y}}_{[W-\overline{W}]}\right)\left|\mathbf{Y}\right.\right]$ and $\phi\left(\overline{\mathbf{Y}}-\mu\right)$ and the concentration of these processes around their means. This is examinated in the two following subsections. The last subsection gives some elements for a wise choice of resampling weight vectors among several classical examples.

2.1 Comparison in expectation

In this section, we compare $\mathbb{E}\phi\left(\overline{\mathbf{Y}}_{[W-\overline{W}]}\right)$ and $\mathbb{E}\phi\left(\overline{\mathbf{Y}}-\mu\right)$. We note that these expectations exist in the Gaussian and the bounded case 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* W *a resampling weight vector. Then, for any measurable positive-homogeneous function* $\phi : \mathbb{R}^K \to \mathbb{R}$*, we have the following equality*

$$B_W \mathbb{E}\phi\left(\overline{\mathbf{Y}} - \mu\right) = \mathbb{E}\phi\left(\overline{\mathbf{Y}}_{[W-\overline{W}]}\right). \tag{8}$$

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 in Tab. 1.*
2. *In a non-Gaussian framework, the constant* B_W *is still relevant, at least asymptotically: in their Theorem 3.6.13, Van der Vaart and Wellner [VdVW96] use the limit of* B_W *when* n *goes to infinity as a normalizing constant.*

When the sample is only symmetric we obtain the following inequalities :

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

(i) We have the general following lower bound :

$$A_W \mathbb{E}\phi\left(\overline{\mathbf{Y}} - \mu\right) \leq \mathbb{E}\phi\left(\overline{\mathbf{Y}}_{[W-\overline{W}]}\right). \tag{9}$$

(ii) Moreover, if the weights satisfy the assumption of (5)*, we have the following upper bound*

$$D_W \mathbb{E}\phi\left(\overline{\mathbf{Y}} - \mu\right) \geq \mathbb{E}\phi\left(\overline{\mathbf{Y}}_{[W-\overline{W}]}\right). \tag{10}$$

Remark 2.7.

1. *The bounds* (9) *and* (10) *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 2.3).*
2. *When* $\mathbf{Y}$ *is not assumed to be symmetric and* $\overline{W} = 1$ *a.s., Proposition 2 in [Fro04] shows that* (9) *holds with* $\mathbb{E}(W_1 - \overline{W})_+$ *instead of* A_W*. Therefore, the symmetry of the sample allows us to get a tighter result (for instance twice sharper with Efron or Random hold-out* (q) *weights).*

2.2 Concentration around the expectations

In this section we present concentration results for the two processes $\phi\left(\overline{\mathbf{Y}}-\mu\right)$ and $\mathbb{E}\left[\phi\left(\overline{\mathbf{Y}}_{[W-\overline{W}]}\right)|\mathbf{Y}\right]$ in the Gaussian framework.

Proposition 2.8. *Let $p \in [1,+\infty]$, $\mathbf{Y}$ a sample satisfying (GA) and let σ be the vector $[\mathrm{Var}^{1/2}(\mathbf{Y}_k^1)]_k$. Let $\phi : \mathbb{R}^K \to \mathbb{R}$ be any subadditive function, bounded by the p-norm.*

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

$$\phi\left(\overline{\mathbf{Y}}-\mu\right) < \mathbb{E}\phi\left(\overline{\mathbf{Y}}-\mu\right) + \frac{\|\sigma\|_p \,\overline{\Phi}^{-1}(\alpha/2)}{\sqrt{n}}\,, \tag{11}$$

and the same bound holds for the corresponding lower deviations.

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

$$\mathbb{E}\left[\phi\left(\overline{\mathbf{Y}}_{[W-\overline{W}]}\right)|\mathbf{Y}\right] < \mathbb{E}\phi\left(\overline{\mathbf{Y}}_{[W-\overline{W}]}\right) + \frac{\|\sigma\|_p \, C_W \overline{\Phi}^{-1}(\alpha/2)}{n}, \tag{12}$$

and the same bound holds for the corresponding lower deviations.

The first bound (11) with a remainder in $n^{-1/2}$ is classical. The last one (12) is much more interesting since it enlights one of the key properties of the resampling idea: the "stabilization". Indeed, the resampling quantity $\mathbb{E}\left[\phi\left(\overline{\mathbf{Y}}_{[W-\overline{W}]}\right)|\mathbf{Y}\right]$ concentrates around its expectation at the rate $C_W n^{-1} = o\left(n^{-1/2}\right)$ for most of the weights (see Section 2.3 and Tab. 1 for more details). Thus, compared to the original process, it is almost deterministic and equal to $B_W \mathbb{E}\phi\left(\overline{\mathbf{Y}}-\mu\right)$.

Remark 2.9. *Combining expression* (8) *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 $1-\alpha$:*

$$\mathbb{E}\left\|\overline{\mathbf{Y}}-\mu\right\|_p < \frac{\mathbb{E}\left[\left\|\overline{\mathbf{Y}}_{[W-\overline{W}]}\right\|_p \middle| \mathbf{Y}\right]}{B_W} + \frac{\|\sigma\|_p \, C_W}{n B_W}\overline{\Phi}^{-1}(\alpha/2)\ , \tag{13}$$

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

Efron	$2\left(1-\frac{1}{n}\right)^n = A_W \leq B_W \leq \sqrt{\frac{n-1}{n}} \quad C_W = 1$
Efr., $n \to +\infty$	$\frac{2}{e} \leq A_W \leq B_W \leq 1 = C_W$
Rademacher	$1-\frac{1}{\sqrt{n}} \leq A_W \leq B_W \leq \sqrt{1-\frac{1}{n}} \quad C_W = 1 \quad D_W \leq 1+\frac{1}{\sqrt{n}}$
Rad., $n \to +\infty$	$A_W = B_W = C_W = D_W = 1$
R. h.-o. (q)	$A_W = 2\left(1-\frac{q}{n}\right) \quad B_W = \sqrt{\frac{n}{q}-1}$
R. h.-o. (q)	$C_W = \sqrt{\frac{n}{n-1}}\sqrt{\frac{n}{q}-1} \quad D_W = \frac{n}{2q} + \left\lvert 1-\frac{n}{2q}\right\rvert$
R. h.-o. $(n/2)$ $(2\vert n)$	$A_W = B_W = D_W = 1 \quad C_W = \sqrt{\frac{n}{n-1}}$
Leave-one-out	$\frac{2}{n} = A_W \leq B_W = \frac{1}{\sqrt{n-1}} \quad C_W = \frac{\sqrt{n}}{n-1} \quad D_W = 1$

Table 1: Resampling constants for classical resampling weight vector.

2.3 Resampling weight vectors

In this section, we consider the question of choosing some appropriate 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**: W has a multinomial distribution with parameters $(n; n^{-1}, \ldots, n^{-1})$.

3. **Random hold-out** (q) (R. h.-o.), $q \in \{1, \ldots, n\}$: $W_i = \frac{n}{q}\mathbb{1}_{i\in I}$, where I is uniformly distributed on subsets of $\{1, \ldots, n\}$ 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 $2|n$). When $q = n-1$, these weights are called **leave-one-out** weights.

For these classical weights, exact or approximate values for the quantities A_W, B_W, C_W and D_W (defined by equations (2) to (5)) can be easily derived (see Tab. 1). However, an exact computation of the resampling estimates $\mathbb{E}\left[\phi\left(\overline{\mathbf{Y}}_{[W-\overline{W}]}\right)|\mathbf{Y}\right]$ using these weights would be time-consuming when n is large. The more standard way to solve this problem is to compute resampling quantities by Monte-Carlo simulations, *i.e.* picking up a small number of weight vectors (see [Hal92], appendix II for a discussion). But we did not yet investigate the analysis of the corresponding thresholds.

Another way to solve this computation time problem is to consider a regular partition $(B_j)_{1\leq j\leq V}$ of $\{1, \ldots, n\}$ (where $V \in \{2, \ldots, n\}$ and $V|n$), and to define the weights $W_i = \frac{V}{V-1}\mathbb{1}_{i\notin B_J}$ with J uniformly distributed on $\{1, \ldots, V\}$. These weights are called the **(regular)** V**-fold cross validation** weights (V-f. c.v.), which are no longer exchangeable but still "piece-wise

exchangeable". Considering the process $(\widetilde{\mathbf{Y}}^j)_{1\leq j\leq K}$ where $\widetilde{\mathbf{Y}}^j = \frac{V}{n}\sum_{i\in B_j}\mathbf{Y}^i$ 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:

$$A_W = \frac{2}{V}\,; \quad B_W = \frac{1}{\sqrt{V-1}}\,; \quad C_W = \sqrt{n}(V-1)^{-1}\,; \quad D_W = 1\ .$$

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.

Note that in the Gaussian framework of (13), V-fold cross-validation weights approximate the estimation risk $\mathbb{E}\left\|\overline{\mathbf{Y}}-\mu\right\|_p$ by $\frac{\sqrt{V-1}}{V^2}\sum_{j=1}^{V}\left\|\widetilde{\mathbf{Y}}^{(-j)}-\widetilde{\mathbf{Y}}^j\right\|_p$, where $\widetilde{\mathbf{Y}}^{(-j)}$ is the mean of the $(\widetilde{\mathbf{Y}}^\ell)_{\ell\neq j}$; which bears a strong analogy with the usual cross-validation philosophy. Actually, the "classical" leave-one-out estimator $\frac{1}{n}\sum_{i=1}^{n}\left\|\widetilde{\mathbf{Y}}^{(-i)}-\mathbf{Y}^i\right\|_p$ approximates a different quantity, the prediction risk $\mathbb{E}\left\|\overline{\mathbf{Y}}-\mathbf{Y}^{n+1}\right\|_p$ for a new independent vector $\mathbf{Y}^{n+1}$. However, under (GA) the two types of risk are proportional, $\sqrt{n+1}\mathbb{E}\left\|\overline{\mathbf{Y}}-\mu\right\|_p = \mathbb{E}\left\|\overline{\mathbf{Y}}-\mathbf{Y}^{n+1}\right\|_p$; taking into account this scaling we conclude that our estimator (with $V=n$) coincides with the classical leave-one-out (up to the factor $\sqrt{1-1/n^2}\sim 1$). To guide our choice for a specific resampling scheme, the first comparison point is that $t_{\alpha,W}(\mathbf{Y})$ should be an accurate upper bound of the ideal threshold. Under the Gaussian assumption, in view of (6), $C_W B_W^{-1}$ appears as a relevant accuracy index for $t_{\alpha,W}$. However, a second comparison point is the price of an exact computation of $t_{\alpha,W}$ in practice. Since one must consider each possible weight vector to compute exactly the threshold, we use the cardinality of the support of $\mathcal{L}(W)$ as a complexity index.

As shown in Tab. 2, there is an *accuracy-complexity trade-off* for choosing the weights. Since for all exchangeable weights $C_W B_W^{-1} \geq \sqrt{n/(n-1)}$, R. h.-o.$(n/2)$ and leave-one-out weights are optimal for accuracy (Rademacher and Efron being "almost optimal"). On the other hand, V-fold c.-v. is less accurate, losing a factor $\sqrt{(n-1)/(V-1)}$. On the computational viewpoint, the leave-one-out is the only reasonable exchangeable procedure (at least when n and K are large), and V-f. c.v. looks even more attractive. Considering that $t_{\alpha,W}$ involves the sum of terms of order $C_W B_W^{-1} n^{-1}$ and $n^{-1/2}$, the best choice of V should be rather small for most applications. We do not give here any universal optimal V since it does not exist, but we suggest to use Tab. 2 to choose it.

Resampling	$C_W B_W^{-1}$ (accuracy)	Card (supp $\mathcal{L}(W)$) (complexity)
Efron	$\leq \frac{1}{2}\left(1-\frac{1}{n}\right)^{-n} \xrightarrow[n\to\infty]{} \frac{e}{2}$	n^n
Rademacher	$\leq \left(1-n^{-1/2}\right)^{-1} \xrightarrow[n\to\infty]{} 1$	2^n
R. h.-o. $(n/2)$	$= \sqrt{\frac{n}{n-1}} \xrightarrow[n\to\infty]{} 1$	$\binom{n}{n/2} \propto n^{-1/2}2^n$
Leave-one-out	$= \sqrt{\frac{n}{n-1}} \xrightarrow[n\to\infty]{} 1$	n
regular V-fold c.-v.	$= \sqrt{\frac{n}{V-1}}$	V

Table 2: Choice of the resampling weight vectors : accuracy-complexity tradeoff.

3 Confidence region using resampled quantiles

In the previous section we have shown how to derive non-asymptotic confidence regions for the mean of a Gaussian (resp. bounded) vector with unknown correlation structure; for this we used a concentration property of the quantities $\phi(\overline{\mathbf{Y}}-\mu)$ and $\mathbb{E}\left[\phi(\overline{\mathbf{Y}}_{[W-\overline{W}]})|\mathbf{Y}\right]$ around their mean. The Gaussian (resp. McDiarmid's) concentration property allowed us to bound deviations from this mean by the deviations of a suitably scaled normal (resp. subgaussian) variable. Through this approach, the level of the confidence region is rigorously controlled for any fixed sample size.

However, the obtained confidence regions are somewhat unsatisfying because they appear to be too conservative in practice. The principal reason for this is that $\phi(\overline{\mathbf{Y}}-\mu)$ is of course not a Gaussian variable (even when $\mathbf{Y}$ is) . Therefore, in spite of the power of the Gaussian concentration property, using Gaussian tails as a bound for the deviations of the above non-Gaussian variable must necessarily result in losing some slack.

On the other hand, in most applications of resampling procedures, it is common to estimate the quantiles of a variable like $\phi\left(\overline{\mathbf{Y}}-\mu\right)$ by the quantiles of the corresponding resampled distribution $\mathcal{L}\left(\phi\left(\overline{\mathbf{Y}}_{[W-\overline{W}]}\right)|\mathbf{Y}\right)$, and to use these quantiles to construct a confidence region. Again, while many asymptotic results are available to justify this method (for instance [VdVW96]), our goal here is to derive a non-asymptotic region based on a similar approach for which the confidence level is proved to hold for any fixed sample size.

For this we apply a principle that is close in spirit to exact tests, *i.e.* by taking advantage of an invariance property (here symmetry around the mean) of the initial distribution and using a resampling scheme that respects this invariance. For this reason the scope of the current section is far less

general: instead of covering generic resampling weights, we only consider the particular Rademacher resampling scheme. Let us define for a function ϕ the resampled empirical quantile:

$$q_\alpha(\phi, \mathbf{Y}) = \inf\left\{x \in \mathbb{R} \text{ s.t. } \mathbb{P}_W\left[\phi(\overline{\mathbf{Y}}_{[W]}) > x\right] \leq \alpha\right\},$$

wherein W is an i.i.d Rademacher weight vector. We now state the main technical result of this section:

Proposition 3.1. *Fix $\delta, \alpha \in (0,1)$. Let $\mathbf{Y}$ be a data sample satisfying assumption (SA). Let $f : \left(\mathbb{R}^K\right)^n \to [0,\infty)$ be a nonnegative (measurable) function on the set of data samples. Let ϕ be a* nonnegative*, subadditive, positive-homogeneous function. Denote $\widetilde{\phi}(x) = \max\left(\phi(x), \phi(-x)\right)$. Finally, for $\eta \in (0,1)$, denote*

$$\overline{\mathcal{B}}(n,\eta) = \min\left\{k \in \{0,\dots,n\} \text{ s.t. } 2^{-n}\sum_{i=k+1}^{n}\binom{n}{i} < \eta\right\},$$

the upper quantile function of a binomial $(n, \frac{1}{2})$ variable. Then we have:

$$\mathbb{P}\left[\phi(\overline{\mathbf{Y}} - \mu) > q_{\alpha(1-\delta)}\left(\phi, \mathbf{Y} - \overline{\mathbf{Y}}\right) + f(\mathbf{Y})\right] \leq \alpha + \mathbb{P}\left[\widetilde{\phi}(\overline{\mathbf{Y}} - \mu) > \frac{n}{2\overline{\mathcal{B}}\left(n, \frac{\alpha\delta}{2}\right) - n} f(\mathbf{Y})\right]$$

Remark 3.2. *By Hoeffding's inequality, $\frac{n}{2\overline{\mathcal{B}}\left(n,\frac{\alpha\delta}{2}\right)-n} \geq \left(\frac{n}{2\ln\left(\frac{2}{\alpha\delta}\right)}\right)^{1/2}$.*

By iteration of this proposition we obtain the following corollary:

Corollary 3.3. *Fix J a positive integer, $(\alpha_i)_{i=0,\dots,J-1}$ a finite sequence in $(0,1)$ and $\beta, \delta \in (0,1)$. Let $\mathbf{Y}$ be a data sample satisfying assumption (SA). Let $\phi : \mathbb{R}^K \to \mathbb{R}$ be a nonnegative, subadditive, positive-homogeneous function and $f : \left(\mathbb{R}^K\right)^n \to [0,\infty)$ be a nonnegative function on the set of data samples. Then the following holds:*

$$\mathbb{P}\left[\phi(\overline{\mathbf{Y}} - \mu) > q_{(1-\delta)\alpha_0}(\phi, \mathbf{Y} - \overline{\mathbf{Y}}) + \sum_{i=1}^{J-1}\gamma_i q_{(1-\delta)\alpha_i}(\widetilde{\phi}, \mathbf{Y} - \overline{\mathbf{Y}}) + \gamma_J f(\mathbf{Y})\right] \leq \sum_{i=0}^{J-1}\alpha_i + \mathbb{P}\left[\widetilde{\phi}(\overline{\mathbf{Y}} - \mu) > f(\mathbf{Y})\right], \quad (14)$$

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

The rationale behind this result is that the sum appearing inside the probability 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 $\alpha_i \gg \exp(-n)$.

Then comes the remaining issue of the trailing term on the right-hand-side. While it is tempting to think that it would be possible to obtain a self-contained result based on the symmetry assumption (SA) alone, we did not succeed in this direction. 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 in the previous section (or apply some other device like Bonferroni's bound (8)). The point is that this bound does not have to be particularly sharp, since we expect (in favorable cases) 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 to be almost negligible.

It seems plausible that at least a minor regularity assumption (supposedly significantly weaker than assuming a Gaussian distribution or bounded data) is actually a necessary condition in addition to (SA) to obtain a self-contained bound and ensure that nothing pathological happens with the extreme quantiles, but this remains as an interesting open issue.

As before, for computational reasons, it might be relevant to consider a block-wise Rademacher resampling scheme.

4 Simulations

For simulations we consider data of the form $Y_t = \mu_t + G_t$, where t belongs to an $m \times m$ discretized 2D torus of $K = m^2$ "pixels", identified with $\mathbb{T}_m^2 = (\mathbb{Z}/m\mathbb{Z})^2$, and G is a centered Gaussian vector obtained by 2D discrete convolution of an i.i.d. standard Gaussian field ("white noise") on $\mathbb{T}_m^2$ with a function $F : \mathbb{T}_m^2 \to \mathbb{R}$ such that $\sum_{t \in \mathbb{T}_m^2} F^2(t) = 1$. 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 $t \in \mathbb{T}_m^2$.

In the simulations below we consider for the function F a "Gaussian" convolution filter of bandwith 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

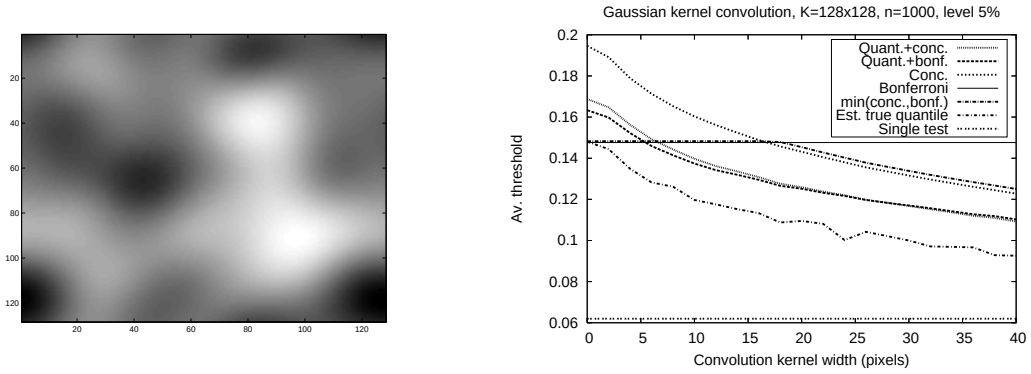

Figure 1: Left: example of a 128x128 pixel image obtained by convolution of Gaussian white noise with a (toroidal) Gaussian filter with width $b = 18$ pixels. Right: average thresholds obtained for the different approaches, see text.

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 the algorithm is the bound on the marginal variance; the form of the function F_b itself is of course unknown.

On Fig. 4 we compare the thresholds obtained when $\phi = \sup|\cdot|$, which corresponds to the two-sided multiple testing situation. 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 $\alpha = 0.05$. For the concentration threshold (6) ('conc.'), we used Rademacher weights. For the "compound" threshold of Corollary 2.2 ('min(conc.,bonf.)'), we used $\delta = 0.1$ and the Bonferroni threshold $t'_{\mathrm{Bonf},0.9\alpha}$ as the deterministic reference threshold. For the quantile approach (14), we used $J = 1$, $\alpha_0 = 0.9\alpha$, $\delta = 0.1$, and the function f is given either by the Bonferroni threshold ('quant.+bonf.') or the concentration threshold ('quant.+conc.'), both at level 0.1α. Each point represents an average over 50 experiments. Finally, we included in the figure the Bonferroni threshold $t'_{\mathrm{Bonf},\alpha}$, the threshold for a single test for comparison, and an estimation of the true quantile (actually, an empirical quantile over 1000 samples).

The quantiles or expectation with Rademacher weights were estimated by Monte-Carlo with 1000 draws. 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).

The overall conclusion of this preliminary 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.) One advantage of the concentration approach is that the 'compound' threshold (7) can "fall back" on the Bonferroni threshold when needed, at the price of a minimal threshold increase.

5 Proofs

Proof of Proposition 2.4. Denoting by $\mathbf{\Sigma}$ the common covariance matrix of the $\mathbf{Y}^i$, we have $\mathcal{L}(\overline{\mathbf{Y}}_{[W-\overline{W}]}|W) = (n^{-1}\sum_{i=1}^n (W_i - \overline{W})^2)^{1/2}\mathcal{N}(0, n^{-1}\mathbf{\Sigma})$, and the result follows because $\mathcal{L}(\overline{\mathbf{Y}}-\mu) = \mathcal{N}(0, n^{-1}\mathbf{\Sigma})$ and ϕ is positive-homogeneous. □ □

Proof of Proposition 2.6. (i). By independence between W and $\mathbf{Y}$, using the positive homogeneity, then convexity of ϕ, for every realization of $\mathbf{Y}$ we have:

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

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

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

We obtain (ii) via the triangle inequality and the same symmetrization trick. □ □

Proof of Proposition 2.8. We denote by $\mathbf{A}$ a square root of the common covariance matrix of the $\mathbf{Y}^i$ and by $(a_k)_{1\leq k\leq K}$ the rows of $\mathbf{A}$. If $\mathbf{G}$ is a $K \times m$

matrix with standard centered i.i.d. Gaussian entries, then $\mathbf{AG}$ has the same distribution as $\mathbf{Y}-\mu$. We let for all $\zeta \in \left(\mathbb{R}^K\right)^n$, $T_1(\zeta) := \phi\left(\frac{1}{n}\sum_{i=1}^n \mathbf{A}\zeta_i\right)$ and $T_2(\zeta) := \mathbb{E}\phi\left(\frac{1}{n}\sum_{i=1}^n (W_i - \overline{W})\mathbf{A}\zeta_i\right)$. From the Gaussian concentration theorem of Cirel'son, Ibragimov and Sudakov (see for example [Mas05], Theorem 3.8), we just need to prove that T_1 (resp. T_2) is a Lipschitz function with constant $\|\sigma\|_p/\sqrt{n}$ (resp. $\|\sigma\|_p C_W/n$), for the Euclidean norm $\|\cdot\|_{2,Kn}$ on $\left(\mathbb{R}^K\right)^n$. Let $\zeta, \zeta' \in \left(\mathbb{R}^K\right)^n$. Using Cauchy-Schwartz's inequality coordinate-wise and $\|a_k\|_2 = \sigma_k$, we deduce

$$|T_1(\zeta) - T_1(\zeta')| \leq \left\| \frac{1}{n}\sum_{i=1}^n \mathbf{A}\left(\zeta_i - \zeta_i'\right) \right\|_p \leq \|\sigma\|_p \left\| \frac{1}{n}\sum_{i=1}^n \left(\zeta_i - \zeta_i'\right) \right\|_2 .$$

Therefore, we get $|T_1(\zeta) - T_1(\zeta')| \leq \frac{\|\sigma\|_p}{\sqrt{n}} \|\zeta - \zeta'\|_{2,Kn}$ by convexity of $x \in \mathbb{R}^K \to \|x\|_2^2$, and we obtain (i). For T_2, we use the same method as for T_1 :

$$\begin{aligned} |T_2(\zeta) - T_2(\zeta')| &\leq \|\sigma\|_p \, \mathbb{E}\left\| \frac{1}{n}\sum_{i=1}^n (W_i - \overline{W})(\zeta_i - \zeta_i') \right\|_2 \\ &\leq \frac{\|\sigma\|_p}{n} \sqrt{\mathbb{E}\left\| \sum_{i=1}^n (W_i - \overline{W})(\zeta_i - \zeta_i') \right\|_2^2} . \end{aligned} \tag{15}$$

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

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

Consequently,

$$\mathbb{E}\left\| \sum_{i=1}^n \left(W_i - \overline{W}\right)\left(\zeta_i - \zeta_i'\right) \right\|_2^2 \leq C_W^2 \sum_{i=1}^n \|\zeta_i - \zeta_i'\|_2^2 \leq C_W^2 \|\zeta - \zeta'\|_{2,Kn}^2 . \tag{16}$$

Combining expression (15) and (16), we find that T_2 is $\|\sigma\|_p C_W/n$-Lipschitz. □ □

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 [Fro04]). The (GA) case is a straightforward consequence of Proposition 2.4 and the proof of Proposition 2.8. □ □

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

Proof of Proposition 3.1. Remember the following inequality coming from the definition of the quantile q_α: for any fixed $\mathbf{Y}$

$$\mathbb{P}_W\left[\phi\left(\overline{\mathbf{Y}}_{[W]}\right) > q_\alpha(\phi, \mathbf{Y})\right] \leq \alpha \leq \mathbb{P}_W\left[\phi\left(\overline{\mathbf{Y}}_{[W]}\right) \geq q_\alpha(\phi, \mathbf{Y})\right], \tag{17}$$

which will be useful in this proof. We have

$$\begin{aligned}
\mathbb{P}_{\mathbf{Y}}\left[\phi(\overline{\mathbf{Y}}-\mu) > q_\alpha(\phi, \mathbf{Y}-\mu)\right] &= \mathbb{E}_W\left[\mathbb{P}_{\mathbf{Y}}\left[\phi\big(\overline{(\mathbf{Y}-\mu)}_{[W]}\big) > q_\alpha(\phi, (\mathbf{Y}-\mu)_{[W]})\right]\right] \\
&= \mathbb{E}_{\mathbf{Y}}\left[\mathbb{P}_W\left[\phi\left(\overline{(\mathbf{Y}-\mu)}_{[W]}\right) > q_\alpha(\phi, \mathbf{Y}-\mu)\right]\right] \\
&\leq \alpha\,. \qquad (18)
\end{aligned}$$

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 $q_\alpha(\phi, (\mathbf{Y}-\mu)_{[W]}) = q_\alpha(\phi, \mathbf{Y}-\mu)$; finally the last inequality comes from (17). Let us define the event

$$\Omega = \left\{\mathbf{Y} \text{ s.t. } q_\alpha(\phi, \mathbf{Y}-\mu) \leq q_{\alpha(1-\delta)}(\phi, \mathbf{Y}-\overline{\mathbf{Y}}) + f(\mathbf{Y})\right\};$$

then we have using (18) :

$$\begin{aligned}
\mathbb{P}\left[\phi(\overline{\mathbf{Y}}-\mu) > q_{\alpha(1-\delta)}(\phi, \mathbf{Y}-\overline{\mathbf{Y}}) + f(\mathbf{Y})\right] &\leq \mathbb{P}\left[\phi(\overline{\mathbf{Y}}-\mu) > q_\alpha(\phi, \mathbf{Y}-\mu)\right] + \mathbb{P}\left[\mathbf{Y} \in \Omega^c\right] \\
&\leq \alpha + \mathbb{P}\left[\mathbf{Y} \in \Omega^c\right]\,. \qquad (19)
\end{aligned}$$

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

$\mathbf{Y} \in \Omega^c$:

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

For the first and last inequalities we have used (17), and for the second inequality the definition of Ω^c. From this we deduce that

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

Now using the homogeneity of ϕ, and the fact that both ϕ and f are non-negative:

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

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

$$\Omega^c \subset \left\{ \mathbf{Y} \text{ s.t. } \widetilde{\phi}(\overline{\mathbf{Y}}-\mu) > \frac{n}{2\overline{\mathcal{B}}\left(n, \frac{\alpha\delta}{2}\right) - n} f(\mathbf{Y}) \right\} ,$$

which, put back in (19), leads to the desired conclusion. □ □

Acknowledgements

We want to thank Pascal Massart for his particulary relevant suggestions.

References

[Efr79] B. Efron. Bootstrap methods: another look at the jackknife. *Ann. Statist.*, 7(1):1–26, 1979.

[Fro04] Magalie Fromont. Model selection by bootstrap penalization for classification. In *Learning theory*, volume 3120 of *Lecture Notes in Comput. Sci.*, pages 285–299. Springer, Berlin, 2004.

[Hal92] Peter Hall. *The bootstrap and Edgeworth expansion*. Springer Series in Statistics. Springer-Verlag, New York, 1992.

[Mas05] Pascal Massart. Concentration inequalities and model selection (lecture notes of the St-Flour probability summer school 2003). Available online at `http://www.math.u-psud.fr/~massart/stf2003_massart.pdf`, 2005.

[VdVW96] Aad W. Van der Vaart and Jon A. Wellner. *Weak convergence and empirical processes*. Springer Series in Statistics. Springer-Verlag, New York, 1996. With applications to statistics.