# Aggregation of Multiple Knockoffs

Tuan-Binh Nguyen, Jérôme-Alexis Chevalier, Bertrand Thirion, Sylvain Arlot

# Aggregation of Multiple Knockoffs

Tuan-Binh Nguyen [*1,2] Jérôme-Alexis Chevalier[2]
Bertrand Thirion[2] Sylvain Arlot[1]

[1] *Université Paris-Saclay, CNRS, Inria, Laboratoire de mathématiques d'Orsay, 91405, Orsay, France*

[2] *Inria, CEA, Université Paris-Saclay, France*

*Abstract.* We develop an extension of the knockoff inference procedure, introduced by Barber and Candès [2015]. This new method, called aggregation of multiple knockoffs (AKO), addresses the instability inherent to the random nature of knockoff-based inference. Specifically, AKO improves both the stability and power compared with the original knockoff algorithm while still maintaining guarantees for false discovery rate control. We provide a new inference procedure, prove its core properties, and demonstrate its benefits in a set of experiments on synthetic and real datasets. [†]

## 1. INTRODUCTION

In many fields, multivariate statistical models are used to *fit* some outcome of interest through a combination of measurements or features. For instance, one might predict the likelihood for individuals to declare a certain type of disease based on genotyping information. Besides prediction accuracy, the inference problem consists in defining which measurements carry useful features for prediction. More precisely, we aim at conditional inference (as opposed to marginal inference), that is, analyzing which features carry information *given* the other features. This inference is however very challenging in high-dimensional settings.

Among the few available solutions, knockoff-based (KO) inference [BC15, CFJL18] consists in introducing noisy copies of the original variables that are independent from the outcome conditional on the original variables, and comparing the coefficients of the original variables to those of the knockoff variables. This approach is particularly

[*] Corresponding email: tuan-binh.nguyen@inria.fr

[†] Code is available at: github.com/ja-che/hidimstat

attractive for several reasons: *i)* it is not tied to a given statistical model, but can work instead for many different multivariate functions, whether linear or not; *ii)* it requires a good generative model for features, but poses few conditions for the validity of inference; and *iii)* it controls the false discovery rate (FDR, [BH95]), a more useful quantity than multiplicity-corrected error rates.

Unfortunately, KO has a major drawback, related to the random nature of the knockoff variables: two different draws yield two different solutions, leading to large, uncontrolled fluctuations in power and false discovery proportion across experiments (see Figure 1 below). This makes the ensuing inference irreproducible. An obvious way to fix the problem is to rely on some type of statistical aggregation, in order to consolidate the inference results. Such procedures have been introduced by [GZ19] and by [EK19], but they have several limitations: the computational complexity scales poorly with the number $B$ of bootstraps, while the power of the method decreases with $B$. In high-dimensional settings that we target, these methods are thus only usable with a limited number of bootstraps.

In this work, we explore a different approach, that we call aggregation of multiple knockoffs (AKO): it rests on a reformulation of the original knockoff procedure that introduces intermediate p-values. As it is possible to aggregate such quantities even without assuming independence [MMB09], we propose to perform aggregation at this intermediate step. We first establish the equivalence of AKO with the original knockoff aggregation procedure in case of one bootstrap (Proposition 1). Then we show that the FDR is also controlled with AKO (Theorem 1). By construction, AKO is more stable than (vanilla) knockoff; we also demonstrate empirical benefits in several examples, using simulated data, but also genetic and brain imaging data. Note that the added knockoff generation and inference steps are embarrassingly parallel, making this procedure no more costly than the original KO inference.

**Notation.** Let $[p]$ denote the set $\{1, 2, \dots, p\}$; for a given set given set $\mathcal{A}$, $|\mathcal{A}| \triangleq \mathbf{card}(\mathcal{A})$; matrices are denoted in bold uppercase letter, while vectors in bold lowercase letter and scalars normal character. An exception for this is the vector of knockoff statistic $\mathbf{W}$, in which we follow the notation from the original paper of [BC15].

## 2. BACKGROUND

**Problem Setting.** Let $\mathbf{X} \in \mathbb{R}^{n \times p}$ be a design matrix corresponding to $n$ observations of $p$ potential explanatory variables $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n \in \mathbb{R}^p$, with its target vector $\mathbf{y} \in \mathbb{R}^n$. To simplify the exposition, we focus on sparse linear models, as [BC15] and [CFJL18]:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta}^* + \sigma\boldsymbol{\epsilon} \tag{1}$$

where $\boldsymbol{\beta}^* \in \mathbb{R}^p$ is the true parameter vector, $\sigma \in \mathbb{R}^+$ the unknown noise magnitude, $\boldsymbol{\epsilon} \in \mathbb{R}^n$ some Gaussian noise vector. Yet, it should be noted that the algorithm does not require linearity or sparsity. Our main interest is in finding an estimate $\widehat{\mathcal{S}}$ of the true support set $\mathcal{S} = \{j \in [p] : \beta^*_j \neq 0\}$, or the set of important features that have an effect on the response. As a consequence, the complementary of the support $\mathcal{S}$, which is denoted $\mathcal{S}^c = \{j \in [p] : \beta^*_j = 0\}$, corresponds to null hypotheses. Identifying the relevant features amounts to simultaneously testing

$$\mathcal{H}_0^j : \beta^*_j = 0 \quad \text{versus} \quad \mathcal{H}_a^j : \beta^*_j \neq 0, \quad \forall j = 1, \ldots, p.$$

Specifically, we want to bound the proportion of false positives among selected variables, that is, control the false discovery rate (FDR, [BH95]) under certain predefined level $\alpha$:

$$\mathrm{FDR} = \mathbb{E}\left[\frac{|\widehat{\mathcal{S}} \cap \mathcal{S}^c|}{|\widehat{\mathcal{S}}| \vee 1}\right] \leq \alpha \in (0, 1)\,.$$

**Knockoff Inference.** Introduced originally by [BC15], the knockoff filter is a variable selection method for multivariate models with theoretical control of FDR. [CFJL18] expanded the method to work in the case of (mildly) high-dimensional data, with the assumption that $\mathbf{x} = (x_1, \ldots, x_p) \sim P_X$ such that $P_X$ is known. The first step of this procedure involves sampling extra null variables that have a correlation structure similar to that of the original variables, with the following formal definition.

Definition 1 (Model-X knockoffs, [CFJL18]). The model-X knockoffs for the family of random variables $\mathbf{x} = (x_1, \ldots, x_p)$ are a new family of random variables $\tilde{\mathbf{x}} = (\tilde{x}_1, \ldots, \tilde{x}_p)$ constructed to satisfy the two properties:

1. For any subset $\mathcal{K} \subset \{1, \ldots, p\}$, $(\mathbf{x}, \tilde{\mathbf{x}})_{\text{swap}(\mathcal{K})} \overset{d}{=} (\mathbf{x}, \tilde{\mathbf{x}})$, where the vector $(\mathbf{x}, \tilde{\mathbf{x}})_{\text{swap}(\mathcal{K})}$ denotes the swap of entries $x_j$ and $\tilde{x}_j$ for all $j \in \mathcal{K}$, and $\overset{d}{=}$ denotes equality in distribution.
2. $\tilde{\mathbf{x}} \perp\!\!\!\perp \mathbf{y} \mid \mathbf{x}\,.$

A test statistic is then calculated to measure the strength of the original variables versus their knockoff counterpart. We call this the knockoff statistic $\mathbf{W} = \{W_j\}_{j=1}^p$, that must fulfill two important properties.

Definition 2 (Knockoff statistic, [CFJL18]). A knockoff statistic $\mathbf{W} = \{W_j\}_{j \in [p]}$ is a measure of feature importance that satisfies the two following properties:

1. It depends only on $\mathbf{X}, \tilde{\mathbf{X}}$ and $\mathbf{y}$

$$\mathbf{W} = f(\mathbf{X}, \tilde{\mathbf{X}}, \mathbf{y})\,.$$

2. Swapping the original variable column $\mathbf{x}_j$ and its knockoff column $\tilde{\mathbf{x}}_j$ switches the sign of $W_j$:

$$W_j([\mathbf{X}, \tilde{\mathbf{X}}]_{swap(S)}, y) = \begin{cases} W_j([\mathbf{X}, \tilde{\mathbf{X}}], \mathbf{y}) \text{ if } j \in \mathcal{S}^c \\ -W_j([\mathbf{X}, \tilde{\mathbf{X}}], y) \text{ if } j \in \mathcal{S} \,. \end{cases}$$

Following previous works on the analysis of the knockoff properties [ACC17, RRJW20], we make the following assumption about the knockoff statistic. This is necessary for our analysis of knockoff aggregation scheme later on.

Assumption 1 (Null distribution of knockoff statistic). The knockoff statistic defined in Definition 2 are such that $\{W_j\}_{j \in \mathcal{S}^c}$, are independent and follow the same distribution $\mathbb{P}_0$.

Remark 1. As a consequence of [CFJL18, Lemma 2] regarding the signs of the null $W_j$ as i.i.d. coin flips, if Assumption 1 holds true, then $\mathbb{P}_0$ is symmetric around zero.

One such example of knockoff statistic is the Lasso-coefficient difference (LCD). The LCD statistic is computed by first making the concatenation of original variable and knockoff variables $[\mathbf{X}, \tilde{\mathbf{X}}] \in \mathbb{R}^{n \times 2p}$, then solving the Lasso problem [Tib96]:

$$\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta} \in \mathbb{R}^{2p}}{\operatorname{argmin}} \left\{ \frac{1}{2} \left\| \mathbf{y} - [\mathbf{X}, \tilde{\mathbf{X}}]\boldsymbol{\beta} \right\|_2^2 + \lambda \left\| \boldsymbol{\beta} \right\|_1 \right\} \tag{2}$$

with $\lambda \in \mathbb{R}$ the regularization parameter, and finally to take:

$$\forall j \in [p] \,, \qquad W_j = |\widehat{\beta}_j| - |\widehat{\beta}_{j+p}| \,. \tag{3}$$

This quantity measures how strong the coefficient magnitude of each original covariate is against its knockoff, hence the name Lasso-coefficient difference. Clearly, the LCD statistic satisfies the two properties stated in Definition 2.

Finally, a threshold for controlling the FDR under given level $\alpha \in (0, 1)$ is calculated:

$$\tau_+ = \min \left\{ t > 0 : \frac{1 + \#\{j : W_j \leq -t\}}{\#\{j : W_j \geq t\} \vee 1} \leq \alpha \right\} , \tag{4}$$

and the set of selected variables is $\widehat{\mathcal{S}} = \{j \in [p] : W_j \geq \tau_+\}$.

**Instability in Inference Results.** Knockoff inference is a flexible method for multivariate inference in the sense that it can use different loss functions (least squares, logistic, etc.), and use different variable importance statistics. However, a major drawback of the method comes from the random nature of the knockoff variables $\tilde{\mathbf{X}}$ obtained by sampling: different draws yield different solutions (see Figure 1 in Section 5.1). This is a major issue in practical settings, where knockoff-based inference is used to prove the conditional association between features and outcome.

# 3. AGGREGATION OF MULTIPLE KNOCKOFFS

## 3.1 Algorithm Description

One of the key factors that lead to the extension of the original (vanilla) knockoff filter stems from the observation that knockoff inference can be formulated based on the following quantity.

Definition 3 (Intermediate p-value). Let $\mathbf{W} = \{W_j\}_{j\in[p]}$ be a knockoff statistic according to Definition 2. For $j = 1, \dots, p$, the intermediate p-value $\pi_j$ is defined as:

$$\pi_j = \begin{cases} \dfrac{1 + \#\{k : W_k \leq -W_j\}}{p} & \text{if} \quad W_j > 0 \\ 1 \quad \text{if} \quad W_j \leq 0\,. \end{cases} \tag{5}$$

We first compute $B$ draws of knockoff variables, and then knockoff statistics. Using Eq. (5), we derive the corresponding empirical p-values $\pi_j^{(b)}$, for all $j \in [p]$ and $b \in [B]$. Then, we aggregate them for each variable $j$ in parallel, using the quantile aggregation procedure introduced by [MMB09]:

$$\bar{\pi}_j = \min\left\{1, \frac{q_\gamma\big(\{\pi_j^{(b)} : b \in [B]\}\big)}{\gamma}\right\} \tag{6}$$

where $q_\gamma(\cdot)$ is the $\gamma$-quantile function. In the experiments, we fix $\gamma = 0.3$ and $B = 25$. The selection of these default values is explained more thoroughly in Section 5.1.

Finally, with a sequence of aggregated p-values $\bar{\pi}_1, \dots, \bar{\pi}_p$, we use Benjamini-Hochberg step-up procedure (BH, [BH95]) to control the FDR.

Definition 4 (BH step-up, [BH95]). Given a list of p-values $\bar{\pi}_1, \dots, \bar{\pi}_p$ and predefined FDR control level $\alpha \in (0, 1)$, the Benjamini-Hochberg step-up procedure comprises three steps:

1. Order p-values such that: $\bar{\pi}_{(1)} \leq \bar{\pi}_{(2)} \leq \dots \leq \bar{\pi}_{(p)}$.
2. Find:

$$\widehat{k}_{BH} = \max\left\{k : \bar{\pi}_{(k)} \leq \frac{k\alpha}{p}\right\}. \tag{7}$$

3. Select $\widehat{\mathcal{S}} = \{j \in [p] : \bar{\pi}_{(j)} \leq \bar{\pi}_{(\widehat{k}_{BH})}\}$.

This procedure controls the FDR, but only under independence or positive-dependence between p-values [BY01]. As a matter of fact, for a strong guarantee of FDR control, one can consider instead a threshold yielding a theoretical control of FDR under arbitrary dependence, such as the one of [BY01]. We call BY step-up the resulting procedure. Yet we use BH step-up procedure in the experiments of Section 5, as we observe empirically that the aggregated p-values $\bar{\pi}_j$ defined in Eq. (5) does not deviate significantly from independence (details in Appendix).

Definition 5 (BY step-up, [BY01]). Given an ordered list of p-values as in step 1 of BH step-up $\bar{\pi}_{(1)} \leq \bar{\pi}_{(2)} \leq \cdots \leq \bar{\pi}_{(p)}$ and predefined level $\alpha \in (0,1)$, the Benjamini-Yekutieli step-up procedure first finds:

$$\widehat{k}_{BY} = \max\left\{k \in [p] : \bar{\pi}_{(k)} \leq \frac{k\beta(p)\alpha}{p}\right\}, \tag{8}$$

with $\beta(p) = (\sum_{i=1}^{p} 1/i)^{-1}$, and then selects

$$\widehat{\mathcal{S}} = \left\{j \in [p] : \bar{\pi}_{(j)} \leq \bar{\pi}_{(\widehat{k}_{BY})}\right\}.$$

[BR09] later on introduced a general function form for $\beta(p)$ to make BY step-up more flexible. However, because we always have $\beta(p) \leq 1$, this procedure leads to a smaller threshold than BH step-up, thus being more conservative.

**Algorithm 1** AKO – Aggregation of multiple knockoffs

```
Input: $\mathbf{X} \in \mathbb{R}^{n\times p}$, $\mathbf{y} \in \mathbb{R}^n$, $B$ – number of bootstraps ; $\alpha \in (0,1)$ – target FDR level
Output: $\widehat{\mathcal{S}}_{AKO}$ – Set of selected variables index
for $b = 1$ to $B$ do
    $\tilde{\mathbf{X}}^{(b)} \leftarrow$ SAMPLING_KNOCKOFF($\mathbf{X}$)
    $\mathbf{W}^{(b)} \leftarrow$ KNOCKOFF_STATISTIC($\mathbf{X}, \tilde{\mathbf{X}}^{(b)}, \mathbf{y}$)
    $\boldsymbol{\pi}^{(b)} \leftarrow$ CONVERT_STATISTIC($\mathbf{W}^{(b)}$) // Using Eq. (5)
end for
for $j = 1$ to $p$ do
    $\bar{\pi}_j \leftarrow$ QUANTILE_AGGREGATION($\left(\{\pi_j^{(b)}\}_{b=1}^{B}\right)$) // Using Eq. (6)
end for
$\widehat{k} \leftarrow$ FDR_THRESHOLD($\alpha, (\bar{\pi}_1, \bar{\pi}_2, \ldots, \bar{\pi}_p)$) // Using either Eq. (7) or Eq. (8)
Return: $\widehat{\mathcal{S}}_{AKO} \leftarrow \left\{j \in [p] : \bar{\pi}_j \leq \bar{\pi}_{\widehat{k}}\right\}$
```

The AKO procedure is summarized in Algorithm 1. We show in the next section that with the introduction of the aggregation step, the procedure offers a guarantee on FDR control under mild hypotheses. Additionally, the numerical experiments of Section 5 illustrate that aggregation of multiple knockoffs indeed improves the stability of the knockoff filter, while bringing significant statistical power gains.

### 3.2 Related Work

To our knowledge, up until now there have been few attempts to stabilize knockoff inference. Earlier work of [SQL15] rests on the same idea of generating multiple knockoff bootstrap as ours, but relies on the linear combination of the so-called *one-bit p-values* (introduced as a means to prove the FDR control in original knockoff work of [BC15]). As such, the method is less flexible since it requires a specific type of knockoff statistic to work. Furthermore, it is unclear how this method would perform in high-dimensional settings, as it was only demonstrated in the case of $n > p$. More recently, the work of [HH18] incorporates directly multiple bootstraps of knockoff statistics for FDR thresholding without the need of p-value conversion. Despite its simplicity and convenience as a way of aggregating knockoffs, our simulation study in Section 5.1 demonstrates that this method somehow fails to control FDR in several settings.

In a different direction, [GZ19] and [EK19] have introduced *simultaneous knockoff* procedure, with the idea of sampling several knockoff copies at the same time instead of doing the process in parallel as in our work. This, however, induces a prohibitive computational cost when the number of bootstraps increases, as opposed to the AKO algorithm that can use parallel computing to sample multiple bootstraps at the same time. In theory, on top of the fact that sampling knockoffs has cubic complexity on runtime with regards to number of variables $p$ (requires covariance matrix inversion), simultaneous knockoff runtime is of $\mathcal{O}(B^3p^3)$, while for AKO, runtime is only of $\mathcal{O}(Bp^3)$ and $\mathcal{O}(p^3)$ with parallel computing. Moreover, the FDR threshold of simultaneous knockoff is calculated in such a way that it loses statistical power as the number of bootstraps increases, when the sampling scheme of vanilla knockoff by [BC15] is used. We have set up additional experiments in the Appendix to illustrate this phenomenon. In addition, the threshold introduced by [EK19] is only proven to have a theoretical control of FDR in the case where $n > p$.

## 4. THEORETICAL RESULTS

We now state our theoretical results about the AKO procedure.

### 4.1 Equivalence of Aggregated Knockoff with Single Bootstrap

First, when $B = 1$ and $\gamma = 1$, we show that AKO+BH is equivalent to vanilla knockoff.

Proposition 1 (Proof in Appendix A.1). Assume that for all $j, j' = 1, \dots, p$,

$$\mathbb{P}(W_j = W_{j'}, \quad W_j \neq 0, \quad W_{j'} \neq 0) = 0$$

that is, non-zero LCD statistics are distinct with probability 1. Then, single bootstrap version of aggregation of multiple knockoffs ($B = 1$), using $\gamma = 1$ and BH step-up procedure in Definition 4 for calculating FDR threshold, is equivalent to the original knockoff inference by [BC15].

Remark 2. Although Proposition 1 relies on the assumption of distinction between non-zero $W_j$s for all $j = 1, \dots, p$, the following lemma establishes that this assumption holds true with probability one for the LCD statistic up to further assumptions.

Lemma 1 (Proof in Appendix A.2). Define the equi-correlation set as:

$$\widehat{J}_\lambda = \left\{ j \in [p] : \mathbf{x}_j^\top (\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}) = \lambda/2 \right\}$$

with $\widehat{\boldsymbol{\beta}}, \lambda$ defined in Eq. (2). Then we have:

$$\mathbb{P}\left( W_j = W_{j'}, W_j \neq 0, W_{j'} \neq 0,\ \operatorname{rank}(X_{\widehat{J}_\lambda}) = |\widehat{J}_\lambda| \right) = 0 \tag{9}$$

for all $j, j' \in [p] : j \neq j'$. In other words, assuming $\mathbf{X}_{\widehat{J}_\lambda}$ is full rank, then the event that LCD statistic defined in Eq. (3) is distinct for all non-zero value happens almost surely.

### 4.2 Validity of Intermediate P-values

Second, the fact that the $\pi_j$ are called "intermediate p-values" is justified by the following lemma.

Lemma 2. If Assumption 1 holds true, and if $|\mathcal{S}^c| \geq 2$, then, for all $j \in \mathcal{S}^c$, the intermediate p-value $\pi_j$ defined by Eq. (5) satisfies:

$$\forall t \in [0, 1] \qquad \mathbb{P}(\pi_j \leq t) \leq \frac{\kappa p}{|\mathcal{S}^c|} t$$

where $\kappa = \dfrac{\sqrt{22} - 2}{7\sqrt{22} - 32} \leq 3.24$.

Proof. The result holds when $t \geq 1$ since $\kappa p \geq p \geq |\mathcal{S}^c|$ and a probability is always smaller than 1. Let us now focus on the case where $t \in [0, 1)$, and define $m = |\mathcal{S}^c| - 1 \geq 1$ by assumption. Let $F_0$ denote the c.d.f. of $\mathbb{P}_0$, the common distribution of the null statistics $\{W_k\}_{k \in \mathcal{S}^c}$, which exists by Assumption 1. Let $j \in \mathcal{S}^c$ be fixed. By definition of $\pi_j$, when $W_j > 0$ we have:

$$\begin{aligned}
\pi_j &= \frac{1 + \#\{k \in [p] : W_k \leq -W_j\}}{p} \\
&= \frac{1 + \#\{k \in \mathcal{S} : W_k \leq -W_j\}}{p} + \frac{\#\{k \in \mathcal{S}^c \setminus \{j\} : W_k \leq -W_j\}}{p} \\
&\quad \text{(since } W_j > 0 > -W_j\text{)} \\
&\geq \frac{m}{p}\widehat{F}_m(-W_j) + \frac{1}{p} \qquad (10)
\end{aligned}$$

where $\forall u \in \mathbb{R}$, $\widehat{F}_m(u) \triangleq \frac{\#\{k \in \mathcal{S}^c \setminus \{j\} : W_k \leq u\}}{m}$ is the empirical cdf of $\{W_k\}_{k \in \setminus \{j\}}$. Therefore, for every $t \in [0, 1)$,

$$\begin{aligned}
\mathbb{P}(\pi_j \leq t) &= \mathbb{P}(\pi_j \leq t \text{ and } W_j > 0) + \underbrace{\mathbb{P}(\pi_j \leq t \text{ and } W_j \leq 0)}_{=0 \text{ since } \pi_j = 1 \text{ when } W_j \leq 0} \\
&= \mathbb{E}\big[\mathbb{P}(\pi_j \leq t \mid W_j)\mathbb{1}_{W_j > 0}\big] \\
&\leq \mathbb{E}\left[\mathbb{P}\left(\frac{m}{p}\widehat{F}_m(-W_j) + \frac{1}{p} \leq t \mid W_j\right)\mathbb{1}_{W_j > 0}\right] \text{ by (10)} \\
&\leq \mathbb{P}\left(\frac{m}{p}\widehat{F}_m(-W_j) + \frac{1}{p} \leq t\right) . \qquad (11)
\end{aligned}$$

Notice that $W_j$ has a symmetric distribution around 0, as shown by Remark 1, that is, $-W_j$ and $W_j$ have the same distribution. Since $W_j$ and $\{W_k\}_{k \in \mathcal{S}^c \setminus \{j\}}$ are independent with the same distribution $\mathbb{P}_0$ by Assumption 1, they have the same joint distribution as $F_0^{-1}(U), F_0^{-1}(U_1), \ldots, F_0^{-1}(U_m)$ where $U, U_1, \ldots, U_m$ are independent random variables with uniform distribution over $[0, 1]$, and $F_0^{-1}$ denotes the generalized inverse of $F_0$. Therefore, Eq. (11) can be rewritten as

$$\mathbb{P}(\pi_j \leq t) \leq \mathbb{P}\left(\frac{m}{p}\widetilde{F}_m(F_0^{-1}(U)) + \frac{1}{p} \leq t\right) \qquad (12)$$

$$\text{where} \qquad \forall v \in \mathbb{R}, \qquad \widetilde{F}_m(v) \triangleq \frac{1}{m}\sum_{k=1}^{m} \mathbb{1}_{F_0^{-1}(U_k) \leq v} .$$

Notice that for every $u \in \mathbb{R}$,

$$\begin{aligned}
\widehat{G}_m(u) &\triangleq \frac{1}{m}\sum_{k=1}^{m} \mathbb{1}_{U_k \leq u} \leq \frac{1}{m}\sum_{k=1}^{m} \mathbb{1}_{F_0^{-1}(U_k) \leq F_0^{-1}(u)} \\
&= \widetilde{F}_m(F_0^{-1}(u))
\end{aligned}$$

since $F_0^{-1}$ is non-decreasing. Therefore, Eq. (12) shows that

$$\begin{aligned} \mathbb{P}(\pi_j \le t) &\le \mathbb{P}\left(m\widehat{G}_m(U) \le tp-1\right) \\ &= \int_0^1 \mathbb{P}\left(m\widehat{G}_m(u) \le tp-1\right) \mathrm{d}u\,. \end{aligned} \tag{13}$$

Now, we notice that for every $u \in (0,1)$, $m\widehat{G}_m(u)$ follows a binomial distribution with parameters $(m,u)$. So, a standard application of Bernstein's inequality [BLM13, Eq. 2.10] shows that for every $0 \le x \le u \le 1$,

$$\begin{aligned} \mathbb{P}\left(m\widehat{G}_m(u) \le mx\right) &\le \exp\left(\frac{-m^2(u-x)^2}{2mu + \frac{m(u-x)}{3}}\right) \\ &= \exp\left(\frac{-3mx\left(\frac{u}{x}-1\right)^2}{\frac{7u}{x}-1}\right). \end{aligned}$$

Note that for every $\lambda \in (0,1/7)$, we have

$$\forall w \ge \frac{1-\lambda}{1-7\lambda} \ge 1\,, \qquad \frac{w-1}{7w-1} \ge \lambda$$

hence $\forall u \ge x\dfrac{1-\lambda}{1-7\lambda}$,

$$\mathbb{P}\left(m\widehat{G}_m(u) \le mx\right) \le \exp\left[-3m\lambda x\left(\frac{u}{x}-1\right)\right].$$

As a consequence, $\forall \lambda \in (0,1/7)$,

$$\begin{aligned} \int_0^1 \mathbb{P}\left(m\widehat{G}_m(u) \le mx\right) \mathrm{d}u &\le \frac{1-\lambda}{1-7\lambda}x + \int_{\frac{1-\lambda}{1-7\lambda}x}^{1} \exp[-3m\lambda(u-x)]\mathrm{d}u \\ &\le \frac{1-\lambda}{1-7\lambda}x + \int_{\frac{6\lambda}{1-7\lambda}x}^{+\infty} \exp(-3m\lambda v)\mathrm{d}v \\ &\le \frac{1-\lambda}{1-7\lambda}x + \frac{1}{3m\lambda}\exp\left(-3m\lambda\frac{6\lambda}{1-7\lambda}x\right) \\ &\le \frac{1-\lambda}{1-7\lambda}x + \frac{1}{3m\lambda}. \end{aligned}$$

Taking $x = (tp-1)/m$, we obtain from Eq. (13) that $\forall \lambda \in (0,1/7)$

$$\begin{aligned} \mathbb{P}(\pi_j \le t) &\le \frac{1-\lambda}{1-7\lambda}\frac{tp-1}{m} + \frac{1}{3m\lambda} \\ &= \frac{1-\lambda}{1-7\lambda}\frac{tp}{m} + \left(\frac{1}{3\lambda} - \frac{1-\lambda}{1-7\lambda}\right)\frac{1}{m}\,. \end{aligned} \tag{14}$$

Choosing $\lambda = (5-\sqrt{22})/3 \in (0,1/7)$, we have $\frac{1}{3\lambda} = \frac{1-\lambda}{1-7\lambda}$ hence the result with

$$\kappa = \frac{1-\lambda}{1-7\lambda} = \frac{\sqrt{22}-2}{7\sqrt{22}-32} \leq 3.24.$$

□

Remark 3. If the definition of $\pi_j$ is replaced by

$$\pi_{j,c} \triangleq \begin{cases} \dfrac{c + \#\{k : W_k \leq -W_j\}}{p} & \text{if} \quad W_j > 0 \\ 1 \quad \text{if} \quad W_j \leq 0 \end{cases} \tag{15}$$

for some $c > 0$, the above proof also applies and yields an upper bound of the form

$$\forall t \geq 0\,, \qquad \mathbb{P}(\pi_{j,c} \leq t) \leq \kappa(c)t$$

for some constant $\kappa(c) > 0$. It is then possible to make $\kappa(c)$ as close to 1 as desired, by choosing $c$ large enough. Lemma 2 corresponds to the case $c = 1$.

Note that we also prove in the Appendix that if $p \to +\infty$ with $|\mathcal{S}| \ll p$, then for every $j \geq 1$ such that $\beta^*_j = 0$, $\pi_j$ is an asymptotically valid p-value, that is,

$$\forall t \in [0,1]\,, \qquad \limsup_{p\to+\infty} \mathbb{P}(\pi_j \leq t) \leq t\,. \tag{16}$$

Yet, proving our main result (Theorem 1) requires a non-asymptotic bound such that the one of Lemma 2.

### 4.3 FDR control for AKO

Finally, the following theorem provides a non-asymptotic guarantee about the FDR of AKO with BY step-up.

Theorem 1. If Assumption 1 holds true and $|\mathcal{S}^c| \geq 2$, then for any $B \geq 1$ and $\alpha \in (0,1)$, the output $\widehat{\mathcal{S}}_{AKO+BY}$ of aggregation of multiple knockoff (Algorithm 1), with the BY step-up procedure, has a FDR controlled as follows:

$$\mathbb{E}\left[\frac{|\widehat{\mathcal{S}}_{AKO+BY} \cap\ \mathcal{S}^c|}{|\widehat{\mathcal{S}}_{AKO+BY}| \vee 1}\right] \leq \kappa\alpha$$

where $\kappa \leq 3.24$ is defined in Lemma 2.

Sketch of the proof. The proof of [MMB09, Theorem 3.3], which itself relies partly on [BY01], can directly be adapted to upper bound the FDR of $\widehat{S}_{AKO+BY}$ in terms of quantities of the form $\mathbb{P}(\pi_j^{(b)} \leq t)$ for $j \in \mathcal{S}^c$ and several $t \geq 0$. Combined with Lemma 2, this yields the result. A full proof is provided in Appendix A.5. □

Note that Theorem 1 loses a factor $\kappa$ compared to the nominal FDR level $\alpha$. This can be solved by changing $\alpha$ into $\alpha/\kappa$ in the definition of $\widehat{S}_{AKO+BY}$. Nevertheless, in our experiments, we do not use this correction and find that the FDR is still controlled at level $\alpha$.

## 5. EXPERIMENTS

**Compared Methods.** We make benchmarks of our proposed method aggregation of multiple knockoffs (AKO) with $B = 25, \gamma = 0.3$ and vanilla knockoff (KO), along with other recent methods for controlling FDR in high-dimensional settings, mentioned in Section 3.2: *simultaneous knockoff*, an alternative aggregation scheme for knockoff inference introduced by [GZ19] (KO-GZ), along with its variant of [EK19] (KO-EK); the *knockoff statistics aggregation* by [HH18] (KO-HH); and *debiased Lasso* (DL-BH) [JJ19].

### 5.1 Synthetic Data

**Simulation Setup.** Our first experiment is a simulation scenario where a design matrix $\mathbf{X}$ $(n = 500, p = 1000)$ with its continuous response vector $\mathbf{y}$ are created following a linear model assumption. The matrix is sampled from a multivariate normal distribution of zero mean and covariance matrix $\mathbf{\Sigma} \in \mathbb{R}^{p\times p}$. We generate $\mathbf{\Sigma}$ as a symmetric Toeplitz matrix that has the structure:

$$\mathbf{\Sigma} = \begin{bmatrix} \rho^0 & \rho^1 & \dots & \rho^{p-1} \\ \rho^1 & \ddots & \dots & \rho^{p-2} \\ \vdots & \dots & \ddots & \vdots \\ \rho^{p-1} & \rho^{p-2} & \dots & \rho^0 \end{bmatrix}$$

where the $\rho \in (0, 1)$ parameter controls the correlation structure of the design matrix. This means that neighboring variables are strongly correlated to each other, and the correlation decreases with the distance between indices. The true regression coefficient $\boldsymbol{\beta}^*$ vector is picked with a sparsity parameter that controls the proportion of non-zero elements with amplitude 1. The noise $\boldsymbol{\epsilon}$ is generated to follow $\mathcal{N}(\boldsymbol{\mu}, \mathbf{I}_n)$ with its magnitude $\sigma = \|\mathbf{X}\boldsymbol{\beta}^*\|_2 / (\text{SNR} \, \|\boldsymbol{\epsilon}\|_2)$ controlled by the SNR parameter. The response vector $\mathbf{y}$ is then sampled according to Eq. (1). In short, the three main parameters controlling this simulation are correlation $\rho$, sparsity degree $k$ and signal-to-noise ratio SNR.

**Aggregation Helps Stabilizing Vanilla Knockoff.** To demonstrate the improvement in stability of the aggregated knockoffs, we first do multiple runs of AKO and KO with $\alpha = 0.05$ under *one simulation* of $\mathbf{X}$ and $\mathbf{y}$. In order to guarantee a fair comparison, we compare 100 runs of AKO, each with $B = 25$ bootstraps, with the corresponding 2500 runs of KO. We then plot the histogram of FDP and power in Figure 1. For the original knockoff, the false discovery proportion varies widely and has a small proportion of FDP above $0.2 = 4\alpha$. Besides, a fair amount of KO runs returns null power.

On the other hand, AKO not only improves the stability in the result for FDP —the FDR being controlled at the nominal level $\alpha = 0.05$— but it also improves statistical power: in particular, it avoids catastrophic behavior (zero power) encountered with KO.

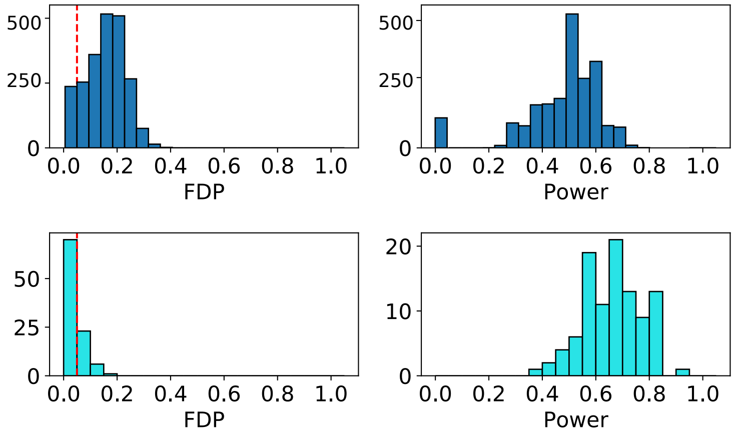

Figure 1: **Histogram of FDP and power for 2500 runs of KO (blue, top row) vs. 100 runs of AKO with B = 25 (teal, bottom row) under <u>the same simulation</u>**. Simulation parameter: SNR $= 3.0, \rho = 0.5,$ sparsity $= 0.06$. FDR is controlled at level $\alpha = 0.05$.

**Inference Results on Different Simulation Settings.** To observe how each algorithm performs under various scenarii, we vary each of the three simulation parameters while keeping the others unchanged at default value. The result is shown in Figure 2. Compared with KO, AKO improves statistical power while still controlling the FDR. Noticeably, in the case of very high correlation between nearby variables ($\rho > 0.7$), KO suffers from a drop in average power. The loss also occurs, but is less severe for AKO.

Moreover, compared with simultaneous knockoff (KO-GZ), AKO gets better control for FDR and a higher average power in the extreme correlation (high $\rho$) case. Knockoff statistics aggregation (KO-HH), contrarily, is spurious: it detects numerous truly significant variables with high average statistical power, but at a cost of failure in FDR control, especially when the correlation parameter $\rho$ gets bigger than 0.6. Debiased Lasso (DL-BH) and KO-EK control FDR well in all scenarii, but are the two most conservative procedures.

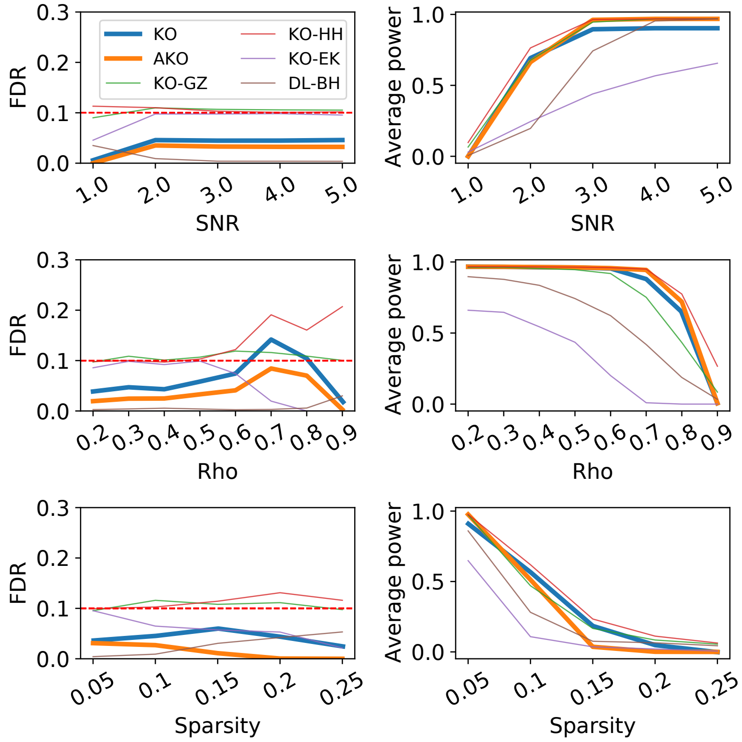

Figure 2: **FDR (left) and average power (right) of several methods for 100 runs with varying simulation parameters**. For each varying parameter, we keep the other ones at default value: SNR $= 3.0, \rho = 0.5,$ sparsity $= 0.06$. FDR is controlled at level $\alpha = 0.1$. The benchmarked methods are: aggregation of multiple knockoffs (AKO – ours); vanilla knockoff (KO); simultaneous knockoff by [GZ19] (KO-GZ) and by [EK19] (KO-EK); knockoff statistics aggregation (KO-HH); debiased-Lasso (DL-BH).

**Choice of B and $\gamma$ for AKO.** Figure 3 shows an experiment when varying $\gamma$ and $B$. FDR and power are averaged across 30 simulations of fixed parameters: SNR=3.0, $\rho = 0.7$, sparsity=0.06. Notably, it seems that there is no further gain in statistical

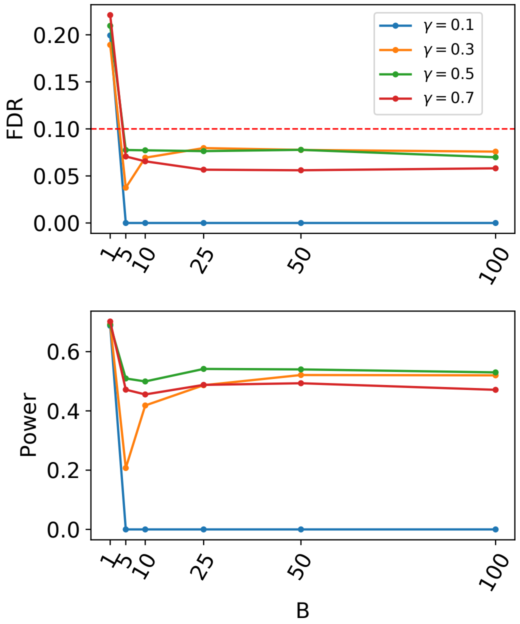

Figure 3: **FDR and average power for 30 simulations of fixed parameters: SNR=3.0, $\rho = 0.7$, sparsity=0.06.** There is virtually no gain in statistical power when $B > 25$ and when $\gamma \geq 0.1$.

power when $B > 25$. Similarly, the power is essentially equal for $\gamma$ values greater than 0.1 when $B \geq 25$. Based on the results of this experiment we set the default value of $B = 25, \gamma = 0.3$.

### 5.2 GWAS on Flowering Phenotype of *Arabidopsis thaliana*

To test AKO on real datasets, we first perform a genome-wide association study (GWAS) on genomic data. The aim is to detect association of each of 174 candidate genes with a phenotype **FT_GH** that describes flowering time of *Arabidopsis thaliana*, first done by [AHV+10]. Preprocessing is done similarly to [AGS+13]: 166 data samples of 9938 binary SNPs located within a ±20−kilobase window of 174 candidate genes that have been selected in previous publications as most likely to be involved in flowering time traits. Furthermore, we apply the same dimension reduction by hierarchical clustering as [SCAV19] to make the final design matrix of size $n = 166$ samples $\times$ $p = 1560$ features. We list the detected genes from each method in Table 1.

The three methods that rely on sampling knockoff variables detect AT2G21070. This gene, which is responsible for the mutant FIONA1, is listed by [KKY+08] to be vital

Table 1
***List of detected genes associated with phenotype FT_GH****. Empty line (—) signifies no detection. Detected genes are listed in well-known studies dated up to 20 years ago.*

| Method | Detected Genes |
|---|---|
| AKO | AT2G21070, AT4G02780, AT5G47640 |
| KO | AT2G21070 |
| KO-GZ | AT2G21070 |
| DL-BH | — |

for regulating period length in the *Arabidopsis* circadian clock. FIONA1 also appears to be involved in photoperiod-dependent flowering and in daylength-dependent seedling growth. In particular, the time for opening of the first flower for FIONA1 mutants are shorter than the ones without under both long and short-day conditions. In addition to FIONA1 mutant, AKO also detects AT4G02780 and AT5G47640. It can be found in studies dating back to the 90s [SCS98] that AT4G02780 encodes a mutation for late flowering. Meanwhile, AT5G47640 mutant delay flowering in long-day but not in short-day experiments [CBE+07].

### 5.3 Functional Magnetic Resonance Imaging (fMRI) analysis on Human Connectome Project Dataset

Human Connectome Project (HCP900) is a collection of neuroimaging and behavioral data on 900 healthy young adults, aged 22–35. Participants were asked to perform different tasks inside an MRI scanner while blood oxygenation level dependent (BOLD) signals of the brain were recorded. The analysis investigates what brain regions are predictive of the subtle variations of cognitive activity across participants, conditional to other brain regions. Similar to genomics data, the setting is high-dimensional with $n = 1556$ samples acquired and 156437 brain voxels. A voxel clustering step that reduces data dimension to $p = 1000$ clusters is done to make the problem tractable.

When decoding brain signals on HCP subjects performing a foot motion experiment (Figure 4, left), AKO recovers an anatomically correct anti-symmetric solution, in the motor cortex and the cerebellum, together with a region in a secondary sensory cortex. KO only detects a subset of those. Moreover, across seven such tasks, the results obtained independently from DL-BH are much more similar to AKO than to KO, as measured with Jaccard index of the resulting maps (Figure 4, right). The maps for the seven tasks are represented in the Appendix. Note that the sign of the effect for significant regions is readily obtained from the regression coefficients, with a voting step for bootstrap-based procedures.

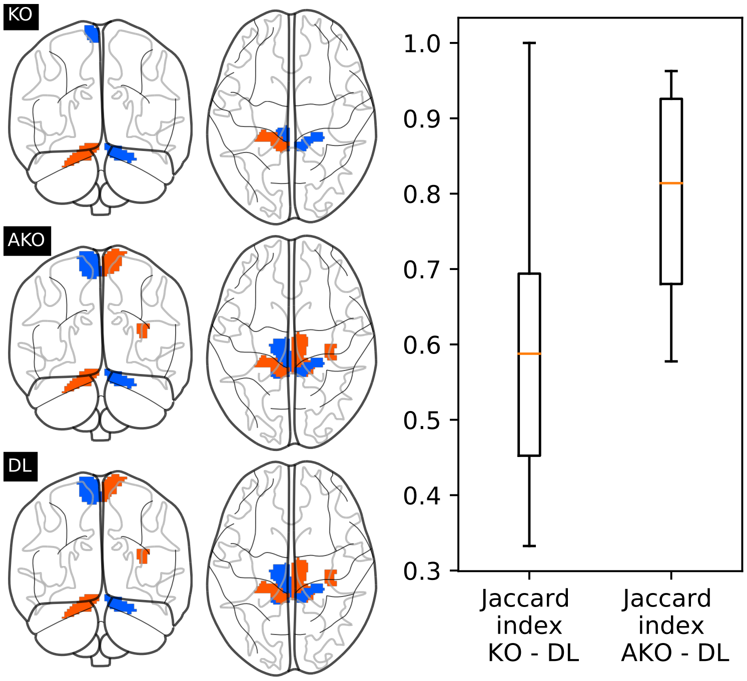

Figure 4: **Detection of significant brain regions for HCP data (900 subjects).** (left) Selected regions in a left or right foot movement task. **Orange**: brain areas with positive sign activation. **Blue**: brain areas with negative sign activation. Here the AKO solution recovers an anatomically correct pattern, part of which is missed by KO. (right) Jaccard index measuring the Jaccard similarity between the KO/AKO solutions on the one hand, and the DL solution on the other hand, over 7 tasks: AKO is significantly more consistent with the DL-BH solution than KO.

## 6. DISCUSSION

In this work, we introduce a p-value to measure knockoff importance and design a knockoffs bootstrapping scheme that leverages this quantity. With this we are able to tame the instability inherent to the original knockoff procedure. Analysis shows that aggregation of multiple knockoffs retains theoretical guarantees for FDR control. However, *i)* the original argument of [BC15] no longer holds (see Appendix); *ii)* a factor $\kappa$ on the FDR control is lost; this calls for tighter FDR bounds in the future, since we always observe empirically that the FDR is controlled without the factor $\kappa$. Moreover, both numerical and realistic experiments show that performing aggregation results in an increase in statistical power and also more consistent results with respect to alternative inference methods.

The quantile aggregation procedure from [MMB09] used here is actually conservative: as one can see in Figure 2, the control of FDR is actually stricter than without the aggregation step. Nevertheless, as often with aggregation-based approaches, the gain in accuracy brought by the reduction of estimator variance ultimately brings more power.

We would like to address here two potential concerns about FDR control for AKO+BH. The first one is when the $\{W_j\}_{j \in \mathcal{S}^c}$ are not independent, hence violating Assumption 1. In the absence of a proof of Theorem 1 that would hold under a general dependency, we first note that several schemes for knockoff construction (for instance, the one of [CFJL18]) imply the independence of $(\mathbf{x}_i - \tilde{\mathbf{x}}_i)_{i \in [p]}$, as well as their pseudo inverse. These observations do not establish the independence of $W_j$. Yet, intuitively, the Lasso coefficient of one variable should be much more associated with its knockoff version than with other variables, so it should not be much affected by these other variables, making the Lasso-coefficient differences weakly correlated if not independent. Moreover, in the proof of Lemma 2 and Theorem 1, Assumption 1 is only used for applying Bernstein's inequality, and several dependent versions of Bernstein's inequality have been proved [Sam00, MPR09, HS17, among others]. Similarly, the proof of Eq. (16) only uses Assumption 1 for applying the strong law of large numbers, a result which holds true for various kinds of dependent variables (for instance, [Abd18], and references therein). Therefore we conjecture that independence in Assumption 1 can be relaxed into some mixing condition. Overall, given that the unstability of KO with respect to the KO randomness is an important drawback (see Figure 1), we consider Assumption 1 as a reasonable price price to pay for correcting it, given that we expect to relax it in future works.

The second potential concern is that Theorem 1 is for AKO with $\widehat{k}$ computed from the BY procedure, while BH step-up may not control the FDR when the aggregated p-values $(\bar{\pi}_j)_{j \in [p]}$ are not independent. We find empirically that the $(\bar{\pi}_j)_{j \in [p]}$ do not exhibit spurious Spearman correlation (Figure B.6 in Appendix) under a setting where the $W_j$ satisfy a mixing condition. This is a mild assumption that should be satisfied, especially when each feature $X_j$ only depends on its "neighbors" (as typically observed on neuroimaging and genomics data). It is actually likely that the aggregation step contributes to reducing the statistical dependencies between the $(\bar{\pi}_j)_{j \in [p]}$. Eventually, it should be noted that BH can be replaced by BY [BY01] in case of doubt.

To conclude on these two potential concerns, let us emphasize that the FDR of AKO+BH with $B > 1$ is always below $\alpha$ (up to error bars) in *all* the experiments we did, including preliminary experiments not shown in this article, which makes us confident when applying AKO+BH on real data such as the ones of Sections 5.2–5.3.

A practical question of interest is to handle the cases where $n \ll p$, that is, the number of features overwhelms the number of samples. Note that in our experiments, we had to resort to a clustering scheme of the brain data and to select some genes. A possible extension is to couple this step with the inference framework, in order to take into account that for instance the clustering used is not given but *estimated* from the data, hence with some level of uncertainty.

The proposed approach introduces two parameters: the number $B$ of bootstrap replications and the $\gamma$ parameter for quantile aggregation. The choice of $B$ is simply driven by a compromise between accuracy (the larger $B$, the better) and computation power, but we consider that much of the benefit of AKO is obtained for $B \approx 25$. Regarding $\gamma$, adaptive solutions have been proposed [MMB09], but we find that choosing a fixed quantile (0.3) yields a good behavior, with little variance and a good sensitivity.

## ACKNOWLEDGEMENTS

The authors thank anonymous reviewers for their helpful comments and suggestions. We are grateful for the help of Lotfi Slim and Chloé-Agathe Azencott on the *Arabidopsis thaliana* dataset, and the people of Human Connectome Project on HCP900 dataset.

This research is supported under funding of French ANR project FastBig (ANR-17-CE23-0011), the KARAIB AI chair and Labex DigiCosme (ANR-11-LABEX-0045-DIGICOSME).

## REFERENCES

[Abd18] A. Abdesselam. The weakly dependent strong law of large numbers revisited. *arXiv e-prints*, page arXiv:1801.09265, January 2018.

[ACC17] E. Arias-Castro and S. Chen. Distribution-free multiple testing. *Electron. J. Statist.*, 11(1):1983–2001, 2017.

[AGS$^+$13] C.-A. Azencott, D. Grimm, M. Sugiyama, Y. Kawahara, and K. M. Borgwardt. Efficient network-guided multi-locus association mapping with graph cuts. *Bioinformatics*, 29(13):i171–i179, 2013.

[AHV$^+$10] S. Atwell, Y. S. Huang, B. J. Vilhjálmsson, G. Willems, M. Horton, Y. Li, D. Meng, A. Platt, A. M. Tarone, T. T. Hu, et al. Genome-wide association study of 107 phenotypes in Arabidopsis thaliana inbred lines. *Nature*, 465(7298):627, 2010.

[BC15] R. F. Barber and E. J. Candès. Controlling the false discovery rate via knockoffs. *The Annals of Statistics*, 43(5):2055–2085, October 2015. arXiv: 1404.5609.

[BH95] Y. Benjamini and Y. Hochberg. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300, 1995.

[BLM13] S. Boucheron, G. Lugosi, and P. Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence.* Oxford University Press, Oxford, 2013.

[BR09] G. Blanchard and É. Roquain. Adaptive false discovery rate control under independence and dependence. *Journal of Machine Learning Research*, 10(Dec):2837–2871, 2009.

[BY01] Y. Benjamini and D. Yekutieli. The control of the false discovery rate in multiple testing under dependency. *Ann. Statist.*, 29(4):1165–1188, 08 2001.

[CBE$^+$07] X. Cai, J. Ballif, S. Endo, E. Davis, M. Liang, D. Chen, D. DeWald, J. Kreps, T. Zhu, and Y. Wu. A Putative CCAAT-Binding Transcription Factor Is a Regulator of Flowering Timing in Arabidopsis. *Plant Physiology*, 145(1):98–105, 2007.

[CFJL18] E. Candès, Y. Fan, L. Janson, and J. Lv. Panning for gold: 'model-X' knockoffs for high dimensional controlled variable selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(3):551–577, 2018.

[EK19] K. Emery and U. Keich. Controlling the FDR in variable selection via multiple knockoffs. *arXiv e-prints*, page arXiv:1911.09442, November 2019.

[Gir14] C. Giraud. *Introduction to high-dimensional statistics.* Chapman and Hall/CRC, 2014.

[GZ19] J. R. Gimenez and J. Zou. Improving the Stability of the Knockoff Procedure: Multiple Simultaneous Knockoffs and Entropy Maximization. In K. Chaudhuri and M. Sugiyama, editors, *Proceedings of Machine Learning Research*, volume 89 of *Proceedings of Machine Learning Research*, pages 2184–2192. PMLR, 16–18 Apr 2019.

[HH18] L. Holden and K. H. Helton. Multiple Model-Free Knockoffs. *arXiv preprint arXiv:1812.04928*, 2018.

[HS17] H. Hang and I. Steinwart. A Bernstein-type inequality for some mixing processes and dynamical systems with an application to learning. *Ann. Statist.*, 45(2):708–743, 2017.

[JJ19] A. Javanmard and H. Javadi. False discovery rate control via debiased lasso. *Electron. J. Statist.*, 13(1):1212–1253, 2019.

[KKY$^+$08] J. Kim, Y. Kim, M. Yeom, J.-H. Kim, and H. G. Nam. FIONA1 Is Essential for Regulating Period Length in the Arabidopsis Circadian Clock. *The Plant Cell*, 20(2):307–319, 2008.

[MMB09] N. Meinshausen, L. Meier, and P. Bühlmann. p-Values for High-Dimensional Regression. *Journal of the American Statistical Association*, 104(488):1671–1681, 2009.

[MPR09] F. Merlevède, M. Peligrad, and E. Rio. Bernstein inequality and moderate deviations under strong mixing conditions. In *High dimensional probability V: the Luminy volume*, volume 5 of *Inst. Math. Stat. (IMS) Collect.*, pages 273–292. Inst. Math. Statist., Beachwood, OH, 2009.

[RD19] J. P. Romano and C. DiCiccio. Multiple Data Splitting for Testing. Technical Report Technical Report 2019-03, Stanford

University, Department of Statistics, April 2019. available at https://statistics.stanford.edu/research/multiple-data-splitting-testing.

[RRJW20] M. Rabinovich, A. Ramdas, M. I. Jordan, and M. J. Wainwright. Optimal Rates and Tradeoffs in Multiple Testing. *Statistica Sinica*, 2020.

[Sam00] P.-M. Samson. Concentration of measure inequalities for Markov chains and Φ-mixing processes. *Ann. Probab.*, 28(1):416–461, 2000.

[SCAV19] L. Slim, C. Chatelain, C.-A. Azencott, and J.-P. Vert. kernelPSI: a Post-Selection Inference Framework for Nonlinear Variable Selection. In *International Conference on Machine Learning*, pages 5857–5865, 2019.

[SCS98] A. L. Silverstone, C. N. Ciampaglio, and T.-p. Sun. The Arabidopsis RGA Gene Encodes a Transcriptional Regulator Repressing the Gibberellin Signal Transduction Pathway. *The Plant Cell*, 10(2):155–169, 1998.

[SQL15] W. Su, J. Qian, and L. Liu. Communication-efficient false discovery rate control via knockoff aggregation. *arXiv preprint arXiv:1506.05446*, 2015.

[Tib96] R. Tibshirani. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58, 1996.

# APPENDIX

The Appendix is organized as follows. First, the main theoretical results of the article are proved:

- Proof of Proposition 1: AKO+BH with $B = 1$ and $\gamma = 1$ is equivalent to vanilla KO.
- Proof of Lemma 1: for Lasso-coefficient differences, the non-zero $W_j$ are distinct.
- Proof that the $\pi_j$ are *asymptotically* valid p-values (without any multiplicative correction): Lemma A.3.
- Statement and proof of a new general result about FDR control with quantile-aggregated p-values: Lemma A.4.
- Proof of Theorem 1.

Second, the results of some additional experiments are reported:

- Additional experiments to show that the KO-GZ alternative aggregation procedure by [GZ19] has decreasing power when the number $\kappa$ of knockoff vectors $\tilde{\mathbf{x}}$ considered simultaneously increases (we compare $\kappa = 2$ with $\kappa = 3$). We show empirically that this is not the case for AKO with respect to $B$.
- Empirical evidence for the near independence of p-values $\pi_j$.
- Additional figures for HCP 900 experiments.

# APPENDIX A: DETAILED PROOFS

## A.1 Proof of Proposition 1

We begin by noticing that the function $f : \mathbb{R}^+ \to \mathbb{Z}^+$, $f(x) = \dfrac{\#\{k : W_k \leq -x\}}{p}$ is decreasing in $x$. This means the first step of both FDR control step-up procedures, that involves ordering the intermediate p-values ascendingly, is the same as arranging the knockoff statistic in descending order: $W_{(1)} \geq W_{(2)} \geq \cdots \geq W_{(p)}$. Therefore from Eq. (7) and the definition of $\pi_j$ we have:

$$\widehat{k} = \max\left\{k : \frac{1 + \#\{i : W_{(i)} \leq -W_{(k)}\}}{p} \leq \frac{k\alpha}{p}\right\}$$

(note that we can exclude all the $\pi_{(k)} = 1$ due to the fact that $\forall\ k \in [p], \alpha \in (0, 1) : k\alpha/p < 1$).

This can be written as:

$$\widehat{k} = \max\left\{k : \frac{1 + \#\{i : W_{(i)} \leq -W_{(k)}\}}{\#\{i : W_{(i)} \geq W_{(k)}\}} \leq \alpha\right\},$$

since $\#\{i : W_{(i)} \geq W_{(k)}\} = k$ because $\{W_{(j)}\}_{j\in[p]}$ is ordered descendingly and because of the assumption that non-zero LCD statistics are distinct. Furthermore, finding the maximum index $k$ of the descending ordered sequence is equivalent to finding the minimum value in that sequence, or

$$\widehat{k} = \min\left\{W_{(k)} > 0 : \frac{1 + \#\{i : W_{(i)} \leq -W_{(k)}\}}{\#\{i : W_{(i)} \geq W_{(k)}\}} \leq \alpha\right\},$$

since all $W_{(j)} \leq 0$ (corresponding with $\pi_{(k)} = 1$) have been excluded. Without loss of generality, we can write:

$$\widehat{t}_+ = \min\left\{t > 0 : \frac{1 + \#\{i : W_i \leq -t\}}{\#\{i : W_i \geq t\}} \leq \alpha\right\}.$$

This threshold $\widehat{t}_+$ is exactly the same as the definition of threshold $\tau_+$ in Eq. (4) from the original KO procedure. □

### A.2 Proof of Lemma 1

**Setting.** Let $\mathbf{X} \in \mathbb{R}^{n\times q}$, $\boldsymbol{\beta}^* \in \mathbb{R}^q, \lambda > 0, \sigma > 0$ be fixed. Define

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta}^* + \boldsymbol{\epsilon}$$

$$\forall \boldsymbol{\beta} \in \mathbb{R}^q, \qquad L(\boldsymbol{\beta}) := \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1 .$$

with $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma\mathbf{I}_n)$ the Gaussian noise and $\|\cdot\|_p$ the $L_p$ norm.

**Classical Optimization Properties.** Since $L$ is convex, non-negative, and tends to $+\infty$ at infinity, its minimum over $\mathbb{R}^q$ exists and is attained (although may not be unique). Since $L$ is convex, its minima are characterized by a first-order condition:

$$\widehat{\boldsymbol{\beta}}_\lambda \in \underset{\boldsymbol{\beta}\in\mathbb{R}^q}{\operatorname{argmin}}\{L(\boldsymbol{\beta})\} \Leftrightarrow \qquad 0 \in \partial L(\boldsymbol{\beta})$$

which is equivalent to

$$\text{(A.17)} \qquad \begin{cases} \exists\ \widehat{\mathbf{z}} \in [-1,1]^q : \mathbf{X}^\top\mathbf{X}\widehat{\boldsymbol{\beta}}_\lambda = \mathbf{X}^\top\mathbf{y} - \dfrac{\lambda}{2}\widehat{\mathbf{z}} \\ \quad \forall j \text{ s.t. } (\widehat{\boldsymbol{\beta}}_\lambda)_j \neq 0,\ \widehat{\mathbf{z}}_j = \operatorname{sign}((\widehat{\boldsymbol{\beta}}_\lambda)_j) \end{cases}$$

As shown by [Gir14, Section 4.5.1] for instance, the fitted value $\widehat{f}_\lambda \in \mathbb{R}^n$ is uniquely defined:

$$\exists!\ \widehat{f}_\lambda \in \mathbb{R}^n \quad \text{such that} \quad \forall \widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}} \in \underset{\boldsymbol{\beta}\in\mathbb{R}^q}{\operatorname{argmin}}\{L(\boldsymbol{\beta})\}, \qquad \widehat{f}_\lambda = \mathbf{X}\widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}} .$$

As a consequence, the equicorrelation set

$$\widehat{J}_\lambda := \Big\{ j \in \{1, \dots, q\} : \Big| \mathbf{x}_j^\top (\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}}) \Big| = \lambda/2 \Big\}$$

is uniquely defined. We also have,

$$\forall\, \widehat{\boldsymbol{\beta}}_\lambda \in \operatorname*{argmin}_{\boldsymbol{\beta} \in \mathbb{R}^q} \{L(\boldsymbol{\beta})\} \,, \qquad \Big\{ j : (\widehat{\boldsymbol{\beta}}_\lambda)_j \neq 0 \Big\} \subset \widehat{J}_\lambda \tag{A.18}$$

(but these two sets are not necessarily equal, and the former set may not be uniquely defined).

Note that for every set $J \subset \{1, \dots, q\}$ such that $\forall j \notin J$, $(\widehat{\boldsymbol{\beta}}_\lambda)_j = 0$, we have $\mathbf{X}\widehat{\boldsymbol{\beta}}_\lambda = \mathbf{X}_J (\widehat{\boldsymbol{\beta}}_\lambda)_J$ so that $(\mathbf{X}^\top \mathbf{X}\widehat{\boldsymbol{\beta}}_\lambda)_J = \mathbf{X}_J^\top \mathbf{X}_J (\widehat{\boldsymbol{\beta}}_\lambda)_J$. As a consequence, taking $J = \widehat{J}_\lambda$, by eq. (A.17) and (A.18), any minimizer $\widehat{\boldsymbol{\beta}}_\lambda$ of $L$ over $\mathbb{R}^q$ satisfies

$$\mathbf{X}_{\widehat{J}_\lambda}^\top \mathbf{X}_{\widehat{J}_\lambda} (\widehat{\boldsymbol{\beta}}_\lambda)_{\widehat{J}_\lambda} = \mathbf{X}_{\widehat{J}_\lambda}^\top \mathbf{y} - \frac{\lambda}{2} \widehat{\mathbf{z}}_{\widehat{J}_\lambda} \tag{A.19}$$

for some $\widehat{\mathbf{z}}_{\widehat{J}_\lambda} \in \{-1, 1\}^{\widehat{J}_\lambda}$.

If the matrix $\mathbf{X}_{\widehat{J}_\lambda}^\top \mathbf{X}_{\widehat{J}_\lambda}$ is non-singular (that is, if $\mathbf{X}_{\widehat{J}_\lambda}$ is of rank $|\widehat{J}_\lambda|$), then the argmin of $L$ is unique [Gir14, Section 4.5.1].

Result A.1. For every $\boldsymbol{\alpha} \in \mathbb{R}^q \backslash \{0\}$, the event

$$\operatorname{rank}(\mathbf{X}_{\widehat{J}_\lambda}) = |\widehat{J}_\lambda| \,, \quad \boldsymbol{\alpha}^\top \widehat{\boldsymbol{\beta}}_\lambda = 0 \quad \text{and} \quad \exists\, j \in \{1, \dots, q\} \,, \quad \alpha_j (\widehat{\boldsymbol{\beta}}_\lambda)_j \neq 0 \,, \tag{A.20}$$

where $\{\widehat{\boldsymbol{\beta}}_\lambda\} = \operatorname{argmin}_{\boldsymbol{\beta} \in \mathbb{R}^q} \{L(\boldsymbol{\beta})\}$ is well-defined by the first property, has probability zero.

Proof. Let $\Omega$ be the event defined by Eq. (A.20). If $\Omega$ holds true, then there exists some $J \subset \{1, \dots, q\}$ and some $\widehat{\mathbf{z}} \in \{-1, 1\}^q$ such that $\operatorname{rank}(X_J) = |J|$, $(\widehat{\boldsymbol{\beta}}_\lambda)_{J^c} = 0$, and

$$\mathbf{X}_J^\top \mathbf{X}_J (\widehat{\boldsymbol{\beta}}_\lambda)_J = \mathbf{X}_J^\top \mathbf{y} - \frac{\lambda}{2} \widehat{\mathbf{z}}_J \,.$$

Indeed, this is a consequence of Eq. (A.18) and (A.19), by taking $J = \widehat{J}_\lambda$ and $\mathbf{z}$ such that $\mathbf{z}_{\widehat{J}_\lambda} = \operatorname{sign}\Big( (\widehat{\boldsymbol{\beta}}_\lambda)_{\widehat{J}_\lambda} \Big)$. Therefore, using that $\mathbf{X}_J^\top \mathbf{X}_J$ is non-singular, we get

$$\begin{aligned}
& (\widehat{\boldsymbol{\beta}}_\lambda)_J = M(J)\boldsymbol{\epsilon} + v(J, \mathbf{z}) \\
\text{where} \quad & M(J) := (\mathbf{X}_J^\top \mathbf{X}_J)^{-1} \mathbf{X}_J^\top \\
\text{and} \quad & v(J, \mathbf{z}) := (\mathbf{X}_J^\top \mathbf{X}_J)^{-1} \mathbf{X}_J^\top \mathbf{X}\boldsymbol{\beta}^* - (\mathbf{X}_J^\top \mathbf{X}_J)^{-1} \frac{\lambda}{2} \widehat{\mathbf{z}}_J \,,
\end{aligned}$$

hence

$$\boldsymbol{\alpha}^\top \widehat{\boldsymbol{\beta}}_\lambda = \boldsymbol{\alpha}_J^\top M(J)\boldsymbol{\epsilon} + \boldsymbol{\alpha}^\top v(J, \mathbf{z})$$

follows a normal distribution with variance $\sigma^2 \boldsymbol{\alpha}_J^\top M(J) M(J)^\top \boldsymbol{\alpha}_J = \sigma^2 \left\| M(J)^\top \boldsymbol{\alpha}_J \right\|^2$. Now, on $\Omega$, we also have the existence of some $j$ such that $\alpha_j (\widehat{\boldsymbol{\beta}}_\lambda)_j \neq 0$. Since $(\widehat{\boldsymbol{\beta}}_\lambda)_{J^c} = 0$, we must have $j \in J$, which shows that $\boldsymbol{\alpha}_J \neq 0$.

Overall, we have proved that

$$\begin{aligned} \Omega \subset & \bigcup_{J \in \mathcal{J}, \mathbf{z} \in \{-1,1\}^q} \Omega_{J,z} \\ \text{where} \qquad & \mathcal{J} := \{ j \in \{1, \ldots, q\} : \operatorname{rank}(X_J) = |J| \text{ and } \alpha_J \neq 0 \} \\ \text{and} \qquad & \Omega_{J,\mathbf{z}} := \left\{ \boldsymbol{\alpha}_J^\top M(J) \boldsymbol{\epsilon} + \alpha^\top v(J, \mathbf{z}) = 0 \right\} . \end{aligned}$$

For every $J \in \mathcal{J}$, $M(J)^\top \boldsymbol{\alpha}_J \neq 0$ since $\alpha_J \neq 0$ and $M(J)$ is of rank $|J|$. As a consequence, for every $J \in \mathcal{J}$ and $\mathbf{z} \in \{-1,1\}^p$, $\mathbb{P}(\Omega_{J,\mathbf{z}})$ is the probability that a Gaussian variable with non-zero variance is equal to zero, so it is equal to zero. We deduce that

$$\mathbb{P}(\Omega) \leq \sum_{J \in \mathcal{J}, \mathbf{z} \in \{-1,1\}^q} \mathbb{P}(\Omega_{J,z}) = 0$$

since the sets $\mathcal{J}$ and $\{-1,1\}^q$ are finite. □

Applying Result A.1 to the case where $\mathbf{X}$ concatenates the original $p$ covariates and their knockoff counterparts (hence $q = 2p$), we get that, apart from the event where $\mathbf{X}_{\widehat{\mathcal{J}_\lambda}}$ is not full rank, for every $j \in \{1, \ldots, p\}$, $W_j$ takes any fixed non-zero value with probability zero (with $\alpha_j = \pm 1$, $\alpha_{j+p} = \pm 1$, $\alpha_k = 0$ otherwise).

Similarly, the above lemma shows that for every $j \neq j' \in \{1, \ldots, p\}$:

$$\mathbb{P}(\mathbf{X}_{\widehat{\mathcal{J}_\lambda}} \text{ is full-rank and } \exists j \neq j', W_j = W_{j'}, W_j \neq 0, W_{j'} \neq 0) = 0.$$

As a consequence, with probability 1, all the non-zero $W_j$ are distinct if $X_{\widehat{\mathcal{J}_\lambda}}$ is full-rank. □

Remark A.4. The proof of Result A.1 is also valid for other noise distributions: it only assumes that the support of the distribution of $\boldsymbol{\epsilon}$ is not included into any hyperplane of $\mathbb{R}^n$.

## A.3 Asymptotic Validity of Intermediate P-values

We consider in this section an asymptotic regime where $p \to +\infty$.

ASSUMPTION A.2 (Asymptotic regime $p \to +\infty$). When $p$ grows to infinity, $n$, $\mathbf{X}$, $\beta^*$, $\boldsymbol{\epsilon}$ and $\mathbf{y}$ all depend on $p$ implicitly. We assume that for every integer $j \geq 1$, $\mathbb{1}_{\beta^*_j = 0}$ does not depend on $p$ (as soon as $p \geq j$), and that

$$\frac{|\mathcal{S}|}{p} = \frac{\left|\{j \in [p] : \beta^*_j \neq 0\}\right|}{p} \xrightarrow[p \to +\infty]{} 0 \,.$$

When making Assumption 1, we also assume that $\mathbb{P}_0$ does not depend on $p$.

LEMMA A.3. If Assumptions 1 and A.2 hold true, then for all $j \geq 1$ such that $\beta^*_j = 0$, the empirical p-value $\pi_j$ defined by Eq. (5) is a valid p-value asymptotically, that is,

$$\forall t \in [0,1] \,, \qquad \lim_{p \to +\infty} \mathbb{P}(\pi_j \leq t) \leq t \,.$$

Note that our proof of Theorem 1 in Section A.5 relies on the use of Lemma 2 with $t$ that can be of order $1/p$. Therefore, Lemma A.3 above is not sufficient for our needs. Nevertheless, it still provides a interesting insight about the $\pi_j$, and justifies (asymptotically) their name, which is why we state and prove this result here.

PROOF. By definition, $\pi_j \leq 1$ almost surely, so the result holds when $t = 1$. Let us now focus on the case where $t \in [0,1)$. Let $F_0$ denote the c.d.f. of $\mathbb{P}_0$, the common distribution of the null statistics $\{W_j\}_{1 \leq j \leq p \,/\, \beta^*_j = 0}$, which exists by Assumption 1. Let $j \geq 1$ such that $\beta^*_j = 0$ be fixed, and assume that $p \geq j$ is large enough so that $|\mathcal{S}^c| \geq 2$. Let $m = |\mathcal{S}^c| - 1 \geq 1$ as in the proof of Lemma 2. Note that $m$ depends on $p$, and $m/p \to 1$ as $p \to +\infty$ by Assumption A.2, hence $m \to +\infty$ as $p \to +\infty$.

By definition of $\pi_j$, when $W_j > 0$ we have:

$$\begin{aligned} \pi_j &= \frac{1 + \#\{k \in [p] : W_k \leq -W_j\}}{p} \\ (\text{since } W_j > 0 > -W_j) &= \frac{1 + \#\{k \in \mathcal{S} : W_k \leq -W_j\} + \#\{k \in \mathcal{S}^c \setminus \{j\} : W_k \leq -W_j\}}{p} \\ &\geq \frac{\#\{k \in \mathcal{S}^c \setminus \{j\} : W_k \leq -W_j\}}{p} \\ &= \frac{\widehat{F}_m(-W_j)}{\alpha_p} \qquad \text{(A.21)} \end{aligned}$$

where $\alpha_p \triangleq \dfrac{p}{m}$ and for all $u \in \mathbb{R}$,

$$\widehat{F}_m(u) \triangleq \frac{\#\{k \in \mathcal{S}^c \setminus \{j\} : W_k \leq u\}}{m}$$

is the empirical cdf of $\{W_k\}_{k\in\mathcal{S}^c\setminus\{j\}}$.

Now, since $\{W_k\}_{k\in\mathcal{S}^c\setminus\{j\}}$ are *i.i.d.* with distribution $\mathbb{P}_0$ by Assumption 1 , the law of large numbers implies that, for all $u \in \mathbb{R}$,

$$\widehat{F}_m(u) \xrightarrow[p\to+\infty]{\text{a.s.}} F_0(u)\,.$$

Since we assume $\lim_{p\to+\infty} |\mathcal{S}|/p = 0$, $\lim_{p\to+\infty} \alpha_p = 1$ and we get that for all $u \in \mathbb{R}$,

$$\frac{1}{\alpha_p}\widehat{F}_m(u) \xrightarrow[p\to+\infty]{\text{a.s.}} F_0(u)\,.$$

Since $W_j$ is independent from $\{W_k\}_{k\in\mathcal{S}^c\setminus\{j\}}$, this result also holds true *conditionally to* $W_j$, with $u = -W_j$. Given that almost sure convergence implies convergence in distribution, we have: conditionally to $W_j$,

$$\frac{1}{\alpha_p}\widehat{F}_m(-W_j) \xrightarrow[p\to+\infty]{\text{(d)}} F_0(-W_j) \overset{(d)}{=} F_0(W_j) \tag{A.22}$$

where the latter equality comes from the fact that $W_j$ has a symmetric distribution, as shown in Remark 1.

So, when $W_j > 0$, for every $t \in [0,1)$,

$$\begin{aligned}\limsup_{p\to+\infty} \mathbb{P}(\pi_j \le t \,|\, W_j) &\le \limsup_{p\to+\infty} \mathbb{P}\left(\frac{\widehat{F}_m(-W_j)}{\alpha_p} \le t \,\middle|\, W_j\right) \qquad \text{by Eq. (A.21)}\\ &\le \mathbb{1}_{F_0(W_j)\le t} \qquad\qquad (\text{A.23})\end{aligned}$$

by Eq. (A.22) combined with the Portmanteau theorem.

Therefore, for every $t \in [0,1)$,

$$\begin{aligned}\limsup_{p\to+\infty} \mathbb{P}(\alpha_p\pi_j \le t) &= \limsup_{p\to+\infty}\Big\{\mathbb{P}(\alpha_p\pi_j \le t \text{ and } W_j > 0) + \underbrace{\mathbb{P}(\alpha_p\pi_j \le t \text{ and } W_j \le 0)}_{\substack{=0 \text{ since } \alpha_p\ge 1>t\\ \text{and } \pi_j=1 \text{ when } W_j\le 0}}\Big\}\\ &= \limsup_{p\to+\infty} \mathbb{E}[\mathbb{P}(\alpha_p\pi_j \le t \mid W_j)\mathbb{1}_{W_j>0}]\\ &\le \mathbb{E}\left[\limsup_{|\mathcal{S}^c|\to+\infty}\{\mathbb{P}(\alpha_p\pi_j \le t \mid W_j)\mathbb{1}_{W_j>0})\}\right]\\ &\le \mathbb{P}(F_0(W_j) \le t) \qquad \text{by Eq. (A.23)}\\ &\le t\,,\end{aligned}$$

which concludes the proof. □

### A.4 A General FDR Control with Quantile-aggregated P-values

The proof of Theorem 1 relies on an adaptation of results proved by [MMB09, Theorems 3.1 and 3.3] about aggregation of p-values. The results of [MMB09], whose proof relies on the proofs of [BY01], are stated for randomized p-values obtained through sample splitting. The following lemma shows that they actually apply to any family of p-values.

Lemma A.4. Let $(\pi_j^{(b)})_{1\leq j\leq p, 1\leq b\leq B}$ be a family of random variables with values in $[0,1]$. Let $\gamma\in(0,1]$, $\bar{\alpha}\in[0,1]$ and $\mathcal{N}\subset[p]$ be fixed. Let us define

$$\forall j\in[p]\,,\quad Q_j \triangleq \frac{p}{\gamma}q_\gamma(\{\pi_j^{(b)} \,:\, 1\leq b\leq B\})\quad\text{where}\quad q_\gamma(\cdot)\text{ is the }\gamma\text{-quantile function,}$$

$$\widehat{h}\triangleq\max\{i\in[p]\,:\,Q_{(i)}\leq i\bar{\alpha}\}\quad\text{where}\quad Q_{(1)}\leq\cdots\leq Q_{(p)}\,,$$

$$\text{and}\quad \widehat{S}\triangleq\{j\in[p]\,:\,Q_j\leq Q_{(\widehat{h})}\}\,.$$

Then,

$$\text{(A.24)}\qquad \mathbb{E}\left[\frac{|\widehat{S}\cap\mathcal{N}|}{|\widehat{S}|\vee 1}\right]\leq\sum_{j=1}^{p-1}\left(\frac{1}{j}-\frac{1}{j+1}\right)F(j)+\frac{F(p)}{p}$$

$$\text{where}\qquad\forall j\in[p]\,,\qquad F(j)\triangleq\frac{1}{\gamma}\frac{1}{B}\sum_{b=1}^{B}\sum_{i\in\mathcal{N}}\mathbb{P}\left(\pi_i^{(b)}\leq\frac{j\bar{\alpha}\gamma}{p}\right)\,.$$

As a consequence, if some $C\geq 0$ exists such that

$$\text{(A.25)}\qquad \forall t\geq 0, \forall b\in[B]\,,\forall i\in\mathcal{N}\,,\qquad \mathbb{P}\left(\pi_i^{(b)}\leq t\right)\leq Ct\,,$$

then we have

$$\text{(A.26)}\qquad \mathbb{E}\left[\frac{|\widehat{S}\cap\mathcal{N}|}{|\widehat{S}|\vee 1}\right]\leq\frac{|\mathcal{N}|C}{p}\left(\sum_{j=1}^{p}\frac{1}{j}\right)\bar{\alpha}\,.$$

Let us emphasize that Lemma A.4 can be useful in general, well beyond knockoff aggregation. To the best of our knowledge, Lemma A.4 is new. In particular, the recent preprint by [RD19], that studies p-values aggregation procedures, focuses on FWER controlling procedures, whereas Lemma A.4 provides an FDR control for a less conservative procedure.

Proof. For every $i,j,k\in[p]$, let us define

$$p_{i,j,k}=\begin{cases}\mathbb{P}\Big(Q_i\in\big((j-1)\bar{\alpha},j\bar{\alpha}\big]\,,\,i\in\widehat{S}\text{ and }|\widehat{S}|=k\Big) & \text{if } j\geq 2\\ \mathbb{P}(Q_i\in[0,\bar{\alpha}]\,,\,i\in\widehat{S}\text{ and }|\widehat{S}|=k) & \text{if } j=1\,.\end{cases}$$

Then,

$$\begin{aligned}\frac{|\widehat{S} \cap \mathcal{N}|}{|\widehat{S}| \vee 1} &= \sum_{k=1}^{p} \mathbb{1}_{|\widehat{S}|=k} \frac{\sum_{i \in \mathcal{N}} \mathbb{1}_{i \in \widehat{S}}}{k} \\ &= \sum_{i \in \mathcal{N}} \sum_{k=1}^{p} \frac{1}{k} \mathbb{1}_{|\widehat{S}|=k \text{ and } i \in \widehat{S}} \\ &= \sum_{i \in \mathcal{N}} \sum_{k=1}^{p} \frac{1}{k} \mathbb{1}_{|\widehat{S}|=k \text{ and } i \in \widehat{S} \text{ and } 0 \leq Q_i \leq k\bar{\alpha}}\end{aligned}$$

since $i \in \widehat{S}$ and $|\widehat{S}| = k$ implies that $Q_i \leq Q_{(\widehat{h})} \leq \widehat{h}\bar{\alpha} = k\bar{\alpha}$. Taking an expectation and writing that

$$\mathbb{1}_{0 \leq Q_i \leq k\bar{\alpha}} = \mathbb{1}_{Q_i \in [0,\bar{\alpha}]} + \sum_{j=2}^{k} \mathbb{1}_{Q_i \in ((j-1)\bar{\alpha}, j\bar{\alpha}]} \, ,$$

we get —following the computations of [MMB09, proof of Theorems 3.3], which themselves rely on the ones of [BY01]—,

$$\begin{aligned}\mathbb{E}\left[\frac{|\widehat{S} \cap \mathcal{N}|}{|\widehat{S}| \vee 1}\right] &\leq \sum_{i \in \mathcal{N}} \sum_{k=1}^{p} \frac{1}{k} \sum_{j=1}^{k} p_{i,j,k} = \sum_{i \in \mathcal{N}} \sum_{j=1}^{p} \sum_{k=j}^{p} \frac{1}{k} p_{i,j,k} \\ &\leq \sum_{i \in \mathcal{N}} \sum_{j=1}^{p} \sum_{k=j}^{p} \frac{1}{j} p_{i,j,k} = \sum_{j=1}^{p} \frac{1}{j} \underbrace{\sum_{i \in \mathcal{N}} \sum_{k=j}^{p} p_{i,j,k}}_{=\overline{F}(j) - \overline{F}(j-1)\mathbb{1}_{j \geq 2}}\end{aligned}$$

$$\text{where} \qquad \forall j \in \{1, \ldots, p\} \, , \qquad \overline{F}(j) \triangleq \sum_{i \in \mathcal{N}} \sum_{j'=1}^{j} \sum_{k=1}^{p} p_{i,j',k} \, .$$

Since the above upper bound is equal to

$$\overline{F}(1) + \sum_{j=2}^{p} \frac{1}{j} \left[ \overline{F}(j) - \overline{F}(j-1) \right] = \sum_{j=1}^{p} \left( \frac{1}{j} - \frac{1}{j+1} \right) \overline{F}(j) + \frac{\overline{F}(p)}{p} \, ,$$

$$\text{(A.27)} \qquad \text{we get that} \qquad \mathbb{E}\left[\frac{|\widehat{S} \cap \mathcal{N}|}{|\widehat{S}| \vee 1}\right] \leq \sum_{j=1}^{p} \left( \frac{1}{j} - \frac{1}{j+1} \right) \overline{F}(j) + \frac{\overline{F}(p)}{p} \, .$$

Notice also that

$$\overline{F}(j) = \sum_{i \in \mathcal{N}} \mathbb{P}(Q_i \leq j\bar{\alpha} \text{ and } i \in \widehat{S}) \leq \sum_{i \in \mathcal{N}} \mathbb{P}(Q_i \leq j\bar{\alpha}) \ \ .$$

Now, as done by [MMB09, proof of Theorems 3.1], we remark that $Q_i \leq j\bar{\alpha}$ is equivalent to

$$\frac{1}{B} \left| \left\{ b \in [B] \, : \, \frac{p\pi_i^{(b)}}{\gamma} \leq j\bar{\alpha} \right\} \right| = \frac{1}{B} \sum_{b=1}^{B} \mathbb{1}_{p\pi_i^{(b)} \leq j\bar{\alpha}\gamma} \geq \gamma$$

so that

$$\begin{aligned}\mathbb{P}(Q_i \leq j\bar{\alpha}) &= \mathbb{P}\left(\frac{1}{B}\sum_{b=1}^{B}\mathbb{1}_{p\pi_i^{(b)}\leq j\bar{\alpha}\gamma} \geq \gamma\right)\\ &\leq \frac{1}{\gamma}\mathbb{E}\left[\frac{1}{B}\sum_{b=1}^{B}\mathbb{1}_{p\pi_i^{(b)}\leq j\bar{\alpha}\gamma}\right] \qquad \text{by Markov inequality}\\ &= \frac{1}{\gamma}\frac{1}{B}\sum_{b=1}^{B}\mathbb{P}\left(p\pi_i^{(b)} \leq j\bar{\alpha}\gamma\right) .\end{aligned}$$

Therefore,

$$\overline{F}(j) \leq \sum_{i\in\mathcal{N}}\frac{1}{\gamma}\frac{1}{B}\sum_{b=1}^{B}\mathbb{P}\left(p\pi_i^{(b)} \leq j\bar{\alpha}\gamma\right) = F(j)\,,$$

so that Eq. (A.27) implies Eq. (A.24).

If condition (A.25) holds true, then, for every $j \in [p]$,

$$F(j) \leq \frac{|\mathcal{N}|C}{\gamma}\frac{j\bar{\alpha}\gamma}{p} = \frac{|\mathcal{N}|C\bar{\alpha}}{p}j\,,$$

hence Eq. (A.24) shows that

$$\mathbb{E}\left[\frac{|\widehat{S}\cap\mathcal{N}|}{|\widehat{S}|\vee 1}\right] \leq \sum_{j=1}^{p-1}\frac{F(j)}{j(j+1)} + \frac{F(p)}{p} \leq \frac{|\mathcal{N}|C\bar{\alpha}}{p}\sum_{j=1}^{p-1}\frac{1}{j+1} + \frac{|\mathcal{N}|C\bar{\alpha}}{p} = \frac{|\mathcal{N}|C\bar{\alpha}}{p}\left(\sum_{j=1}^{p}\frac{1}{j}\right)\,,$$

which is the desired result. □

### A.5 Proof of Theorem 1

We can now prove Theorem 1. We apply Lemma A.4 with $\bar{\alpha} = \beta(p)\alpha$, $\mathcal{N} = \mathcal{S}^c$, so that $\widehat{S} = \widehat{\mathcal{S}}_{AKO+BY}$. Since the $\pi_j^{(b)}$, $b = 1, \dots, B$, have the same distribution as $\pi_j$, by Lemma 2, condition (A.25) holds true with $C = \kappa p/|\mathcal{S}^c|$, and Eq. (A.26) yields the desired result. □

Note that an FDR control for AKO such as Theorem 1 cannot be obtained straightforwardly from the arguments of [BC15] and [CFJL18]. One key reason for this is that their proof relies on a reordering of the features according to the values of $(|W_j|)_{j\in[p]}$ [BC15, Section 5.2], such a reordering being permitted since the signs of the $W_j$ are iid coin flips *conditionally to* the $(|W_j|)_{j\in[p]}$ [CFJL18, Lemma 2]. In the case of AKO, we must handle the $(W_j^{(b)})_{j\in[p]}$ *simultaneously for all* $b \in [B]$, and conditioning with respect to $(|W_j^{(b)}|)_{j\in[p],b\in[B]}$ may reveal some information about the signs of the $(W_j^{(b)})_{j\in[p]}$ as soon as $B > 1$. At least, it does not seem obvious to us that the key result of [CFJL18, Lemma 2] can be proved *conditionally to* the $(|W_j^{(b)}|)_{j\in[p],b\in[B]}$ when $B > 1$, so that the proof strategy of [CFJL18] breaks down in the case of AKO with $B > 1$.

## APPENDIX B: ADDITIONAL EXPERIMENTAL RESULTS

### B.1 Demonstration of Aggregated Multiple Knockoff vs. Simultaneous Knockoff

Using the same simulation settings as in Section 5.1 with $n = 500, p = 1000$ and varying simulation parameters to generate Figure 2 in the main text, we benchmark only aggregation of multiple knockoffs (AKO) with 5 and 10 bootstraps ($B = 5$ and $B = 10$) and compare with simultaneous knockoffs with 2 and 3 bootstraps ($\kappa = 2$ and $\kappa = 3$). Results in Figure B.5 show that while increasing the number of knockoff bootstraps makes simultaneous knockoffs more conservative, doing so with AKO makes the algorithm more powerful (and in the worst case retains the same power with fewer bootstraps).

### B.2 Empirical Evidence on the Independence of Aggregated P-values $\bar{\pi}$

Using the same simulation settings as in Section 5.1 with $n = 500, p = 1000, \rho = 0.6, \mathrm{snr} = 3.0, \mathrm{sparsity} = 0.06$ we generate 100 observations of *aggregated* p-values $\bar{\boldsymbol{\pi}}$. Then, we compute the Spearman rank-order correlation coefficient of the *Null* $\bar{\pi}_j$ for these 100 observations along with their two-sided p-value (for a hypothesis test whose null hypothesis is that two sets of data are uncorrelated).

The results are illustrated in Figure B.6: the Spearman correlation values are concentrated around zero, while the distribution of associated p-values seems to be a mixture between a uniform distribution and a small mixture component consisting of mostly non-null p-values. This indicates near independence between the aggregated p-values using quantile-aggregation [MMB09], hence justifies our use of BH step-up procedure for selecting FDR controlling threshold in the AKO algorithm.

Remark B.5. Again, it is worth noticing that the empirical evidence we have shown is only done in a setting with a Toeplitz structure for the covariance matrix. However, as explained in the main text, this correlation setting is usually found in neuroimaging and genomics data. Hence, we believe that assuming short-distance correlations between the $(X_j)_{j\in[p]}$ is a mild assumption, which should be satisfied in the practical scenarios where we want to apply our algorithm.

The decoding maps returned by the KO, AKO and DL inference procedures are presented in Figure B.7. As quantified by the Jaccard index in the main text, we observe that the AKO solution is typically closer to an external method based on the desparsified lasso (DL). Moreover, AKO is also typically more sensitive than KO alone.

The seven classification problems are the following:

- Emotion: predict whether the participant watches an angry face or a geometric shape.
- Gambling: predict whether the participant gains or loses gambles.
- Motor foot: predict whether the participant moves the left or right foot.
- Motor hand: predict whether the participant moves the left or right hand.
- Relational: predict whether the participant matches figures or identified feature similarities.
- Social: predict whether the participant watches a movie with social behavior or not.
- Working Memory: predict whether the participant does a 0-back or a 2-back task.

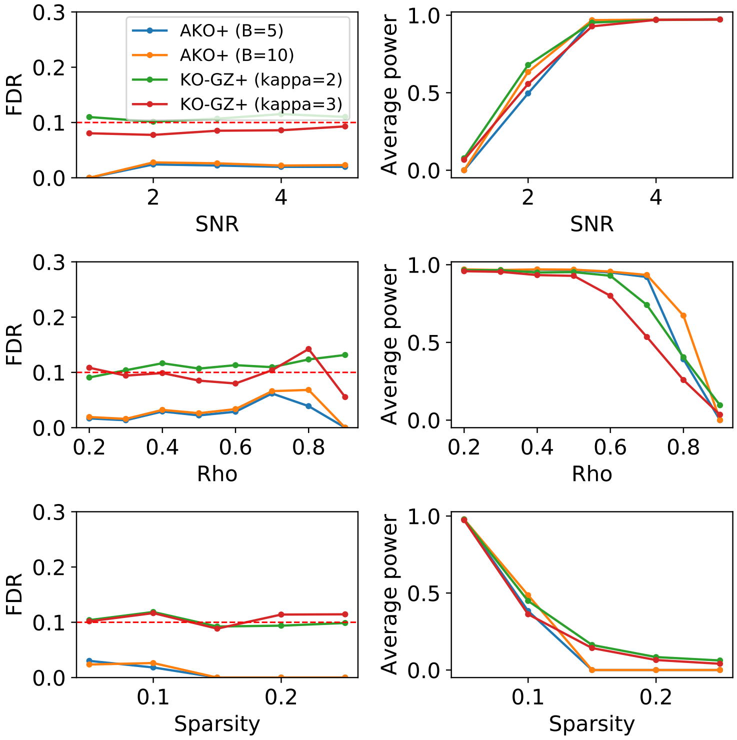

Figure B.5: **Aggregation of multiple knockoffs** ($B = 5$ **and** $B = 10$) **vs. simultaneous knockoffs** ($\kappa = 2$ **and** $\kappa = 3$)**.** A clear loss in statistical power is demonstrated in the latter method when increasing the number of bootstraps $\kappa$, while the former (AKO) shows the opposite: with $B = 10$ bootstraps there are small, yet consistent gains in the number of true detections compared to using only $B = 5$ bootstraps.

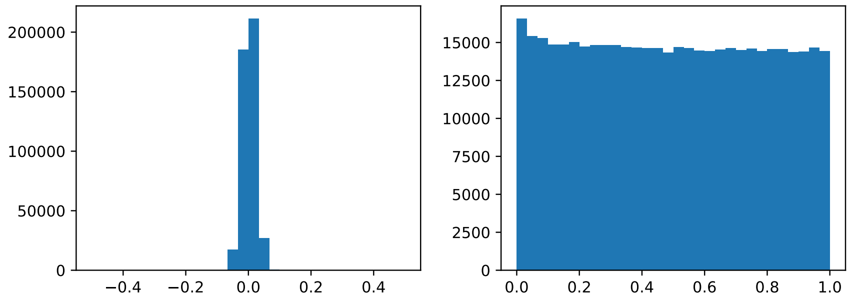

Figure B.6: **Left: Histogram of Spearman correlation values for 100 samples of null aggregated p-values $\bar{\pi}_j$. Right: Histogram of corresponding p-values for the Spearman correlation**

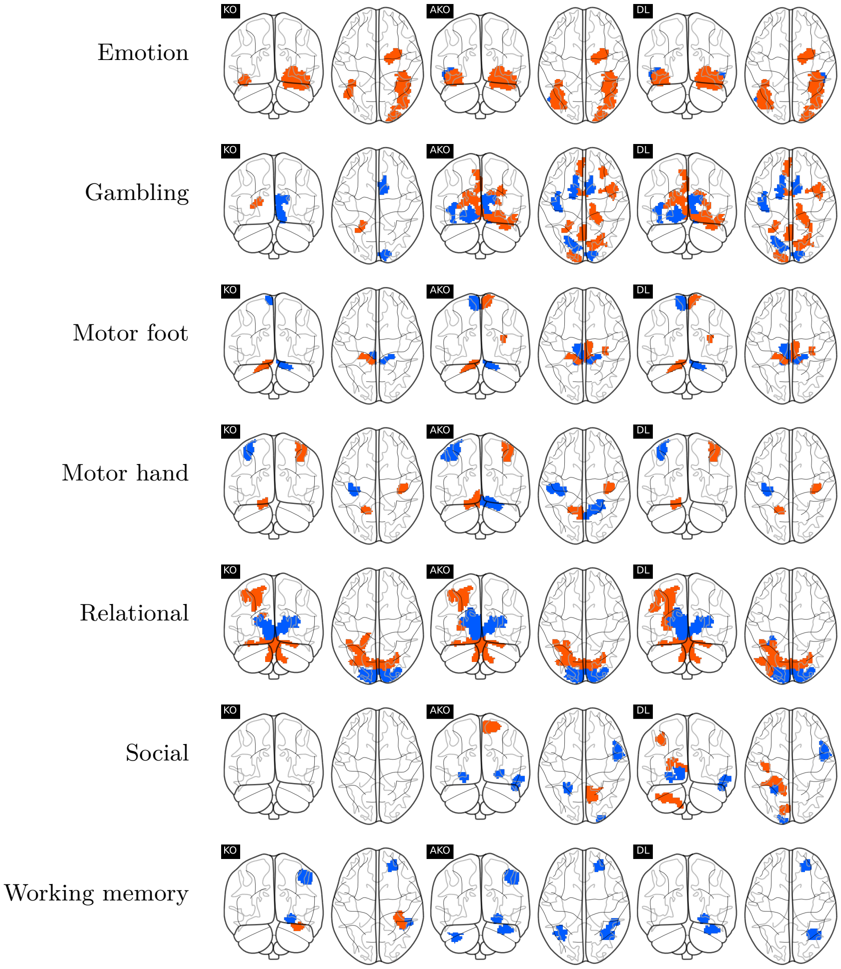

Figure B.7: **Decoding maps obtained for seven classification tasks.** Emotion, gambling, motor foot, motor hand, relational, social and working memory refer to 7 binary tasks that were considered based on the HCP900 dataset. We observe that AKO is typically more sensitive than KO, and yields solution closer to that of an independent solution based on a desparsified-Lasso (DL) estimator.