The Annals of Statistics
2021, Vol. 49, No. 1, 272–294
https://doi.org/10.1214/20-AOS1956

ESTIMATING MINIMUM EFFECT WITH OUTLIER SELECTION

BY ALEXANDRA CARPENTIER[1], SYLVAIN DELATTRE[2], ETIENNE ROQUAIN[3] AND NICOLAS VERZELEN[4]

[1]*Otto-von-Guericke-Universität Magdeburg, Fakultät für Mathematik (FMA), Institut für Mathematische Stochastik (IMST), alexandra.carpentier@ovgu.de*

[2]*Laboratoire de Probabilités Statistique et Modélisation, Université de Paris, CNRS, sylvain.delattre@univ-paris-diderot.fr*

[3]*Laboratoire de Probabilités Statistique et Modélisation, Sorbonne Université, CNRS, etienne.roquain@upmc.fr*

[4]*INRAE, SUPAGRO, Univ. Montpellier, UMR 729, MISTEA, nicolas.verzelen@inrae.fr*

We introduce one-sided versions of Huber's contamination model, in which corrupted samples tend to take larger values than uncorrupted ones. Two intertwined problems are addressed: estimation of the mean of the uncorrupted samples (minimum effect) and selection of the corrupted samples (outliers). Regarding estimation of the minimum effect, we derive the minimax risks and introduce estimators that are adaptive with respect to the unknown number of contaminations. The optimal convergence rates differ from the ones in the classical Huber contamination model. This fact uncovers the effect of the one-sided structural assumption of the contaminations. As for the problem of selecting the outliers, we formulate the problem in a multiple testing framework for which the location and scaling of the null hypotheses are unknown. We rigorously prove that estimating the null hypothesis while maintaining a theoretical guarantee on the amount of the falsely selected outliers is possible, both through false discovery rate (FDR) and through post hoc bounds. As a by-product, we address a long-standing open issue on FDR control under equi-correlation, which reinforces the interest of removing dependency in such a setting.

1. Introduction. We are considering statistical problems for which some of the available data have been corrupted. Such issues have been addressed by different fields in statistics, depending on how one defines and considers the corruption. Two examples are robust estimation and sparse modeling. In the former, Huber's contamination model [41, 42] is the prototypical setting for handling this problem. It assumes that among n observations $Y_1, \dots, Y_n$, most of them follow some normal distribution $\mathcal{N}(\theta, \sigma^2)$ whereas the corrupted data are arbitrarily distributed. In sparse modeling, one typically assumes that the data $Y_1, \dots, Y_n$ are normally distributed with mean γ_i, where $\gamma_i = \theta$ for uncorrupted samples and $\gamma_i \neq \theta$ is arbitrary for corrupted samples (see [8] for a related model).

However, in some practical problems, corrupted samples do not take arbitrary values and instead satisfy a structural assumption. Consider for instance the situation where the Y_i's are concentration measurements of a pollutant, coming from n sensors spread out at n locations of a city. This pollutant has a background concentration value θ in the city, but due to local pollution effects, the sensors may record larger values at some locations. Health authorities are then interested in evaluating the degree of background pollution and in finding the most affected regions in the city.

In this work, we introduce one-sided contamination models to account for the structural assumption that corrupted samples tend to take larger values than uncorrupted ones. Then we

Received September 2018; revised January 2020.
MSC2020 subject classifications. Primary 62G10; secondary 62C20.
Key words and phrases. Contamination, equicorrelation, false discovery rate, Hermite polynomials, minimax rate, moment matching, multiple testing, post hoc, selective inference, sparsity.

consider the twin problems of estimating the distribution of the uncorrupted samples and of identifying the corrupted samples.

1.1. *Models and objectives.*

1.1.1. *One-Sided Contamination model* (*OSC*). We first introduce a one-sided counterpart of Huber's contamination model for which some of the Y_i's follow a $\mathcal{N}(\theta, \sigma^2)$ distribution, whereas the remaining samples are positively contaminated, that is, they have a distribution that stochastically dominates a $\mathcal{N}(\theta, \sigma^2)$ distribution, but is otherwise arbitrary.

More formally, we assume that

$$Y_i = \theta + \sigma \varepsilon_i, \quad 1 \le i \le n, \tag{1}$$

where $\sigma > 0$ is a standard deviation parameter (either equal to 1 or unknown), $\theta \in \mathbb{R}$ is a fixed *minimum effect* and the ε_i's are independent noise random variables. Denoting by π_i the unknown distribution of the noise, we assume that, for some k, the distribution $\pi = \bigotimes_{i=1}^n \pi_i$ of ε belongs to the set

$$\overline{\mathcal{M}}_k = \left\{ \pi = \bigotimes_{i=1}^n \pi_i : \pi_i \succeq \mathcal{N}(0,1),\ 1 \le i \le n,\ \sum_{i=1}^n \mathbb{1}_{\{\pi_i \succ \mathcal{N}(0,1)\}} \le k \right\}, \tag{2}$$

where $\succeq$ (resp., $\succ$) denotes the stochastic domination (resp., strict stochastic domination); see Section 1.4 below for a definition. In $\overline{\mathcal{M}}_k$, at most k of the distributions π_i are allowed to strictly dominate the Gaussian measure. Model (1) satisfies the heuristic explanation described above: if $\pi \in \overline{\mathcal{M}}_k$, then at least $n - k$ samples are noncontaminated and follow a $\mathcal{N}(\theta, \sigma^2)$ distribution, whereas the remaining contaminated samples stochastically dominate this distribution.

In this model, henceforth referred to as the One-Sided Contamination (OSC) model, the parameter θ corresponds to the expectation of the non-contaminated samples. If $k \le n - 1$, it also satisfies

$$\theta = \min_{1 \le i \le n} \mathbb{E}(Y_i), \tag{3}$$

and can therefore be interpreted as a minimum theoretical effect. In particular, θ is identifiable for $k \in [n/2, n-1]$, whereas this is not the case in the classical Huber's model.

Throughout the paper, probabilities (resp., expectations) in model (1) are denoted by $\mathbb{P}_{\theta,\pi,\sigma}$ (resp., $\mathbb{E}_{\theta,\pi,\sigma}$). The parameter σ is dropped in the notation whenever $\sigma = 1$.

1.1.2. *One-Sided Gaussian Contamination model* (*gOSC*). As in the sparse Gaussian vector model, we also consider a specific case of the OSC model for which the contaminated samples are still assumed to be normally distributed, that is, the π_i's are Gaussian distributions with unit variance and positive mean μ_i/σ where $\mu \in \mathbb{R}_+^n$ is a *contamination effect*. In that case, the model can be rewritten as

$$Y_i = \theta + \mu_i + \sigma \xi_i, \quad 1 \le i \le n, \tag{4}$$

where the ξ_i's are i.i.d. $\mathcal{N}(0,1)$ distributed and $\mu \in \mathbb{R}_+^n$ is unknown. Defining the mean vector

$$\gamma = (\theta + \mu_i)_{1 \le i \le n}, \tag{5}$$

we deduce that Y follows a normal distribution with unknown mean γ and variance $\sigma^2 I_n$, while θ corresponds to $\min_{1 \le i \le n}\{\gamma_i\}$, that is, the minimum component of the mean vector.

To formalize the connection with the OSC model, let $\varepsilon_i = \mu_i/\sigma + \xi_i$ and $\pi_i = \mathcal{N}(\mu_i/\sigma, 1)$ for all $i \in \{1, \dots, n\}$. Then (4) is a particular case of (1) because $\mathcal{N}(\mu_i/\sigma, 1) \succeq \mathcal{N}(0,1)$. As

in the OSC model, we prescribe the number of contaminated samples to be less or equal to k, by defining

$$\mathcal{M}_k = \left\{ \mu \in \mathbb{R}_+^n : \sum_{i=1}^n \mathbb{1}_{\{\mu_i \neq 0\}} \leq k \right\}. \tag{6}$$

In what follows, we refer to model (4) as the One-Sided Gaussian Contamination (gOSC) model. The probabilities (resp., expectations) in this model are denoted by $\mathbb{P}_{\theta,\mu,\sigma}$ (resp., $\mathbb{E}_{\theta,\mu,\sigma}$). Whenever we assume that the variance parameter σ is known and is equal to 1, the subscript σ is dropped in the above notation.

1.1.3. *Objectives*. We are interested in the two following intertwined problems:

– *Objective 1: optimal estimation of the minimum effect.* We aim at establishing the minimax estimation rates of θ, both in the OSC (1) and in the gOSC (4) models. In particular, we explore the role of the one-sided contamination assumption for the computation of such estimation rates. As explained below, this problem is at the crossroads of several lines of research, including robust estimation and nonsmooth linear functional estimation.

– *Objective 2: controlled selection of the outliers.* Here, we are interested in finding the contaminated samples. In the Gaussian case (gOSC), this is equivalent to selecting the positive entries of μ in (5). Adopting a multiple testing framework, we aim at building a selection procedure with suitable false discovery rate (FDR) control [4] and providing a valid post hoc bound [36, 37]. The difficulty stems from the fact the minimum effect θ is unknown. In contrast to objective 1 where the contaminated samples were considered as nuisance quantities, for this second objective the contaminated samples are now interpreted as the signal whereas θ is a nuisance parameter.

Objective 2 is intrinsically connected to the problem of removing the correlation when making (one-sided) multiple tests from Gaussian equicorrelated test statistics: when the equicorrelation is carried by the latent factor θ, we can remove this correlation by subtracting an estimator of θ from the test statistics. Although this simple strategy is quite common (see, e.g., [33] and references therein), assessing whether the theoretical performance of the resulting procedure remains suitable is a longstanding issue in multiple testing literature. In this work, we provide positive results for this problem, by showing that it is possible to (asymptotically) control the FDR while having (at least) the same power as if the test statistics had been independent.

In the remainder of the Introduction, we first describe our contribution for minimum effect estimation and then turn to outlier selection.

1.2. *Optimal estimation of the minimum effect*. Given $\sigma^2 = 1$ and a sparsity parameter $k \in \{1, \dots, n-1\}$, we define the L_1 minimax estimation risk of θ in both gOSC (4) and OSC (1) models:

$$\mathcal{R}[k,n] = \inf_{\widehat{\theta}} \sup_{(\theta,\mu)\in\mathbb{R}\times\mathcal{M}_k} \mathbb{E}_{\theta,\mu}\big[|\widehat{\theta} - \theta|\big]; \qquad \overline{\mathcal{R}}[k,n] = \inf_{\widehat{\theta}} \sup_{\theta\in\mathbb{R},\pi\in\overline{\mathcal{M}}_k} \mathbb{E}_{\theta,\pi}\big[|\widehat{\theta} - \theta|\big]. \tag{7}$$

We characterize these minimax risks by deriving matching (up to numerical constants) lower and upper bounds, uniformly over the parameter k; see Sections 2 and 3. The results are summarized in Table 1. It is interesting to compare these orders of magnitude with those derived for the Huber contamination model with at most k contaminated samples. From, for example, Section 2 of [14], the minimax risk is of order $\max(n^{-1/2}, \frac{k}{n})$ in Huber's model (the results in [14] are proved for a model where the number of contaminated samples follows a Binomial distribution with parameters $(n, k/n)$, but the proofs straightforwardly extend to

TABLE 1
Minimax estimation risks of θ (up to numerical constants)

	General bound	$1 \le k \le \sqrt{n}$	$\sqrt{n} \le k \le n/2$	$n/2 \le k \le n-1$
$\overline{\mathcal{R}}[k,n]$	$\frac{\log(\frac{n}{n-k})}{\log^{1/2}(1+\frac{k^2}{n})}$	$n^{-1/2}$	$\frac{k/n}{\log^{1/2}(k^2/n)}$	$\frac{\log(\frac{n}{n-k})}{\log^{1/2} n}$
$\mathcal{R}[k,n]$	$\frac{\log^2(1+\sqrt{\frac{k}{n-k}})}{\log^{3/2}(1+(\frac{k}{\sqrt{n}})^{2/3})}$	$n^{-1/2}$	$\frac{k/n}{\log^{3/2}(k^2/n)}$	$\frac{\log^2(\frac{n}{n-k})}{\log^{3/2} n}$
Huber	$\max(n^{-1/2}, \frac{k}{n})$	$n^{-1/2}$	k/n	∞

the case of a fixed parameter k). For $k \le \sqrt{n}$, the rate is parametric in all three models. For $k \in (\sqrt{n}, n/2)$, the assumption of one-sided contamination leads to a $\log^{1/2}(k^2/n)$ gain over the Huber model, while assuming that the contaminations are Gaussian lead to an additional logarithmic gain. For $k \in [n/2, n-1]$, recall that Huber's model is not identifiable while the one-sided contamination model is and we identify various minimax rates. For a fixed proportion (k/n) of contaminated samples, the optimal rate converges to 0 at a polylogarithmic rate. For a proportion $(n-k)/n$ of noncontaminated samples slowly decaying with n, the estimation rate still goes to 0.

For both models (OSC and gOSC), we also provide estimation procedures that are adaptive to the unknown number k of contaminated samples. Finally, in Section 4, we consider the case where the noise level σ in (4) is unknown. We prove, in the OSC model, that adaptation to unknown σ is possible, and we characterize the optimal estimation risk for σ.

OSC: *Technical aspects and connection to robust estimation*. As explained earlier, the OSC model (1) is a one-sided counterpart of Huber's contamination model [41, 42]; see also [58] for the historical reference on the concept of contamination and [50, 53] for more recent reviews. From a technical perspective, minimax bounds for the OSC model proceed from the same general ideas as for Huber's contamination model, but with a twist. In Huber's model, the empirical median turns out to be optimal [42], while in the OSC model, there is a benefit to using other empirical quantiles. Since the contaminations are one-sided, the left tail is less perturbed than the right tail. Correcting for the bias and choosing a suitable quantile, we prove that the resulting estimator achieves (up to a constant) the optimal rate $\overline{\mathcal{R}}[k,n]$. Adaptation to unknown k is performed via Lepski's method, while adaptation to unknown σ is based on a difference of empirical quantiles.

gOSC: *Technical aspects and connection to nonsmooth functional estimation*. Pinpointing the minimax risk in the Gaussian contamination model (gOSC) is much more technical. Indeed, standard estimators, such as those based on quantiles, are not optimal in this setting. The key idea of our upper bound is to invert a collection of local tests of the form "$\theta \ge u$" versus "$\theta < u$" for $u \in \mathbb{R}$, following an approach from [13] developed for sparsity testing. Recall that γ_i in (5) stands for the expectation of Y_i; then under the null "$\theta \ge u$", we have $\sum_{i=1}^n \mathbb{1}_{\gamma_i < u} = 0$, whereas under the alternative "$\theta < u$" we have $\sum_{i=1}^n \mathbb{1}_{\gamma_i < u} \ge n-k$. Thus, this approach boils down to estimating the nonsmooth-functional $u \mapsto \sum_{i=1}^n \mathbb{1}_{\gamma_i < u}$.

Starting from the seminal works [23, 43] (for respectively linear and quadratic functionals), an extensive literature on estimating smooth functionals of the mean of a Gaussian vector has been developed. Under a sparsity assumption, this problem has been investigated in [9, 15, 17, 66], and shares some strong connections with the problem of signal detection, as studied in, for example, [2, 45].

However, estimation of nonsmooth functionals (e.g., $\sum_{i=1}^n |\gamma_i|^q$ with $q \in (0,1]$) is significantly more involved, even without sparsity assumptions; see, for example, [8, 10, 13, 16, 40, 46–49, 55, 68]. For this class of problems, one powerful approach, known as the polynomial approximation method [40, 55], amounts to building a suitable polynomial approximation of the nonsmooth function and plugging in unbiased estimators of the moments $\sum_{i=1}^n \gamma_i^s$ for some integers $s \in \{1, \dots, s_{\max}\}$. Unfortunately, we cannot rely on this strategy for estimating $\sum_{i=1}^n \mathbb{1}_{\gamma_i < u}$, mainly because the contaminated γ_i's may be arbitrarily large. In a related setting, where the contaminated means $\gamma_i \neq \theta$ are distributed according to some smooth prior distributions supported on $\mathbb{R}$, Cai and Jin [8] have pinpointed the optimal rate by using the empirical Fourier transform (see also [22]). However, this approach also falls down in our framework, again because the contaminated γ_i's are arbitrary. In this work, we introduce a new strategy that combines polynomial approximation methods with the empirical Laplace transform. Indeed, the empirical Laplace transform has the virtue of being almost insensitive to arbitrarily large values of γ_i.

As for the minimax lower bound, we rely on moment matching techniques, following the approach of [55] that has been recently applied to other non-smooth functional models [10, 13, 40, 68].

1.3. *Controlled selection of the outliers.* Turning to the second objective, we now detail our contributions and discuss the relevant literature. Our approach falls with the multiple testing paradigm and builds upon some of our estimators for θ (and for σ).

1.3.1. *Multiple testing formulation.* Recall that our second objective was to identify the active set of outliers in the general model (1). Again, we emphasize that what we described before as outliers is in this part the quantities of interest (e.g., the city locations with abnormal pollutant concentration in our motivating example). In the OSC model, we formulate this selection problem as n simultaneous tests of

$$H_{0,i} : \text{“}\pi_i = \mathcal{N}(0,1)\text{”} \quad \text{against} \quad H_{1,i} : \text{“}\pi_i \succ \mathcal{N}(0,1)\text{”}, \quad 1 \le i \le n. \tag{8}$$

(Recall that “$\succ$” stands for strict stochastic domination.)

In the specific case of the gOSC model (4), this problem amounts to simultaneously testing

$$H_{0,i} : \text{“}\mu_i = 0\text{”} \quad \text{against} \quad H_{1,i} : \text{“}\mu_i > 0\text{”}, \quad 1 \le i \le n.$$

We denote the set of nonoutlier coordinates by $\mathcal{H}_0(\pi) = \{1 \le i \le n : \pi_i = \mathcal{N}(0,1)\}$, and the set of outlier coordinates by $\mathcal{H}_1(\pi) = \{1 \le i \le n : \pi_i \succ \mathcal{N}(0,1)\}$.

The cardinality of $\mathcal{H}_0(\pi)$ (resp., $\mathcal{H}_1(\pi)$) is denoted by $n_0(\pi)$ (resp., $n_1(\pi)$). Hence, $\pi \in \overline{\mathcal{M}_k}$, means that the number of outliers is $n_1(\pi) \le k$. Our selection problem thus amounts to estimating $\mathcal{H}_1(\pi)$ (or equivalently $\mathcal{H}_0(\pi)$). The dependence on π of $\mathcal{H}_0(\pi)$, $\mathcal{H}_1(\pi)$, $n_0(\pi)$, $n_1(\pi)$ is sometimes suppressed for convenience.

A multiple-testing procedure is a data-driven set $R \subset \{1, \dots, n\}$ of proposed outliers. For a given R, we classically quantify the amount of false positives of R by its false discovery proportion [4],

$$\mathrm{FDP}(\pi, R) = \frac{|R \cap \mathcal{H}_0(\pi)|}{|R| \vee 1}, \tag{9}$$

which records the proportion of errors among the set R of selected outliers. The expectation of this quantity $\mathbb{E}_{\theta,\pi,\sigma}[\mathrm{FDP}(\pi, R)]$ is the false discovery rate, which can be considered as the standard generalization of the single testing type I error rate to large scale multiple testing. The true discovery proportion is then defined by

$$\mathrm{TDP}(\pi, R) = \frac{|R \cap \mathcal{H}_1(\pi)|}{n_1(\pi) \vee 1}, \tag{10}$$

and corresponds to the proportion of (correctly) selected outliers among the set of true outliers. The expectation of this quantity $\mathbb{E}_{\theta,\pi,\sigma}[\mathrm{TDP}(\pi, R)]$ is a widely used analogue of the power in single testing; see, for example, [1, 59, 63]. Our contribution falls into two frameworks:

- *Multiple testing*: find a procedure selecting a subset $R \subset \{1, \dots, n\}$ as close as possible to $\mathcal{H}_1(\pi)$, that is, that has a TDP as high as possible while maintaining a controlled FDR.
- *Post hoc bound*: provide a confidence bound on $\mathrm{FDP}(\pi, S)$, uniformly valid over all possible selection $S \subset \{1, \dots, n\}$.

While the first objective is classical in the multiple testing field (see, e.g., [4, 5, 32, 34]), the second objective was proposed more recently in [35–37]. It is connected to the burgeoning research field of selective inference; see, for example, [6] and references therein. The rationale behind developing such a bound is that, since the control is uniform, the probability coverage is guaranteed even if S is itself data-dependent. The obtained bound is therefore also valid when practitioners reuse the same dataset, possibly several times, in order to design S. We denote the outlier selected set either by R or S depending on the considered setting: R is typically a procedure designed by the statistician, whereas S is chosen by the user.

1.3.2. *Relation to the first objective and to previous literature.* In the OSC model (1), solving the above multiple testing issues is challenging primarily because the parameters θ and σ are unknown. Indeed, this entails that the scaling of the null distribution (i.e., the distribution under the null hypothesis) is unknown. A natural idea is to design a two-stage procedure: first, we estimate θ and σ by some estimators $\widehat{\theta}$ and $\widehat{\sigma}$ (as we do in the first part of this paper). Then, in a testing stage, we apply a standard multiple testing procedure to the rescaled observation $Y_i' = (Y_i - \widehat{\theta})/\widehat{\sigma}$.

Estimating the null distribution in a multiple testing context has been popularized in a series of works by Efron; see [25, 27, 28]. Through careful data analyses, Efron noticed that the theoretical null distribution often turns out to be wrong in practical situations, which can lead to an uncontrolled increase in false positives. To address this issue, Efron recommends estimating the scaling parameters of the null distribution (θ, σ here) by "central matching," that is, by fitting a parametric curve to the trimmed data. In his work, Efron provides compelling empirical evidence for his approach. However, to our knowledge, the FDP and TDP of such two-stage testing procedures have never been theoretically controlled. Note that estimating the null in a multiple testing context was also the motivation of the minimax results of [8, 48], although the corresponding multiple testing procedure was not studied. We recall that these previous studies are all developed in the two-sided context, whereas our focus is on a one-sided shape constraint.

Finally, let us mention that a procedure learning part of the null distribution while maintaining FDR control and a form of power optimality has been proposed in [1, 3], often referred to as the Barber–Candès procedure. In the one-sided case, it is able to learn the unknown null density when this is assumed to be symmetric around 0. However, this assumption is not satisfied in our context because the symmetry point of the null distribution (θ here) is unknown.

1.3.3. *Summary of our results.* In Sections 5 and S-1, we show that a fair modification of the quantile-based estimators $\widehat{\theta}, \widehat{\sigma}$, introduced for the OSC model, can be used to estimate the null distribution to rescale the p-value process, and can then be suitably combined with classical multiple testing procedures:

1. A new $(\widehat{\theta}, \widehat{\sigma})$-rescaled Benjamini–Hochberg procedure R is defined that satisfies the following FDR controlling property: in the general model (1), for any $\pi \in \overline{\mathcal{M}}_k$,

with $k = \lfloor 0.9n \rfloor$,

$$\left(\mathbb{E}_{\theta,\pi,\sigma}\big(\mathrm{FDP}(\pi, R)\big) - \frac{n_0}{n}\alpha\right)_+ \lesssim \log(n)/n^{1/16}.$$

In addition, we derive a power result showing that the power (the expectation of the TDP) of this procedure is close to that of the (θ, σ)-rescaled Benjamini–Hochberg procedure, under mild conditions. The latter is an oracle benchmark that would require the exact knowledge of θ and σ.

2. A new $(\widehat{\theta}, \widehat{\sigma})$-rescaled post hoc bound $\overline{\mathrm{FDP}}(\cdot)$ is proposed, satisfying, for any $\pi \in \overline{\mathcal{M}}_k$, with $k = \lfloor 0.9n \rfloor$,

$$\big(1 - \alpha - \mathbb{P}_{\theta,\pi,\sigma}\big(\forall S \subset \{1, \dots, n\}, \mathrm{FDP}(\pi, S) \le \overline{\mathrm{FDP}}(S)\big)\big)_+ \lesssim \log(n)/n^{1/16}.$$

To the best of our knowledge, these are the first results that theoretically validate Efron's principle of empirical null correction in a specific multiple testing context.

For bounding the type I error rates, the technical argument used in our proof is close in spirit to recent studies [44, 57] (among others): the idea is to divide the data into two "orthogonal" parts (small or large Y_i's), the first part being used for the rescaling and the second one for testing. For the power result, our formal argument is entirely new to our knowledge; see also Remark 5.3.

Finally, let us mention that the above rate $\log(n)/n^{1/16}$ is certainly not optimal and could be improved by using more involved estimators $\widehat{\theta}$, $\widehat{\sigma}$. Indeed, our concern here regarding the multiple testing procedure is to show that it is consistent with respect to the FDP/TDP metrics, this for a wide range of values of k, including the dense case where k is of the order of n. Furthermore, despite the fact that this rate is relatively slow, the new rescaled BH procedure improves on the vanilla BH procedure, as illustrated in the numerical experiments of Section S-1.3 (see also Figure 1).

1.3.4. *Application to decorrelation in multiple testing.* It is well known that Efron's methodology on empirical null correction can be applied to reduce the effect of correlations between the tests, as noted by Efron himself [26, 29] where he mentioned that "there is a lot at stake here." Several following works supported this assertion, especially by decomposing the covariance matrix of the data into factors; see [30, 31, 33, 54]. However, strong theoretical results on such corrected multiple testing procedures are still not available.

Meanwhile, another branch of the literature aims at incorporating known and unknown dependence into multiple testing procedures, for instance, by resampling-based approaches [3,

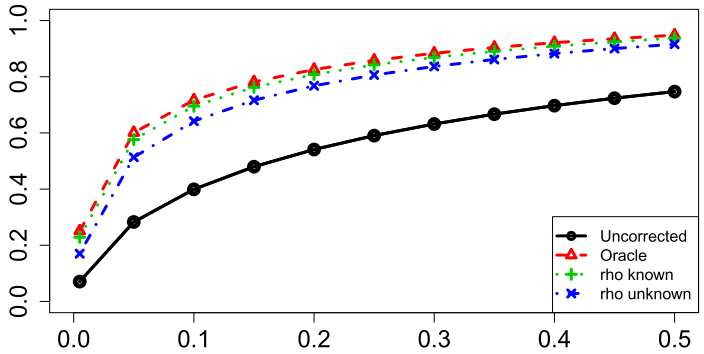

FIG. 1. *X-axis*: *targeted FDR level* $\alpha \in \{0.005, 0.05, 0.1, 0.15, 0.2, 0.25, 0.3, 0.35, 0.4, 0.45, 0.5\}$, *Y-axis*: *TDP* (*power*) *averaged over* 100 *replications for four different procedures* (*see text in Section S*-1.3). *The model is the one-sided Gaussian one with equicorrelation* $\rho = 0.3$. *The parameters used are* $n = 10^6$, $\Delta = 2.5$, $k/n = 0.1$.

24, 60–62, 67] or by directly plugging the known dependence structure [7, 19, 39]. However, as noted for instance in the discussion of [64], even for very simple correlation structures, no multiple testing procedure has yet been proved to control the FDR while having an optimal expected TDP.

In Section S-1.2, we apply our two-step procedure to address the multiple testing problem in the contaminated one-factor model. For instance, this model encompasses the one-sided Gaussian equi-correlation case (with nonnegative equicorrelation ρ) which is often used as a concrete test bed in multiple testing literature; see, for example, [18, 20, 51] among others. It turns out that the contaminated one-factor model can be written in the form of the OSC model (1) with a random value of θ and an unknown variance $\sigma^2 = 1 - \rho$. Hence, we can directly apply our $(\widehat{\theta}, \widehat{\sigma})$-rescaled Benjamini–Hochberg procedure introduced above to solve the problem: we show that the new procedure has performances close to the BH procedure under independence (and even with a slight increase of the signal to noise ratio). Even if the model is somewhat specific, this shows that correcting the dependence can be fully theoretically justified. To illustrate numerically the benefit of such an approach, Figure 1 displays a ROC-type curve for four different versions of corrected BH procedure (in the Gaussian equicorrelated case). A full description of the simulation setting and additional experiments are provided in Section S-1.3.

1.4. *Notation.* For $x > 0$, we write $\lfloor x \rfloor^{(\log_2)}$ (resp., $\lceil x \rceil^{(\log_2)}$) for $2^{\lfloor \log_2(x) \rfloor}$ (resp., $2^{\lceil \log_2(x) \rceil}$), that is, the largest (resp., smallest) dyadic number no larger (resp., no smaller) than x. Similarly, $\lfloor x \rfloor_{\text{even}}$ is the largest even integer which is not larger than x.

For $x \in \mathbb{R}^n$, we denote by $x_{(k)}$ the kth smallest element of $\{x_i, 1 \le i \le n\}$. We also write $x_{(\ell:m)}$ for the ℓth smallest element among $\{x_i, 1 \le i \le m\}$, for some integer $1 \le m \le n$.

In what follows, c, c' denote numerical positive constants whose values may vary from line to line. For two sequences $(u_t)_{t \in \mathcal{T}}$ and $(v_t)_{t \in \mathcal{T}}$, we write that for all $t \in \mathcal{T}$, $u_t \lesssim v_t$ (resp., for all $t \in \mathcal{T}$, $u_t \gtrsim v_t$), if there exists a universal constant $c > 0$ such that for all $t \in \mathcal{T}$, $u_t \le c v_t$ (resp., for all $t \in \mathcal{T}$, $u_t \ge c v_t$). We write $u_t \asymp v_t$ if $u_t \lesssim v_t$ and $v_t \lesssim u_t$.

For X, Y two real random variables with respective cumulative distribution functions F_X, F_Y, we write $X \succeq Y$ if for all $x \in \mathbb{R}$, we have $F_X(x) \le F_Y(x)$. We write $X \succ Y$ if $X \succeq Y$ and if there exists $x \in \mathbb{R}$, such that $F_X(x) < F_Y(x)$. We also denote $P \succeq Q$ (resp., $P \succ Q$) whenever $X \succeq Y$ (resp., $X \succ Y$) for $X \sim P$ and $Y \sim Q$.

We write Φ for the cumulative distribution function of the standard normal distribution, we write $\bar{\Phi} = 1 - \Phi$ and ϕ for its usual density.

For space reasons, part of the outlier selection results, the numerical experiments and all the proofs of the paper are deferred to a Supplementary Material [11]. For clarity, the sections, equations and results of this supplement are referred to with an additional symbol "S-" in the numbering.

2. Estimation of θ in the gOSC model with known variance.

In this section, we consider the problem of estimating θ in the Gaussian contamination model (4) and investigate the L_1 minimax risk defined in (7). We assume throughout this section that $\sigma^2 = 1$. The case of unknown variance is discussed in Section 6.

2.1. *Lower bound for the gOSC model with known variance.*

THEOREM 2.1. *There exists a universal constant $c > 0$ such that for any positive integer n and for any integer $k \in [1, n-1]$,*

$$\mathcal{R}[k, n] \ge c \frac{\log^2(1 + (\frac{k}{n-k})^{1/2})}{\log^{3/2}(1 + (\frac{k^2}{n})^{1/3})}. \tag{11}$$

The proof of this theorem is given in Section S-2.1. The main tool for proving this lower bound is moment matching: we build two priors on the parameter γ that correspond to distant values of θ while they have the same $\log n$ (or so) first moments. This is done in an implicit way using the Hahn–Banach theorem together with properties of Chebychev polynomials, by using techniques close to [13, 49].

Let us distinguish between the following three regimes (see also Table 1):

- for $k \le \sqrt{n}$, the lower bound (11) is of order $n^{-1/2}$, which is the parametric rate that would hold in the case of no contamination (i.e., $k = 0$);
- for $k \in (\sqrt{n}, \zeta n)$ with $\zeta \in (0, 1)$, the lower bound is of the order $(k/n)\log^{-3/2}(k^2/n)$. In particular, in the nonsparse case $k = \lceil n/2 \rceil$, we obtain a rate $\log^{-3/2} n$;
- for $k \in [\zeta n, n-1]$ with $\zeta \in (0, 1)$, the lower bound on the minimax risk is of order $\log^2(\frac{n}{n-k}) \log^{-3/2}(n)$. In particular, for $k = n-1$, the lower bound is of order $\log^{1/2} n$.

In the remainder of this section, we match these lower bounds by considering three different estimators of θ, corresponding to the three regimes discussed above. They are then combined to derive an adaptive estimator.

2.2. *Upper bound for small and large k in the gOSC model with known variance.* For small and for large values of k, the optimal risk is achieved by simple quantile estimators. For $k \le n^{1/2}$, we consider the empirical median defined by

$$\widehat{\theta}_{\mathrm{med}} = Y_{(\lceil n/2 \rceil)}. \tag{12}$$

The following result holds for $\widehat{\theta}_{\mathrm{med}}$ (note that it is stated in the more general OSC model (1)).

PROPOSITION 2.2. *Consider the OSC model (1) with $\sigma = 1$. Then there exist universal positive constants c_1, c_2 and a universal positive integer n_0 such that the following holds. For any $n \ge n_0$, any $k \le n/10$, any $\pi \in \overline{\mathcal{M}}_k$ and any $\theta \in \mathbb{R}$, we have*

$$\mathbb{P}_{\theta,\pi}\left[|\widehat{\theta}_{\mathrm{med}} - \theta| \ge \frac{3(k+1)}{2(n-k)} + 3\frac{\sqrt{(n+1)x}}{n-k} \right] \le e^{-x} \quad \text{for all } x \le c_1 n,$$

$$\mathbb{E}_{\theta,\pi}\big[|\widehat{\theta}_{\mathrm{med}} - \theta| \big] \le \frac{3(k+1)}{2(n-k)} + \frac{c_2}{\sqrt{n}}.$$

A proof is provided in Section S-3.1. A consequence is that, for $k \le \sqrt{n}$, the empirical median $\widehat{\theta}_{\mathrm{med}}$ achieves the parametric rate $n^{-1/2}$, which turns out to be optimal in this regime; see Theorem 2.1. Note that in the regime $k \le \sqrt{n}$, the empirical median was already known to achieve this parametric rate in the more general Huber's contamination model (which allows for two-sided contaminations).

When k is really close to n, there are very few noncontaminated data. Since $\theta = \min_{1 \le i \le n}\{\gamma_i\}$ in the model (4), we consider a debiased empirical minimum estimator

$$\widehat{\theta}_{\mathrm{min}} = Y_{(1)} + \overline{\Phi}^{-1}(1/n), \tag{13}$$

where we recall that $\overline{\Phi}^{-1}(1/n) = \sqrt{2\log(n)} + O(1)$; see Section S-5.2. The following result holds for $\widehat{\theta}_{\mathrm{min}}$ (note that it is again stated in the more general OSC model (1)).

PROPOSITION 2.3. *Consider the OSC model (1) with $\sigma = 1$. Then there exists some universal positive integer n_0 such that for any $n \ge n_0$, any $\pi \in \overline{\mathcal{M}}_{n-1}$ and any $\theta \in \mathbb{R}$, the estimator $\widehat{\theta}_{\mathrm{min}}$ satisfies*

$$\mathbb{P}_{\theta,\pi}\big[|\widehat{\theta}_{\mathrm{min}} - \theta| \ge 2\sqrt{2\log n} \big] \le \frac{2}{n}; \qquad \mathbb{E}_{\theta,\pi}\big[|\widehat{\theta}_{\mathrm{min}} - \theta| \big] \le 2\sqrt{2\log n} + 1.$$

A proof is provided in Section S-3.1. From Theorem 2.1, the estimator $\widehat{\theta}_{\min}$ turns out to be optimal when k is very close to n, that is, when k is larger than $n - n^{\epsilon}$ for a fixed $\epsilon \in (0, 1)$ (very few samples are noncontaminated).

2.3. *Upper bound in the intermediate regime in the gOSC model with known variance.* In the previous section, we have introduced estimators that are optimal in the regimes where $k \leq \sqrt{n}$ and where k is very close to n, respectively. The intermediate case turns out to be much more involved.

Let $q \geq 2$ be an even integer whose value will be fixed below. Let $a = 3[1 + \log(3 + 2\sqrt{2})] \approx 8.29$ and $q_{\max} = \lfloor \frac{1}{2a} \log n \rfloor_{\text{even}} - 2$, where $\lfloor \cdot \rfloor_{\text{even}}$ is defined in Section 1.4. Let us also introduce two rough estimators $\widehat{\theta}_{\text{up}}$ and $\widehat{\theta}_{\text{low},q}$ such that θ is proved to belong to $[\widehat{\theta}_{\text{low},q}, \widehat{\theta}_{\text{up}}]$ with high probability. Let $\widehat{\theta}_{\text{up}} = Y_{(1)} + 2\sqrt{\log n}$. For any positive and even integer q, define $\widehat{\theta}_{\text{low},q} = \widehat{\theta}_{\text{med}} - \overline{v}$ with $\overline{v} = \pi^2/(144 q_{\max}^{3/2})$ if $q \leq \frac{3}{10a} \log n$ and $\widehat{\theta}_{\text{low},q} = -\infty$ for larger q.

To explain the intuition behind our procedure, assume for the purpose of discussion that we have access to the mean $\gamma_i = \theta + \mu_i$, and that instead of estimating θ we simply want to test whether θ is greater than u or not. Thus, our aim is to define a suitable function $\psi_{q,\lambda}(u)$ of u and the γ_i's which is close to zero when all γ_i's are higher than u and is as large as possible when many γ_i's are smaller than u. Since at least $n - k$ of the γ_i's are equal to θ, having large values of $\psi_{q,\lambda}(u)$ means that $\theta < u$. Assuming without loss of generality that $u = 0$, this can be achieved by constructing an auxiliary function $g_q : \mathbb{R} \mapsto \mathbb{R}$ such that $|g_q(x)| \leq 1$ for $x \in (-\infty, 0]$ and $g_q(x)$ large for $x > 0$. If the interval $(-\infty, 0]$ were replaced by $[-1, 1]$ and the function g_q was restricted to be a polynomial, constructing such a g_q would look like a polynomial extremum problem, which is solved by a Chebychev polynomial (see Section S-5.1 for some definitions and properties). To handle the non-bounded interval $(-\infty, 0]$, we map $(-\infty, 0]$ to $(-1, 1]$ using the function $x \mapsto 2e^x - 1$. Denoting by T_q the Chebychev polynomial of degree q, this leads us to considering the function

$$g_q(x) = T_q\big(2e^x - 1\big) = \sum_{j=0}^{q} a_{j,q} e^{xj}, \quad x \in \mathbb{R}, \tag{14}$$

where the coefficients $a_{j,q}$ are defined in (S-65). It follows from the definition of Chebychev polynomials that $g_q(x)$ belongs to $[-1, 1]$ for $x \leq 0$ and $g_q(x) = \cosh[q \operatorname{arccosh}(2e^x - 1)]$ for $x > 0$.

Now, for $\lambda > 0$ and $u \in \mathbb{R}$, consider the function $\psi_{q,\lambda}(u)$ defined by

$$\psi_{q,\lambda}(u) = \frac{1}{n} \sum_{i=1}^{n} g_q\big(\lambda(u - \gamma_i)\big) = \frac{1}{n} \sum_{i=1}^{n} g_q\big(\lambda(u - \theta - \mu_i)\big). \tag{15}$$

This function depends on the γ_i's. Since all μ_i's are nonnegative, it follows from the above observation that $|\psi_{q,\lambda}(u)| \leq 1$ for all $u \leq \theta$. Conversely, for $u \geq \theta$, $\psi_{q,\lambda}(u)$ is lower bounded as follows:

$$\psi_{q,\lambda}(u) \geq -\frac{k}{n} + \frac{n - k}{n} g_q\big(\lambda(u - \theta)\big), \tag{16}$$

hence it is bounded away from 1 as long as $u - \theta$ is large enough. As a consequence, the smallest number u_* that satisfies $\psi_{q,\lambda}(u_*) > 1$ should be close (in some sense) to θ.

Obviously, we do not have access to the function $\psi_{q,\lambda}$ as it requires the knowledge of the γ_i's (or, more precisely, of quantities of the form $e^{-j\lambda\gamma_i}$). Nevertheless, we can still build

an unbiased estimator of such quantities using the empirical Laplace transform of Y. Given $\lambda > 0$ and $u \in \mathbb{R}$, define

$$\widehat{\eta}_\lambda(u) = n^{-1}\sum_{i=1}^{n} e^{\lambda(u-Y_i)-\lambda^2/2}, \qquad \eta_\lambda(u) = n^{-1}\sum_{i=1}^{n} e^{\lambda(u-\theta-\mu_i)}. \tag{17}$$

Since all Y_i's are independent and normally distributed with unit variance, we have $\mathbb{E}[\widehat{\eta}_\lambda(u)] = \eta_\lambda(u)$. This leads us to considering the statistic

$$\widehat{\psi}_{q,\lambda}(u) = \sum_{j=0}^{q} a_{j,q}\widehat{\eta}_{j\cdot\lambda}(u), \tag{18}$$

which is an unbiased estimator of $\psi_{q,\lambda}(u)$ for any fixed $\lambda > 0$ and $u \in \mathbb{R}$. Since $\widehat{\psi}_{q,\lambda}(u)$ approximates $\psi_{q,\lambda}(u)$, we roughly want to take $\widehat{\theta}_q$ to be the smallest value such that $\widehat{\psi}_{q,\lambda}(u)$ is bounded away from 1. This is why we define $\widehat{\theta}_q$ by inverting the function $\widehat{\psi}_{q,\lambda}(\cdot)$. More precisely, for an even integer $q \le q_{\max}$, taking $\lambda_q = \sqrt{2/q}$, we define the estimator $\widehat{\theta}_q$ by

$$\widehat{\theta}_q = \inf\left\{u \in [\widehat{\theta}_{\mathrm{low},q}, \widehat{\theta}_{\mathrm{up}}] : \widehat{\psi}_{q,\lambda_q}(u) > 1 + \frac{e^{aq}}{\sqrt{n}}\right\}, \tag{19}$$

with the convention $\inf \varnothing = \widehat{\theta}_{\mathrm{up}}$.

THEOREM 2.4. *Consider the gOSC model* (4) *with known variance* $\sigma^2 = 1$. *There exist universal positive constants* c_1, c_2, c_3 *and* n_0 *such that the following holds for any* $n \ge n_0$, *any integer* $k \in [e^{2a}\sqrt{n}, n - 64n^{1-1/(4a)})$, *any* $\mu \in \mathcal{M}_k$ *and any* $\theta \in \mathbb{R}$. *The estimator* $\widehat{\theta}_{q_k}$ *defined by* (19) *with* $q_k = \lfloor \frac{1}{a}\log(\frac{k}{\sqrt{n}})\rfloor_{\mathrm{even}} \wedge q_{\max}$ *satisfies*

$$\mathbb{P}_{\theta,\mu}\left(\widehat{\theta}_{q_k} \notin \left[\theta, \theta + c_1\frac{\log^2(1+\sqrt{\frac{k}{n-k}})}{\log^{3/2}(\frac{k^2}{n})}\right]\right) \le c_3\left(\frac{\sqrt{n}}{k}\right)^{4/3}\log^3\left(\frac{k^2}{n}\right) \tag{20}$$

and

$$\mathbb{E}_{\theta,\mu}\left[|\widehat{\theta}_{q_k} - \theta|\right] \le c_2\frac{\log^2(1+\sqrt{\frac{k}{n-k}})}{\log^{3/2}(\frac{k^2}{n})}. \tag{21}$$

A proof is provided in Section S-3.3. This result shows that $\widehat{\theta}_{q_k}$ has a maximum risk of order $\frac{k}{n}\log^{-3/2}(k^2/n)$ in the regime $k \in [e^{2a}\sqrt{n}, n - 64n^{1-1/(4a)})]$. Combined with the lower bound of Theorem 2.1, we have shown that $\widehat{\theta}_{q_k}$ is minimax in the intermediate regime.

REMARK 2.5. Let us emphasize that in the regime $e^{2a}\sqrt{n} \le k \le \lfloor n/2\rfloor$, the minimax risk is of order $(k/n)\log^{-3/2}(n)$, which is faster than the minimax rate $(k/n)\log^{-1/2}(n)$ that we would obtain in a two-sided deconvolution problem, as in [8] where $k/n \propto n^{-\beta}$ (and by considering the extreme case where there is no regularity assumption, that is, $\alpha = 0$ with their notation).

REMARK 2.6. If we are only interested in a probability bound (20) and not in the moment bound (21), the preliminary estimators $\widehat{\theta}_{\mathrm{low},q}$ and $\widehat{\theta}_{\mathrm{up}}$ are not needed: the estimator could be computed by taking the minimum over $\mathbb{R}$ in (19).

REMARK 2.7. The behavior of the estimator family $\{\widehat{\theta}_q\}_q$ given by (19) is illustrated in Section S-6.3 by using numerical experiments. While a fine tuning of the constants is required, this reinforces the theoretical finding that q should be chosen as a function of k in order to provide a small risk.

2.4. *Adaptative estimation in the gOSC model with known variance.* In this section, we combine the three estimators studied in the above section to obtain an estimator that is adaptive with respect to the parameter k. The method stems from the Goldenshluger–Lepski principle; see, for example, [38, 52, 56].

To unify notation, we henceforth write $\widehat{\theta}_0$ for the median estimator $\widehat{\theta}_{\mathrm{med}}$ and $\widehat{\theta}_{q_{\max}+2}$ for the minimum estimator $\widehat{\theta}_{\min}$. In order to obtain an adaptive procedure, we select one of the estimators $\{\widehat{\theta}_q, q \in \{0, 2, \dots, q_{\max}, q_{\max}+2\}\}$ as follows:

$$\widehat{q} = \min\bigl\{q \in \{0, \dots, q_{\max}+2\} \text{ s.t. } |\widehat{\theta}_q - \widehat{\theta}_{q'}| \le \delta_{q'} \text{ for all } q' > q\bigr\}, \tag{22}$$

where the thresholds are chosen such that $\delta_q = 10\frac{e^{a(q+2)}}{\sqrt{n}q^{3/2}}$ for $q \in \{2, \dots, q_{\max}-2\}$, $\delta_{q_{\max}} = \frac{25}{q_{\max}^{3/2}}$ and $\delta_{q_{\max}+2} = 4\sqrt{2\log n}$ (the value of a being the same as in Section 2.3).

THEOREM 2.8. *Consider the gOSC model* (4) *with known variance* $\sigma^2 = 1$. *There exist universal positive constants* c_1, c_2, c_3 *and* n_0 *such that the following holds. For any* $n \ge n_0$, *for any integer* $k \in [1, n-1]$, *any* $\theta \in \mathbb{R}$, *and any* $\mu \in \mathcal{M}_k$, *the adaptive estimator* $\widehat{\theta}_{\mathrm{ad}} = \widehat{\theta}_{\widehat{q}}$ *satisfies*

$$\mathbb{P}_{\theta,\mu}\left[|\widehat{\theta}_{\mathrm{ad}} - \theta| > c_1 \frac{\log^2(1+\sqrt{\frac{k}{n-k}})}{\log^{3/2}(1+(\frac{k^2}{n})^{1/3})}\right] \le c_2\left(\frac{\sqrt{n}}{k \vee \sqrt{n}}\right)^{4/3} \log^3\left(\frac{k \vee (2\sqrt{n})}{\sqrt{n}}\right) \tag{23}$$

and

$$\mathbb{E}_{\theta,\mu}\bigl[|\widehat{\theta}_{\mathrm{ad}} - \theta|\bigr] \le c_3 \frac{\log^2(1+\sqrt{\frac{k}{n-k}})}{\log^{3/2}(1+(\frac{k^2}{n})^{1/3})}. \tag{24}$$

A proof is given in Section S-3.4. The risk bound in (24) matches the minimax lower bound of Theorem 2.1 for all $k \in [1, n-1]$. The estimator $\widehat{\theta}_{\mathrm{ad}}$ is therefore minimax adaptive with respect to k. The qualitative behavior of $\widehat{\theta}_{\mathrm{ad}} = \widehat{\theta}_{\widehat{q}}$ is illustrated in Section S-6.3 via numerical experiments.

REMARK 2.9. Theorem 2.8 implies that the estimation rate is not affected by knowledge of k. This contrasts with other statistical setting where costless adaptation is impossible. For instance, the optimal adaptative rate is slower than the minimax rate in signal detection for the smooth Gaussian white noise model; see [65]. A related phenomenon appears also in extreme value theory, for the problem of adaptively estimating the tail coefficient [12]. In both cases, in the nonadaptive setting, the bias and deviations are of the same order. Intuitively, the deviations have to be slightly inflated to enable adaptivity, which leads to slower rates. This does not happen in our problem because the bias of the statistic turns out to be significantly larger than the deviations (see the proof in the Supplementary Material for more details) and, therefore, a slight inflation of the deviations to account for adaptivity has no impact on the rate.

3. Estimation of θ in the general OSC model with known variance. In this section, we study the estimation problem in the general OSC model (1). Hence, the contaminations are no longer assumed to be Gaussian. Throughout this section, σ is still assumed to be known and equal to 1 (for unknown variance, see Section 4). Recall that the L_1 minimax risk is defined in (7).

3.1. *Lower bound in the OSC model with known variance.* We first show that estimating θ is more difficult in this model than for the gOSC case.

THEOREM 3.1. *There exists a universal positive constant c such that for any positive integer n and for any integer $k \in [1, n-1]$,*

$$\overline{\mathcal{R}}[k,n] \geq c \frac{\log(\frac{n}{n-k})}{\log^{1/2}(1+\frac{k^2}{n})}. \tag{25}$$

A proof is provided in Section S-2.2. Let us comment briefly the order of this lower bound, by considering again the three aforementioned regimes (see also Table 1):

- for $k \leq \sqrt{n}$, the lower bound (25) is of order $n^{-1/2}$, which is the parametric rate, hence is the same as for the Gaussian case;
- for $k \in (\sqrt{n}, \zeta n)$ with $\zeta \in (0,1)$, the lower bound is of the order $(k/n)\log^{-1/2}(k^2/n)$, and so is strictly slower than with the Gaussian assumption (by an additional factor of order $\log(k^2/n)$). In particular, in the nonsparse case $k = \lceil n/2 \rceil$, this gives a lower bound of order $\log^{-1/2}(n)$ (by contrast to the $\log^{-3/2}(n)$ bound in the Gaussian model);
- for $k \in [\zeta n, n-1]$ with $\zeta \in (0,1)$, the lower bound is of order $\log(n/(n-k))\log^{-1/2}(n)$. Compared to the gOSC model, there is an additional factor of order $\log(n)/\log(n/(n-k))$. Nevertheless, in the extreme case $k = n-1$, the two lower bounds are both of order $\log^{1/2}(n)$.

3.2. *Upper bound in the OSC model with known variance.* In this subsection, we introduce a bias-corrected quantile estimator that matches the minimax lower bound of Theorem 3.1. Consider some $\pi \in \overline{\mathcal{M}}_k$. Let $\xi = (\xi_1, \ldots, \xi_n)$ denote a standard Gaussian vector. The starting point is the following: on the one hand, all random variables $Y_i - \theta$ stochastically dominate ξ_i so that $Y_{(q)} - \theta \succeq \xi_{(q)}$. On the other hand, $Y_{(q)}$ is stochastically dominated by the qth smallest observation among the *noncontaminated* data Y_j. As a consequence, we have

$$\xi_{(q)} \preceq Y_{(q)} - \theta \preceq \xi_{(q:(n-k))}, \tag{26}$$

where we recall that $\xi_{(q:(n-k))}$ is the qth largest observation among the $n-k$ first observations of ξ. Since $\xi_{(q)}$ is concentrated around $\overline{\Phi}^{-1}(q/n)$, this leads to introducing the debiased estimator

$$\widetilde{\theta}_q = Y_{(q)} + \overline{\Phi}^{-1}(q/n), \quad 1 \leq q \leq \lceil n/2 \rceil. \tag{27}$$

In view of (12), we have that $\widetilde{\theta}_1 = \widehat{\theta}_{\min}$ while $\widetilde{\theta}_{\lceil n/2 \rceil}$ is almost equal to the empirical median $\widehat{\theta}_{\mathrm{med}}$ (up to the additive $\overline{\Phi}^{-1}(\lceil n/2 \rceil/n)$ term which is of order $1/n$ and so is negligible). The following theorem bounds the error of $\widetilde{\theta}_q$ for a wide range of q.

THEOREM 3.2. *Consider the OSC model (1) with known variance $\sigma^2 = 1$. There exist universal positive constants c_1, c_2, c_2', c_3, c_4 such that the following holds. For any integer $k \in [1, n-1]$, any integer q such that $c_4 \log n \leq q \leq (0.7(n-k)) \wedge \lceil n/2 \rceil$, any $\theta \in \mathbb{R}$ and any $\pi \in \overline{\mathcal{M}}_k$, the estimator $\widetilde{\theta}_q$ satisfies*

$$\mathbb{P}_{\theta,\pi}\left[-c_2 \sqrt{\frac{x}{q[\log(\frac{n-k}{q}) \vee 1]}} \leq \widetilde{\theta}_q - \theta \leq c_1 \frac{\log(\frac{n}{n-k})}{\sqrt{\log(\frac{n-k}{q}) \vee 1}} + c_2 \sqrt{\frac{x}{q[\log(\frac{n-k}{q}) \vee 1]}} \right] \geq 1 - 2e^{-x} \tag{28}$$

for all $0 < x < c_3 q$, and

$$\mathbb{E}_{\theta,\pi}[|\widetilde{\theta}_q - \theta|] \leq c_1 \frac{\log(\frac{n}{n-k})}{\sqrt{\log(\frac{n-k}{q}) \vee 1}} + c_2' \frac{1}{\sqrt{q[\log(\frac{n-k}{q}) \vee 1]}}. \tag{29}$$

A proof is given in Section S-3.5. The risk bound in (29) exhibits a bias/variance trade-off as a function of q via the quantities

$$b(q) = \frac{\log(\frac{n}{n-k})}{\sqrt{\log(\frac{n-k}{q}) \vee 1}}; \qquad s(q) = \frac{1}{\sqrt{q[\log(\frac{n-k}{q}) \vee 1]}}.$$

The quantity $s(q)$ is a deviation term that decreases with q and whose minimum is of the order of $n^{-1/2}$. This minimum is achieved for $q = \lceil n/2 \rceil$ and the corresponding estimator is close to the empirical median. The quantity $b(q)$ is a bias term which increases slowly with q. Its minimum is of the order of $\log(\frac{n}{n-k}) \log^{-1/2}(n-k)$ and is achieved for q constant (or of the order of $\log n$). The corresponding estimators are extreme quantiles such as $\widetilde{\theta}_1 = \widehat{\theta}_{\min}$.

Note also that the condition $c_4 \log(n) \le q \le 0.7(n-k)$ cannot be met when k is too close to n (i.e., $n - k < (c_4/0.7)\log(n)$). Hence, Theorem 3.2 is silent in that regime. Nevertheless, this case is addressed by the minimum estimator $\widetilde{\theta}_1 = \widehat{\theta}_{\min}$ already studied in Proposition 2.3.

To achieve the minimax risk, it remains to suitably choose q as a function of k. Examining $b(\cdot)$ and $s(\cdot)$, we see that when k increases we should decrease q (and, therefore, use a more extreme quantile) in order to decrease the bias. More precisely, we define

$$q_k = \begin{cases} \lceil n/2 \rceil & \text{if } k \in [1, 4\sqrt{n}); \\ \left\lceil \dfrac{n^{5/4}}{k^{1/2}} \right\rceil^{(\log_2)} & \text{if } k \in [4\sqrt{n}, n - n^{4/5}]; \\ 1 & \text{if } k \in (n - n^{4/5}, n-1]. \end{cases} \tag{30}$$

In the very sparse situation ($k \le 4\sqrt{n}$), $\widetilde{\theta}_{q_k}$ corresponds to the empirical median. As k increases toward n, q_k goes smoothly to $n^{1/4}$. Finally, when k is very close to n, we consider the minimum estimator $\widetilde{\theta}_1$. Other choices of q_k may also lead to optimal risk bounds and the choice (30) is made to simplify the proofs.

COROLLARY 3.3. *Consider the OSC model* (1) *with known variance* $\sigma^2 = 1$. *There exist universal positive constants* c *and* n_0 *such that the following holds. For any integer* $n \ge n_0$, *any integer* $k \in [1, n-1]$, *any* $\theta \in \mathbb{R}$ *and any* $\pi \in \overline{\mathcal{M}}_k$, *the estimator* $\widetilde{\theta}_{q_k}$ *satisfies*

$$\mathbb{E}_{\theta,\pi}\big[|\widetilde{\theta}_{q_k} - \theta|\big] \le c \frac{\log(\frac{n}{n-k})}{\log^{1/2}(1 + \frac{k^2}{n})}. \tag{31}$$

A proof is given in Section S-3.5. The estimation rate of the estimator $\widetilde{\theta}_{q_k}$ matches the minimax lower bound given in Theorem 3.1. However, it is not adaptive because it uses the value of k.

3.3. *Adaptive estimation in the OSC model with known variance.* We now provide a procedure that adapts to k, by following a Goldenshluger–Lepski approach. Let $\mathcal{Q}$ denote the collection of values of q_k when k goes from 1 to $n-1$. This collection contains 1, $\lceil n/2 \rceil$ and a dyadic sequence from $n^{1/4}$ to $n/2$ (roughly). To build an adaptive procedure, we select among the estimators $\{\widetilde{\theta}_q, q \in \mathcal{Q}\}$ in the following way:

$$\widehat{q} = \max\big\{q \in \mathcal{Q} \text{ s.t. } |\widehat{\theta}_q - \widehat{\theta}_{q'}| \le \delta_{q'} \text{ for all } q' < q\big\}, \tag{32}$$

where

$$\delta_q = c_0 \begin{cases} \sqrt{\log(n)} & \text{if } q < \sqrt{2} n^{1/4}; \\ \dfrac{n^{1/6}}{q^{2/3}\sqrt{\log(\frac{n}{q}) \vee 1}} & \text{otherwise,} \end{cases} \tag{33}$$

where the constant c_0 is large enough (depending on the constants c_1 and c_2' of Theorem 3.2). Note that at most three elements in $\mathcal{Q}$ are less than $\sqrt{2}n^{1/4}$.

PROPOSITION 3.4. *Consider the OSC model* (1) *with known variance* $\sigma^2 = 1$. *There exist universal positive constants* c *and* n_0 *such that the following holds. For any integer* $n \geq n_0$, *any integer* $k \in [1, n-1]$, *any* $\theta \in \mathbb{R}$, *and any* $\pi \in \overline{\mathcal{M}}_k$, *the estimator* $\widetilde{\theta}_{\mathrm{ad}} = \widetilde{\theta}_{\widehat{q}}$ (*see* (30) *and* (32)) *satisfies*

$$\mathbb{E}_{\theta,\pi}\big[|\widetilde{\theta}_{\mathrm{ad}} - \theta|\big] \leq c \frac{\log(\frac{n}{n-k})}{\log^{1/2}(1 + \frac{k^2}{n})}.$$

A proof is given in Section S-3.5. The above result shows that, as in the Gaussian case, adaptation with respect to k can be achieved without any loss.

4. Estimation in the OSC model with unknown variance. In this section, we consider the OSC model (1) for which the noise variance σ^2 is unknown. We derive the minimax risks and estimators for θ and σ in that setting.

4.1. *Lower bound in the OSC model with unknown variance.* First, note that, obviously, the lower bound (25) for estimating θ is also a valid lower bound for the minimax risk

$$\inf_{\widetilde{\theta}} \sup_{\theta \in \mathbb{R}, \sigma > 0, \pi \in \overline{\mathcal{M}}_k} \mathbb{E}_{\theta,\pi,\sigma}\left[\frac{|\widetilde{\theta} - \theta|}{\sigma}\right],$$

corresponding to the OSC model (1) where σ is unknown.

Now, let us provide a lower bound for the estimation risk of σ. Since the problem of estimating σ when θ is unknown is at least as difficult as the problem when it is known that $\theta = 0$, we provide a lower bound for the latter. Interestingly, the procedure that will be presented in Section 4.2 achieves the same risk even in the more general setting where $\theta \in \mathbb{R}$ is unknown. We introduce the minimax risk

$$\overline{\mathcal{R}}_v[k, n] = \inf_{\widetilde{\sigma}} \sup_{\sigma > 0, \pi \in \overline{\mathcal{M}}_k} \mathbb{E}_{0,\pi,\sigma}\left[\frac{|\widetilde{\sigma} - \sigma|}{\sigma}\right]. \tag{34}$$

The following theorem provides a lower bound for $\overline{\mathcal{R}}_v[k, n]$ (and, therefore, also a lower bound on the minimax risk with arbitrary unknown θ).

THEOREM 4.1. *There exists a universal positive constant* c *such that for any integer* $n \geq 2$ *and any integer* $k \in [1, n-2]$, *we have*

$$\overline{\mathcal{R}}_v[k, n] \geq c \frac{\log(\frac{n}{n-k})}{\log(1 + \frac{k}{n^{1/2}})}. \tag{35}$$

A proof is given in Section S-2.3. For $k \leq \sqrt{n}$, the lower bound (35) is of order $n^{-1/2}$. For $k \in [\sqrt{n}, \zeta n]$ (with $\zeta \in (0, 1)$ fixed), the risk is of order $k/[n \log(k^2/n)]$ which is faster by a $\log^{1/2}(k^2/n)$ term than for mean estimation. When $n - k = n^\gamma$ with $\gamma \in (0, 1)$ (almost no uncontaminated data), the relative risk of convergence is at least of constant order.

In the next section, we prove that these lower bounds on θ and σ are all sharp (up to numerical constants).

4.2. *Upper bound for the OSC model with unknown variance.* Since the model is translation invariant, the variance can be estimated without knowing θ. This is done by considering a rescaled difference of empirical quantiles. More precisely, for two positive integers $1 \le q' \le q \le n$, let

$$\widetilde{\sigma}_{q,q'} = \frac{Y_{(q)} - Y_{(q')}}{\overline{\Phi}^{-1}(q'/n) - \overline{\Phi}^{-1}(q/n)}, \tag{36}$$

with the convention $0/0 = 0$. When $k = 0$ (no contamination), $Y_{(q)}$ (resp., $Y_{(q')}$) should be close to $\theta - \sigma\overline{\Phi}^{-1}(q/n)$ (resp. $\theta - \sigma\overline{\Phi}^{-1}(q'/n)$) so that, intuitively, $\widetilde{\sigma}_{q,q'}$ should be close to σ. Then, to estimate θ, we simply plug $\widetilde{\sigma}_{q,q'}$ into the quantile estimator considered in Section 3.2. More precisely, we consider

$$\widetilde{\theta}_{q,q'} = Y_{(q)} + \widetilde{\sigma}_{q,q'}\overline{\Phi}^{-1}\left(\frac{q}{n}\right). \tag{37}$$

Given $k \in \{1, \dots, n-1\}$, q_k is taken as in (30) and

$$q'_k = \begin{cases} \lceil n/3 \rceil & \text{if } k \in [1, 4\sqrt{n}); \\ \left\lfloor \dfrac{n^{7/4}}{k^{3/2}} \right\rfloor^{(\log_2)} & \text{if } k \in \left[4\sqrt{n}, n - n^{4/5}\right]; \\ 1 & \text{if } k \in \left(n - n^{4/5}, n - 2\right]. \end{cases} \tag{38}$$

For sparse contaminations ($k < 4\sqrt{n}$), $\widetilde{\sigma}_{q_k,q'_k}$ is a rescaled difference of the empirical median and the empirical quantile of order $1/3$. For a larger number of contaminations, more extreme quantiles are considered. For $k \ge n - n^{4/5}$, we simply take $\widetilde{\sigma}_{q_k,q'_k} = 0$.

PROPOSITION 4.2. *Consider the OSC model* (1) *with unknown variance* σ^2 *and the quantities* q_k *and* q'_k *defined in* (30) *and* (38). *There exist universal positive constants* c, c' *and* n_0 *such that the following holds. For any* $n \ge n_0$, *for any integer* $k \in [1, n-2]$, *any* $\theta \in \mathbb{R}$, *any* $\sigma > 0$ *and any* $\pi \in \overline{\mathcal{M}}_k$, *we have*

$$\mathbb{E}_{\theta,\pi,\sigma}\left[|\widetilde{\sigma}_{q_k,q'_k} - \sigma|/\sigma\right] \le c\frac{\log(\frac{n}{n-k})}{\log(1 + \frac{k}{n^{1/2}})}; \tag{39}$$

$$\mathbb{E}_{\theta,\pi,\sigma}\left[|\widetilde{\theta}_{q_k,q'_k} - \theta|/\sigma\right] \le c'\frac{\log(\frac{n}{n-k})}{\log^{1/2}(1 + \frac{k^2}{n})}. \tag{40}$$

A proof is given in Section S-3.5. The above proposition together with the lower bounds of Section 4.1 implies that $\widetilde{\sigma}_{q_k,q'_k}$ and $\widetilde{\theta}_{q_k,q'_k}$ are minimax estimator of σ and θ, respectively.

Interestingly, the estimators $\widetilde{\sigma}_{q_k,q'_k}$ and $\widetilde{\theta}_{q_k,q'_k}$ do not require the knowledge of either θ or σ whereas our lower bounds were respectively restricted to known θ and known σ settings. This entails that the knowledge of one parameter (θ or σ) does not significantly ease the estimation of the other.

5. Controlled selection of outliers. In this section, we consider the OSC model (1) (with unknown variance) and now turn to the identification of the outliers. As described in Section 1.3, this can be reformulated as a multiple testing problem (see also the notation therein).

5.1. *Rescaled p-values.* As already discussed in Section 1.3, ensuring good multiple testing properties in the OSC model is challenging because the scaling parameters θ and σ are unknown. A natural approach is then to use the rescaled observations $Y_i' = (Y_i - \widehat{\theta})/\widehat{\sigma}$, $1 \le i \le n$, where $\widehat{\theta}, \widehat{\sigma}$ are some suitable estimators of θ and σ. To formalize further this idea, let us consider the corrected p-values

$$p_i(u,s) = \overline{\Phi}\left(\frac{Y_i - u}{s}\right), \quad u \in \mathbb{R}, s > 0, 1 \le i \le n. \tag{41}$$

The perfectly corrected p-values thus correspond to

$$p_i^\star = p_i(\theta, \sigma), \quad 1 \le i \le n. \tag{42}$$

These oracle p-values cannot be used in practice, because they depend on the unknown parameters θ and σ. Our general aim is to build estimators $\widehat{\theta}$, $\widehat{\sigma}$ such that the theoretical performance of the corrected p-values $p_i(\widehat{\theta}, \widehat{\sigma})$ mimic those of the oracle p-values $p_i^\star$, when plugged into standard multiple testing or post hoc procedures. Although the use of modified p-values and plug-in estimators has often been advocated since the seminal work of Efron [25], proving that the behavior of the plug-in procedure is asymptotically similar to that of the oracle procedure is, to our knowledge, new. The challenge is to precisely quantify how the estimation error affects the FDP/TDP metrics. For this, a key point is the following relation between $p_i(u,s)$ and $p_i^\star$:

$$\{p_i(u,s) \le t\} = \{p_i^\star \le U_{u,s}(t)\}, \quad i \in \{1, \dots, n\}, t \in [0,1], \tag{43}$$

where

$$U_{u,s}(t) = \overline{\Phi}\left(\frac{s}{\sigma}\overline{\Phi}^{-1}(t) + \frac{u-\theta}{\sigma}\right); \qquad U_{u,s}^{-1}(v) = \overline{\Phi}\left(\frac{\sigma}{s}\overline{\Phi}^{-1}(v) + \frac{\theta - u}{s}\right). \tag{44}$$

Furthermore, a useful property is that the order of the p-values does not change after rescaling. We will denote

$$0 = p_{(0)}(u,s) \le p_{(1)}(u,s) \le \cdots \le p_{(n)}(u,s), \tag{45}$$

the ordered elements of $\{p_i(u,s), 1 \le i \le n\}$. We also denote by $0 = p_{(0:\mathcal{H}_0)}(u,s) \le p_{(1:\mathcal{H}_0)}(u,s) \le \cdots \le p_{(n_0:\mathcal{H}_0)}(u,s)$ the ordered elements of the subset $\{p_i(u,s), i \in \mathcal{H}_0\}$, that is, of the p-value set corresponding to false outliers (or, equivalently, true null hypotheses).

5.2. *Upper-biased estimators.* This section provides estimators $\widetilde{\theta}_+$, $\widetilde{\sigma}_+$ that will be suitable to make the p-value rescaling. They are similar to the estimators introduced in Sections 3.2 and 4.2. However, since minimax estimation and false outliers control do not use the same risk metrics, we need to slightly modify these estimators, especially by making them upper-biased (which roughly means that the null hypotheses are favored).

For $q_n = \lfloor n^{3/4} \rfloor$ and $q_n' = \lfloor n^{1/4} \rfloor$, let us consider

$$\begin{cases} \widetilde{\theta}_+ = Y_{(q_n)} + \widetilde{\sigma}_+ \overline{\Phi}^{-1}\left(\dfrac{q_n}{n}\right); \\ \widetilde{\sigma}_+ = \dfrac{Y_{(q_n)} - Y_{(q_n')}}{\overline{\Phi}^{-1}(q_n'/(n-k_0)) - \overline{\Phi}^{-1}(q_n/n)} \end{cases} \tag{46}$$

for some parameter $k_0 \le \lfloor 0.9n \rfloor$. The key difference with the estimators $\widetilde{\theta}_{q,q'}, \widetilde{\sigma}_{q,q'}$ of Section 4 is the quantity k_0 appearing in the denominator of $\widetilde{\sigma}_+$. The following result holds.

PROPOSITION 5.1. *Consider the OSC model* (1) *with unknown variance* σ^2. *Then there exist two universal positive constants* c, c' *such that the following holds for any positive integer* n, *for any* $\theta \in \mathbb{R}$, $\sigma > 0$, *for any* $\pi \in \overline{\mathcal{M}}_k$ *with* $k = \lfloor 0.9n \rfloor$. *Choosing* k_0 *such that* $n_1(\pi) \le k_0 \le \lfloor 0.9n \rfloor$ *within the estimators* $\widetilde{\theta}_+$, $\widetilde{\sigma}_+$ (46), *we have*

$$\mathbb{P}_{\theta,\pi,\sigma}\big(\widetilde{\theta}_+ - \theta \le -\sigma n^{-1/16}\big) \le c/n; \tag{47}$$

$$\mathbb{P}_{\theta,\pi,\sigma}\big(\widetilde{\sigma}_+ - \sigma \le -\sigma n^{-1/16}\big) \le c/n; \tag{48}$$

$$\mathbb{P}_{\theta,\pi,\sigma}\big(|\widetilde{\theta}_+ - \theta| \ge \sigma\big(c'(k_0/n)\log^{-1/2}(n) + n^{-1/16}\big)\big) \le c/n; \tag{49}$$

$$\mathbb{P}_{\theta,\pi,\sigma}\big(|\widetilde{\sigma}_+ - \sigma| \ge \sigma\big(c'(k_0/n)\log^{-1}(n) + n^{-1/16}\big)\big) \le c/n. \tag{50}$$

Proposition 5.1 is proved in Section S-3.5. It is closely related to Theorem 3.2 above, although the statement is slightly different because of the introduced bias in the estimators. Inequalities (47)–(48) entail that the estimators are (with high probability) above the targeted quantity minus a polynomial term, which will be particularly suitable for obtaining a control on the false positives (FDR control and post hoc bounds). Inequalities (49)–(50) are two-sided, which is useful for studying the power (TDP) of the rescaled procedures: there is an additional error term of order $(k_0/n)\log^{-a}(n)$, $a \in \{1/2, 1\}$, where k_0 corresponds to a known upper bound of the number of contaminated coordinates of π.

The assumption $n_1(\pi) \le k_0 \le \lfloor 0.9n \rfloor$ in Proposition 5.1 is not very restrictive: it means that the number of outliers is bounded by above by some quantity k_0, which is used in the definition of the estimators (46). For instance, taking $k_0 = \lfloor 0.7n \rfloor$ means that we assume that there is no more than 70% of outliers in the data.

Finally, note that our multiple testing analysis will only rely on the deviation bounds (47)–(50). As a consequence, other estimators satisfying these properties can be used for scaling the p-values.

5.3. *FDR control and power optimality for selected outliers.* The Benjamini–Hochberg (BH) procedure, introduced in [4], has probably been the most widely used multiple testing procedure. Here, the rescaled BH procedure (of nominal level α), denoted $\mathrm{BH}_\alpha(u, s)$ is defined from the p-value family $p_i(u, s)$, $1 \le i \le n$, as follows:

- order the p-values as in (45);
- consider $\widehat{\ell}_\alpha(u, s) = \max\{\ell \in \{0, 1, \dots, n\} : p_{(\ell)}(u, s) \le \alpha\ell/n\}$;
- reject $H_{0,i}$ for any i such that $p_i(u, s) \le \widehat{t}_\alpha(u, s)$, for $\widehat{t}_\alpha(u, s) = \alpha\widehat{\ell}_\alpha(u, s)/n$.

The procedure, identified to the set of selected outliers, is then given by

$$\mathrm{BH}_\alpha(u, s) = \big\{1 \le i \le n : p_i(u, s) \le \widehat{t}_\alpha(u, s)\big\}. \tag{51}$$

The classical FDR-controlling result of [4, 5] can be re-interpreted as follows: the BH procedure using the perfectly corrected p-values (42), that is, $\mathrm{BH}^\star_\alpha = \mathrm{BH}_\alpha(\theta, \sigma)$, satisfies

$$\mathbb{E}_{\theta,\pi,\sigma}\big(\mathrm{FDP}\big(\pi, \mathrm{BH}^\star_\alpha\big)\big) = \frac{n_0}{n}\alpha \le \alpha \quad \text{for all } \theta, \pi, \sigma.$$

This comes from the fact that the perfectly corrected p-values (42) are independent, with uniform marginal distributions under the null hypothesis.

Recall the estimators $\widetilde{\theta}_+$ and $\widetilde{\sigma}_+$ defined in (46) with the tuning parameter k_0. The next result gives the behavior of the rescaled procedure $\mathrm{BH}_\alpha(\widetilde{\theta}_+, \widetilde{\sigma}_+)$ both in terms of FDP and TDP.

THEOREM 5.2. *Consider the OSC model (1) with unknown variance σ^2. There exists a universal positive constant c such that the following holds. Considering the estimators $\widetilde{\theta}_+, \widetilde{\sigma}_+$ given by (46), for any $\alpha \in (0, 0.4)$, $\theta \in \mathbb{R}$, $\sigma > 0$, and any $\pi \in \overline{\mathcal{M}}_{\lfloor 0.9n \rfloor}$, if k_0 in $\widetilde{\sigma}_+$ satisfies $n_1(\pi) \le k_0 \le \lfloor 0.9n \rfloor$, we have*

$$\left(\mathbb{E}_{\theta,\pi,\sigma}\big(\mathrm{FDP}(\pi, \mathrm{BH}_\alpha(\widetilde{\theta}_+, \widetilde{\sigma}_+))\big) - \frac{n_0}{n}\alpha\right)_+ \le c \log(n)/n^{1/16}. \tag{52}$$

Additionally, for any sequence $\epsilon_n \in (0, 1)$ tending to zero with $\epsilon_n \gg \log^{-1/2}(n)$, for any sequence $\pi = \pi_n$, $k_0 = k_{0,n}$ with $n_1(\pi_n) \le k_{0,n} \le \lfloor 0.9n \rfloor$ and $n_1(\pi_n)/n \asymp k_{0,n}/n$, we have for all $\theta \in \mathbb{R}$, $\sigma > 0$,

$$\begin{aligned}\limsup_n \big\{&\mathbb{E}_{\theta,\pi_n,\sigma}\big(\mathrm{TDP}(\pi_n, \mathrm{BH}^\star_\alpha)\big)\\ &- \mathbb{E}_{\theta,\pi_n,\sigma}\big(\mathrm{TDP}\big(\pi_n, \mathrm{BH}_{\alpha(1+\epsilon_n)}(\widetilde{\theta}_+, \widetilde{\sigma}_+)\big)\big)\big\} \le 0.\end{aligned} \tag{53}$$

In a nutshell, inequalities (52) and (53) show that the selection procedure $\mathrm{BH}_\alpha(\widetilde{\theta}_+, \widetilde{\sigma}_+)$ behaves similar to the oracle procedure $\mathrm{BH}^\star_\alpha$, both in terms of false discovery rate control and power. Note that the power result (53) is shown to be valid only when k_0/n is taken of the same order as $n_1(\pi)/n$, that is, for a procedure using information on $n_1(\pi)/n$. On the positive side, it entails an optimal procedure even without sparsity: if we know that $n_1(\pi)/n$ is of the order of a constant, then using the estimators $\widetilde{\theta}_+, \widetilde{\sigma}_+$ for say $k_0 = \lfloor 0.5n \rfloor$, we will always achieve both (52) and (53). In particular, Theorem 5.2 can be seen as a first step toward the validation of Efron's empirical null principle, in a specific one-sided situation.

The proof of this theorem is given in Section S-4.1. Compared to the usual FDR proofs of the existing literature, there are two additional difficulties: first, the independence assumption between the corrected p-values is not satisfied anymore, because the correction terms are random; second, the quantity $\mathrm{FDP}(\pi, \mathrm{BH}_\alpha(\widetilde{\theta}_+, \widetilde{\sigma}_+))$ is not monotone in the estimators $\widetilde{\theta}_+$, $\widetilde{\sigma}_+$, because of the denominator of the FDP. However, the specific properties of $\widetilde{\theta}_+$, $\widetilde{\sigma}_+$ given in Proposition 5.1 will be enough to get our result: first, these estimators are biased upwards with an error term vanishing at a polynomial rate $n^{-1/16}$, which is enough for false positive control. As for the power, we consider the bias downwards, which is of order $(k_0/n)\log^{-a}(n)$, $a \in \{1/2, 1\}$. It turns out that the error term induced in the power is of order $(k_0/n)\log^{-a}(n)\log(n/n_1)$, which tends to 0 when $k_0/n \asymp n_1/n$ both in the sparse and nonsparse cases.

The other multiple testing inferences described in the Introduction (Section 1.3): post hoc bounds, decorrelation and the corresponding numerical experiments are postponed to the Supplementary Material (see Section S-1).

REMARK 5.3. Let us define the risk of a procedure as FDR + FNR, where FNR denotes the average of 1 − TDP. Theorem 5.2 shows that the risk of the selection procedure $\mathrm{BH}_{\alpha(1+\epsilon_n)}(\widetilde{\theta}_+, \widetilde{\sigma}_+)$ is smaller than the risk $\mathcal{R}^*(n, \pi_n, \alpha)$ of the oracle procedure $\mathrm{BH}^\star_\alpha$ plus some vanishing terms. Interestingly, the risk $\mathcal{R}^*(n, \pi_n, \alpha)$ has been studied in [1] for a particular choice of π_n coming from the classical Gaussian sequence model in the specific Donoho–Jin sparse regime [22]. It is shown that $\mathcal{R}^*(n, \pi_n, \alpha_n)$ tends to zero when the signal strength is above the estimation boundary (as defined, e.g., in [22]), for some proper calibration of α_n, typically $\alpha_n = 1/(\log n)$. Inspecting our proofs (see Section S-4.1), we can derive that the procedure $\mathrm{BH}_{\alpha_n(1+\epsilon_n)}(\widetilde{\theta}_+, \widetilde{\sigma}_+)$ also has a risk tending to zero in that regime. In addition, the work [59] provides a polynomial decaying rate for $\mathcal{R}^*(n, \pi_n, \alpha_n)$ when $\alpha_n = n^{-\kappa}$ for some $\kappa > 0$. Again, inspecting our proofs (see Section S-4.1), this also holds for the risk of $\mathrm{BH}_{\alpha_n(1+\epsilon_n)}(\widetilde{\theta}_+, \widetilde{\sigma}_+)$, as our remainder term also vanishes at a polynomial rate in that regime, although our exponent may be smaller (e.g., n^{-c} with $c < 1/16$). Finally, we emphasize the fact that Theorem 5.2 does not rely on the Donoho–Jin sparse regime. Hence, this type of power result is new in multiple testing literature to the best of our knowledge.

6. Discussion. In this section, we describe some possible extensions and future research directions.

Knowledge of the variance in the gOSC model. In the gOSC model (Section 2), we assumed exact knowledge of the variance. Unfortunately, plugging an estimator of the variance into the empirical Laplace transform from (17) seems to deteriorate the deviations of the corresponding estimator $\widehat{\psi}_{q,\lambda}(u)$. For this reason, we were not able to derive an estimator of θ with σ unknown achieving the optimal rate $\mathcal{R}[k,n]$ in the intermediate regime where $k \in [\sqrt{n}; cn]$, $c \in (0,1)$. Since the gOSC model is a specific case of the OSC model, it is possible to achieve the rate $\overline{\mathcal{R}}[k,n]$ using $\widetilde{\theta}_{q_k,q'_k}$ (37). Since there is $\log(k^2/n)$ gap between the gOSC and OSC models, it still remains unclear if perfect adaptation to unknown σ is possible in the gOSC model or if a logarithmic price must be paid.

Simultaneous adaptation to k and σ in the OSC model. In Sections 3 and 4, we assumed that either σ or k is known. If both quantities are unknown, one needs to select a suitable estimator $\widetilde{\theta}_{q,q'}$ in the collection $\{\widetilde{\theta}_{q_k,q'_k}\}$ introduced in Section 4. This turns out to be possible using a Goldenshluger–Lepski scheme in the spirit of the one studied in Section 3.3 provided that we have at our disposal an estimator $\widetilde{\sigma}$ of σ satisfying $|\widetilde{\sigma}^2 - \sigma^2|/\sigma^2 \in [1/2, 2]$ with high probability. In view of Section 4, this is the case when k is not close to n ($k \le n - n/e^{c\log^{1/2}(n)}$). As a result, simultaneous adaptation to k and σ is possible provided that the true number of contaminations is not too large. If k is allowed to be as large as $n-2$, it remains an open question whether there is price to pay.

Multivariate one-sided contaminations. Throughout this manuscript, we assumed that the observations are univariate. One could think of a multivariate extension: $Y_i = \theta + \sigma\varepsilon_i$ where $Y_i \in \mathbb{R}^d$ and ε_i is either distributed as a standard normal distribution or, for a contaminated observation, satisfies $\varepsilon_{i,j} \succ \mathcal{N}(0,1)$ for each coordinate $j = 1, \dots, d$. In Huber's original multivariate contamination model, it is known that estimating each coordinate θ independently is suboptimal. This led Huber to introduce Tukey depth estimators [41]. More recently, polynomial time estimators were proved to achieve similar performances [21] (up to some logarithmic loss). It would be interesting to investigate whether it is possible to build upon the one-sided assumption to improve their convergence rates.

Acknowledgments. We would like to warmly thank the Associate Editor and the reviewers for their constructive remarks. We also thank Kweku Abraham and James Cheshire for their comments that have much improved the presentation of the manuscript. The work of A. Carpentier is partially supported by the Deutsche Forschungsgemeinschaft (DFG) Emmy Noether grant MuSyAD (CA 1488/1-1), by the DFG—314838170, GRK 2297 MathCoRe, by the DFG GRK 2433 DAEDALUS, by the DFG CRC 1294 "Data Assimilation," Project A03 and by the UFA-DFH through the French–German Doktorandenkolleg CDFA 01-18. This work has also been supported by ANR-16-CE40-0019 (SansSouci), ANR-17-CE40-0001 (BASICS) and by the GDR ISIS through the "projets exploratoires" program (project TASTY).

SUPPLEMENTARY MATERIAL

Supplement to "Estimating minimum effect with outlier selection" (DOI: 10.1214/20-AOS1956SUPP; .pdf). This supplement contains additional materials for the outlier selection problem, the proofs of the paper, auxiliary results and numerical experiments.

REFERENCES

[1] ARIAS-CASTRO, E. and CHEN, S. (2017). Distribution-free multiple testing. *Electron. J. Stat.* **11** 1983–2001. MR3651021 https://doi.org/10.1214/17-EJS1277

[2] BARAUD, Y. (2002). Non-asymptotic minimax rates of testing in signal detection. *Bernoulli* **8** 577–606. MR1935648

[3] BARBER, R. F. and CANDÈS, E. J. (2015). Controlling the false discovery rate via knockoffs. *Ann. Statist.* **43** 2055–2085. MR3375876 https://doi.org/10.1214/15-AOS1337

[4] BENJAMINI, Y. and HOCHBERG, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B* **57** 289–300. MR1325392

[5] BENJAMINI, Y. and YEKUTIELI, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Ann. Statist.* **29** 1165–1188. MR1869245 https://doi.org/10.1214/aos/1013699998

[6] BLANCHARD, G., NEUVIAL, P. and ROQUAIN, E. (2020). Post hoc confidence bounds on false positives using reference families. *Ann. Statist.* **40** 1281–1303. MR4124323 https://doi.org/10.1214/19-AOS1847

[7] BOGDAN, M., VAN DEN BERG, E., SABATTI, C., SU, W. and CANDÈS, E. J. (2015). SLOPE—Adaptive variable selection via convex optimization. *Ann. Appl. Stat.* **9** 1103–1140. MR3418717 https://doi.org/10.1214/15-AOAS842

[8] CAI, T. T. and JIN, J. (2010). Optimal rates of convergence for estimating the null density and proportion of nonnull effects in large-scale multiple testing. *Ann. Statist.* **38** 100–145. MR2589318 https://doi.org/10.1214/09-AOS696

[9] CAI, T. T. and LOW, M. G. (2005). Nonquadratic estimators of a quadratic functional. *Ann. Statist.* **33** 2930–2956. MR2253108 https://doi.org/10.1214/009053605000000147

[10] CAI, T. T. and LOW, M. G. (2011). Testing composite hypotheses, Hermite polynomials and optimal estimation of a nonsmooth functional. *Ann. Statist.* **39** 1012–1041. MR2816346 https://doi.org/10.1214/10-AOS849

[11] CARPENTIER, A., DELATTRE, S., ROQUAIN, E. and VERZELEN, N. (2021). Supplement to "Estimating minimum effect with outlier selection." https://doi.org/10.1214/20-AOS1956SUPP

[12] CARPENTIER, A. and KIM, A. K. H. (2015). Adaptive and minimax optimal estimation of the tail coefficient. *Statist. Sinica* **25** 1133–1144. MR3410301

[13] CARPENTIER, A. and VERZELEN, N. (2019). Adaptive estimation of the sparsity in the Gaussian vector model. *Ann. Statist.* **47** 93–126. MR3909928 https://doi.org/10.1214/17-AOS1680

[14] CHEN, M., GAO, C. and REN, Z. (2018). Robust covariance and scatter matrix estimation under Huber's contamination model. *Ann. Statist.* **46** 1932–1960. MR3845006 https://doi.org/10.1214/17-AOS1607

[15] COLLIER, O., COMMINGES, L. and TSYBAKOV, A. B. (2017). Minimax estimation of linear and quadratic functionals on sparsity classes. *Ann. Statist.* **45** 923–958. MR3662444 https://doi.org/10.1214/15-AOS1432

[16] COLLIER, O., COMMINGES, L. and TSYBAKOV, A. B. (2018). On estimation of nonsmooth functionals of sparse normal means. Preprint. Available at arXiv:1805.10791.

[17] COLLIER, O., COMMINGES, L., TSYBAKOV, A. B. and VERZELEN, N. (2018). Optimal adaptive estimation of linear functionals under sparsity. *Ann. Statist.* **46** 3130–3150. MR3851767 https://doi.org/10.1214/17-AOS1653

[18] DELATTRE, S. and ROQUAIN, E. (2011). On the false discovery proportion convergence under Gaussian equi-correlation. *Statist. Probab. Lett.* **81** 111–115. MR2740072 https://doi.org/10.1016/j.spl.2010.09.025

[19] DELATTRE, S. and ROQUAIN, E. (2015). New procedures controlling the false discovery proportion via Romano–Wolf's heuristic. *Ann. Statist.* **43** 1141–1177. MR3346700 https://doi.org/10.1214/14-AOS1302

[20] DELATTRE, S. and ROQUAIN, E. (2016). On empirical distribution function of high-dimensional Gaussian vector components with an application to multiple testing. *Bernoulli* **22** 302–324. MR3449784 https://doi.org/10.3150/14-BEJ659

[21] DIAKONIKOLAS, I., KAMATH, G., KANE, D. M., LI, J., MOITRA, A. and STEWART, A. (2017). Being robust (in high dimensions) can be practical. In *Proceedings of the* 34*th International Conference on Machine Learning*. *Proceedings of Machine Learning Research* **70** 999–1008.

[22] DONOHO, D. and JIN, J. (2004). Higher criticism for detecting sparse heterogeneous mixtures. *Ann. Statist.* **32** 962–994. MR2065195 https://doi.org/10.1214/009053604000000265

[23] DONOHO, D. L. and NUSSBAUM, M. (1990). Minimax quadratic estimation of a quadratic functional. *J. Complexity* **6** 290–323. MR1081043 https://doi.org/10.1016/0885-064X(90)90025-9

[24] DUDOIT, S. and VAN DER LAAN, M. J. (2008). *Multiple Testing Procedures with Applications to Genomics*. *Springer Series in Statistics*. Springer, New York. MR2373771 https://doi.org/10.1007/978-0-387-49317-6

[25] EFRON, B. (2004). Large-scale simultaneous hypothesis testing: The choice of a null hypothesis. *J. Amer. Statist. Assoc.* **99** 96–104. MR2054289 https://doi.org/10.1198/016214504000000089

[26] EFRON, B. (2007). Correlation and large-scale simultaneous significance testing. *J. Amer. Statist. Assoc.* **102** 93–103. MR2293302 https://doi.org/10.1198/016214506000001211

[27] EFRON, B. (2007). Doing thousands of hypothesis tests at the same time. *Metron* **LXV** 3–21.
[28] EFRON, B. (2009). Empirical Bayes estimates for large-scale prediction problems. *J. Amer. Statist. Assoc.* **104** 1015–1028. MR2562003 https://doi.org/10.1198/jasa.2009.tm08523
[29] EFRON, B. (2010). Correlated z-values and the accuracy of large-scale statistical estimates. *J. Amer. Statist. Assoc.* **105** 1042–1055. MR2752597 https://doi.org/10.1198/jasa.2010.tm09129
[30] FAN, J. and HAN, X. (2017). Estimation of the false discovery proportion with unknown dependence. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **79** 1143–1164. MR3689312 https://doi.org/10.1111/rssb.12204
[31] FAN, J., HAN, X. and GU, W. (2012). Estimating false discovery proportion under arbitrary covariance dependence. *J. Amer. Statist. Assoc.* **107** 1019–1035. MR3010887 https://doi.org/10.1080/01621459.2012.720478
[32] FINNER, H., DICKHAUS, T. and ROTERS, M. (2007). Dependency and false discovery rate: Asymptotics. *Ann. Statist.* **35** 1432–1455. MR2351092 https://doi.org/10.1214/009053607000000046
[33] FRIGUET, C., KLOAREG, M. and CAUSEUR, D. (2009). A factor model approach to multiple testing under dependence. *J. Amer. Statist. Assoc.* **104** 1406–1415. MR2750571 https://doi.org/10.1198/jasa.2009.tm08332
[34] GAVRILOV, Y., BENJAMINI, Y. and SARKAR, S. K. (2009). An adaptive step-down procedure with proven FDR control under independence. *Ann. Statist.* **37** 619–629. MR2502645 https://doi.org/10.1214/07-AOS586
[35] GENOVESE, C. and WASSERMAN, L. (2004). A stochastic process approach to false discovery control. *Ann. Statist.* **32** 1035–1061. MR2065197 https://doi.org/10.1214/009053604000000283
[36] GENOVESE, C. R. and WASSERMAN, L. (2006). Exceedance control of the false discovery proportion. *J. Amer. Statist. Assoc.* **101** 1408–1417. MR2279468 https://doi.org/10.1198/016214506000000339
[37] GOEMAN, J. J. and SOLARI, A. (2011). Multiple testing for exploratory research. *Statist. Sci.* **26** 584–597. MR2951390 https://doi.org/10.1214/11-STS356
[38] GOLDENSHLUGER, A. and LEPSKI, O. (2011). Bandwidth selection in kernel density estimation: Oracle inequalities and adaptive minimax optimality. *Ann. Statist.* **39** 1608–1632. MR2850214 https://doi.org/10.1214/11-AOS883
[39] GUO, W., HE, L. and SARKAR, S. K. (2014). Further results on controlling the false discovery proportion. *Ann. Statist.* **42** 1070–1101. MR3210996 https://doi.org/10.1214/14-AOS1214
[40] HAN, Y., JIAO, J. and WEISSMAN, T. (2016). Minimax estimation of KL divergence between discrete distributions. Preprint. Available at arXiv:1605.09124.
[41] HUBER, P. J. (1964). Robust estimation of a location parameter. *Ann. Math. Stat.* **35** 73–101. MR0161415 https://doi.org/10.1214/aoms/1177703732
[42] HUBER, P. J. (2011). Robust statistics. In *International Encyclopedia of Statistical Science* 1248–1251. Springer, Berlin.
[43] IBRAGIMOV, I. A. and KHASMINSKII, R. Z. (1985). On nonparametric estimation of the value of a linear functional in Gaussian white noise. *Theory Probab. Appl.* **29** 18–32.
[44] IGNATIADIS, N. and HUBER, W. (2017). Covariate powered cross-weighted multiple testing. Preprint. Available at arXiv:1701.05179.
[45] INGSTER, YU. I. and SUSLINA, I. A. (2012). *Nonparametric Goodness-of-Fit Testing Under Gaussian Models. Lecture Notes in Statistics* **169**. Springer, New York. MR1991446 https://doi.org/10.1007/978-0-387-21580-8
[46] JIAO, J., HAN, Y. and WEISSMAN, T. (2016). Minimax estimation of the L_1 distance. In 2016 *IEEE International Symposium on Information Theory* (*ISIT*) 750–754.
[47] JIN, J. (2008). Proportion of non-zero normal means: Universal oracle equivalences and uniformly consistent estimators. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **70** 461–493. MR2420411 https://doi.org/10.1111/j.1467-9868.2007.00645.x
[48] JIN, J. and CAI, T. T. (2007). Estimating the null and the proportional of nonnull effects in large-scale multiple comparisons. *J. Amer. Statist. Assoc.* **102** 495–506. MR2325113 https://doi.org/10.1198/016214507000000167
[49] JUDITSKY, A. and NEMIROVSKI, A. (2002). On nonparametric tests of positivity/monotonicity/convexity. *Ann. Statist.* **30** 498–527. MR1902897 https://doi.org/10.1214/aos/1021379863
[50] JUREČKOVÁ, J., SEN, P. K. and PICEK, J. (2012). *Methodology in Robust and Nonparametric Statistics*. CRC Press, Boca Raton, FL. MR2963549
[51] KORN, E. L., TROENDLE, J. F., MCSHANE, L. M. and SIMON, R. (2004). Controlling the number of false discoveries: Application to high-dimensional genomic data. *J. Statist. Plann. Inference* **124** 379–398. MR2080371 https://doi.org/10.1016/S0378-3758(03)00211-8
[52] LACOUR, C. and MASSART, P. (2016). Minimal penalty for Goldenshluger–Lepski method. *Stochastic Process. Appl.* **126** 3774–3789. MR3565477 https://doi.org/10.1016/j.spa.2016.04.015

[53] LANCASTER, T. (2000). The incidental parameter problem since 1948. *J. Econometrics* **95** 391–413. MR1752336 https://doi.org/10.1016/S0304-4076(99)00044-5

[54] LEEK, J. T. and STOREY, J. D. (2008). A general framework for multiple testing dependence. *Proc. Natl. Acad. Sci. USA* **105** 18718–18723. https://doi.org/10.1073/pnas.0808709105

[55] LEPSKI, O., NEMIROVSKI, A. and SPOKOINY, V. (1999). On estimation of the L_r norm of a regression function. *Probab. Theory Related Fields* **113** 221–253. MR1670867 https://doi.org/10.1007/s004409970006

[56] LEPSKIĬ, O. V. (1990). A problem of adaptive estimation in Gaussian white noise. *Teor. Veroyatn. Primen.* **35** 459–470. MR1091202 https://doi.org/10.1137/1135065

[57] LI, A. and BARBER, R. F. (2019). Multiple testing with the structure-adaptive Benjamini–Hochberg algorithm. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **81** 45–74. MR3904779

[58] NEYMAN, J. and SCOTT, E. L. (1948). Consistent estimates based on partially consistent observations. *Econometrica* **16** 1–32. MR0025113 https://doi.org/10.2307/1914288

[59] RABINOVICH, M., RAMDAS, A., JORDAN, M. I. and WAINWRIGHT, M. J. (2020). Optimal rates and tradeoffs in multiple testing. *Statist. Sinica.* **30** 741–762. https://doi.org/10.5705/ss.202017.0468

[60] ROMANO, J. P., SHAIKH, A. M. and WOLF, M. (2008). Control of the false discovery rate under dependence using the bootstrap and subsampling. *TEST* **17** 417–442. MR2470085 https://doi.org/10.1007/s11749-008-0126-6

[61] ROMANO, J. P. and WOLF, M. (2005). Exact and approximate stepdown methods for multiple hypothesis testing. *J. Amer. Statist. Assoc.* **100** 94–108. MR2156821 https://doi.org/10.1198/016214504000000539

[62] ROMANO, J. P. and WOLF, M. (2007). Control of generalized error rates in multiple testing. *Ann. Statist.* **35** 1378–1408. MR2351090 https://doi.org/10.1214/009053606000001622

[63] ROQUAIN, E. and VAN DE WIEL, M. A. (2009). Optimal weighting for false discovery rate control. *Electron. J. Stat.* **3** 678–711. MR2521216 https://doi.org/10.1214/09-EJS430

[64] SARKAR, S. K. (2008). Rejoinder: On methods controlling the false discovery rate [MR2551809; MR2551810; MR2551811]. *Sankhyā* **70** 183–185. MR2551812

[65] SPOKOINY, V. G. (1996). Adaptive hypothesis testing using wavelets. *Ann. Statist.* **24** 2477–2498. MR1425962 https://doi.org/10.1214/aos/1032181163

[66] VERZELEN, N. (2012). Minimax risks for sparse regressions: Ultra-high dimensional phenomenons. *Electron. J. Stat.* **6** 38–90. MR2879672 https://doi.org/10.1214/12-EJS666

[67] WESTFALL, P. H. and YOUNG, S. S. (1993). *Resampling-Based Multiple Testing*: *Examples and Methods for P-Value Adjustment*. Wiley, New York.

[68] WU, Y. and YANG, P. (2019). Chebyshev polynomials, moment matching, and optimal estimation of the unseen. *Ann. Statist.* **47** 857–883. MR3909953 https://doi.org/10.1214/17-AOS1665