Supplementary material to: "Aggregation of Multiple Knockoffs"

Outline. The supplementary material 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 π_j are *asymptotically* valid p-values (without any multiplicative correction): Lemma A.1.
- Statement and proof of a new general result about FDR control with quantile-aggregated p-values: Lemma A.2.
- 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 Gimenez and Zou (2019) has decreasing power when the number κ 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 π_j.
- Additional figures for HCP 900 experiments.

A. Detailed Proofs

A.1. Proof of Proposition 1

We begin by noticing that the function $f : \mathbb{R}^+ \to \mathbb{Z}^+$, $f(x) = \dfrac{\#\{k : W_k \leqslant -x\}}{p}$ 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: $W_{(1)} \geqslant W_{(2)} \geqslant \cdots \geqslant W_{(p)}$. Therefore from Eq. (7) and the definition of π_j we have:

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

(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:

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

since $\#\{i : W_{(i)} \geqslant W_{(k)}\} = k$ 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

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

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

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

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

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

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

As shown by Giraud (2014, 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

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

is uniquely defined. We also have,

$$\forall\, \widehat{\boldsymbol{\beta}}_\lambda \in \underset{\boldsymbol{\beta} \in \mathbb{R}^q}{\operatorname{argmin}} \{L(\boldsymbol{\beta})\}\ , \qquad \left\{ j : (\widehat{\boldsymbol{\beta}}_\lambda)_j \neq 0 \right\} \subset \widehat{J}_\lambda \tag{A.2}$$

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

Note that for every set $J \subset \{1, \ldots, q\}$ 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.1) and (A.2), any minimizer $\widehat{\boldsymbol{\beta}}_\lambda$ of L over $\mathbb{R}^q$ satisfies

$$\mathbf{X}_{\widehat{J}_\lambda}^\top \mathbf{X}_{\widehat{J}_\lambda} (\widehat{\boldsymbol{\beta}}_\lambda)_{\widehat{J}_\lambda} = \mathbf{X}_{\widehat{J}_\lambda}^\top \mathbf{y} - \frac{\lambda}{2} \widehat{\mathbf{z}}_{\widehat{J}_\lambda} \tag{A.3}$$

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 (Giraud, 2014, Section 4.5.1).

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

$$\operatorname{rank}(\mathbf{X}_{\widehat{J}_\lambda}) = |\widehat{J}_\lambda|\,, \quad \boldsymbol{\alpha}^\top \widehat{\boldsymbol{\beta}}_\lambda = 0 \quad \text{and} \quad \exists\, j \in \{1, \ldots, q\}\ , \quad \alpha_j (\widehat{\boldsymbol{\beta}}_\lambda)_j \neq 0\,, \tag{A.4}$$

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 Ω be the event defined by Eq. (A.4). If Ω holds true, then there exists some $J \subset \{1, \ldots, q\}$ 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.2) and (A.3), by taking $J = \widehat{J}_\lambda$ and $\mathbf{z}$ such that $\mathbf{z}_{\widehat{J}_\lambda} = \operatorname{sign}\left((\widehat{\boldsymbol{\beta}}_\lambda)_{\widehat{J}_\lambda}\right)$. Therefore, using that $\mathbf{X}_J^\top \mathbf{X}_J$ is non-singular, we get

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

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 Ω, 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

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

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 $\mathbf{X}_{\widehat{J}_\lambda}$ 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 $j \neq j' \in \{1, \ldots, p\}$:

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

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

Remark A.1. 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.1 (Asymptotic regime $p \to +\infty$)**.** When p grows to infinity, n, $\mathbf{X}$, β^*, $\boldsymbol{\epsilon}$ and $\mathbf{y}$ all depend on p implicitly. We assume that for every integer $j \geqslant 1$, $\mathbb{1}_{\beta_j^* = 0}$ does not depend on p (as soon as $p \geqslant j$), 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.1. If Assumptions 1 and A.1 hold true, then for all $j \geqslant 1$ such that $\beta_j^* = 0$, the empirical p-value π_j defined by Eq. (5) is a valid p-value asymptotically, that is,

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

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.1 above is not sufficient for our needs. Nevertheless, it still provides a interesting insight about the π_j, and justifies (asymptotically) their name, which is why we state and prove this result here.

Proof. By definition, $\pi_j \leqslant 1$ 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 $\{W_j\}_{1 \leqslant j \leqslant p \,/\, \beta_j^* = 0}$, which exists by Assumption 1. Let $j \geqslant 1$ such that $\beta_j^* = 0$ be fixed, and assume that $p \geqslant j$ is large enough so that $|\mathcal{S}^c| \geqslant 2$. Let $m = |\mathcal{S}^c| - 1 \geqslant 1$ 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.1, hence $m \to +\infty$ as $p \to +\infty$.

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

$$
\begin{aligned}
\pi_j &= \frac{1 + \#\{k \in [p] : W_k \leqslant -W_j\}}{p} \\
&= \frac{1 + \#\{k \in \mathcal{S} : W_k \leqslant -W_j\} + \#\{k \in \mathcal{S}^c \setminus \{j\} : W_k \leqslant -W_j\}}{p} \quad \text{(since } W_j > 0 > -W_j\text{)} \\
&\geqslant \frac{\#\{k \in \mathcal{S}^c \setminus \{j\} : W_k \leqslant -W_j\}}{p} \qquad \text{(A.5)} \\
&= \frac{\widehat{F}_m(-W_j)}{\alpha_p} \qquad \text{(A.6)}
\end{aligned}
$$

$$
\text{where} \qquad \alpha_p \triangleq \frac{p}{m}
$$

$$
\text{and} \qquad \forall u \in \mathbb{R}, \qquad \widehat{F}_m(u) \triangleq \frac{\#\{k \in \mathcal{S}^c \setminus \{j\} : W_k \leqslant u\}}{m}
$$

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,

$$
\frac{1}{\alpha_p} \widehat{F}_m(-W_j) \xrightarrow[p \to +\infty]{\text{(d)}} F_0(-W_j) \overset{(d)}{=} F_0(W_j) \qquad \text{(A.7)}
$$

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 $t \in [0, 1)$,

$$
\begin{aligned}
\limsup_{p \to +\infty} \mathbb{P}(\pi_j \leqslant t \,|\, W_j) &\leqslant \limsup_{p \to +\infty} \mathbb{P}\left(\frac{\widehat{F}_m(-W_j)}{\alpha_p} \leqslant t \,\middle|\, W_j \right) \qquad \text{by Eq. (A.6)} \\
&\leqslant \mathbb{1}_{F_0(W_j) \leqslant t} \qquad \text{(A.8)}
\end{aligned}
$$

by Eq. (A.7) combined with the Portmanteau theorem.

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

$$\begin{aligned}\limsup_{p\to+\infty} \mathbb{P}(\alpha_p \pi_j \leqslant t) &= \limsup_{p\to+\infty} \Big\{ \mathbb{P}(\alpha_p \pi_j \leqslant t \text{ and } W_j > 0) + \underbrace{\mathbb{P}(\alpha_p \pi_j \leqslant t \text{ and } W_j \leqslant 0)}_{=0 \text{ since } \alpha_p \geqslant 1 > t \text{ and } \pi_j = 1 \text{ when } W_j \leqslant 0} \Big\} \\ &= \limsup_{p\to+\infty} \mathbb{E}[\mathbb{P}(\alpha_p \pi_j \leqslant t \mid W_j) \mathbb{1}_{W_j>0}] \\ &\leqslant \mathbb{E}\left[\limsup_{|\mathcal{S}^c|\to+\infty} \{\mathbb{P}(\alpha_p \pi_j \leqslant t \mid W_j) \mathbb{1}_{W_j>0})\} \right] \\ &\leqslant \mathbb{P}\big(F_0(W_j) \leqslant t\big) \qquad \text{by Eq. (A.8)} \\ &\leqslant t \,,\end{aligned}$$

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 Meinshausen et al. (2009, Theorems 3.1 and 3.3) about aggregation of p-values. The results of Meinshausen et al. (2009), whose proof relies on the proofs of Benjamini and Yekutieli (2001), 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.2. Let $(\pi_j^{(b)})_{1\leqslant j\leqslant p, 1\leqslant b\leqslant B}$ 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

$$\begin{aligned} \forall j \in [p]\,, \qquad & Q_j \triangleq \frac{p}{\gamma} q_\gamma\big(\{\pi_j^{(b)} \,:\, 1 \leqslant b \leqslant B\}\big) \qquad \text{where} \qquad q_\gamma(\cdot) \text{ is the } \gamma\text{-quantile function,} \\ & \widehat{h} \triangleq \max\big\{ i \in [p] \,:\, Q_{(i)} \leqslant i\bar{\alpha} \big\} \qquad \text{where} \qquad Q_{(1)} \leqslant \cdots \leqslant Q_{(p)}\,, \\ \text{and} \qquad & \widehat{S} \triangleq \big\{ j \in [p] \,:\, Q_j \leqslant Q_{(\widehat{h})} \big\}\,. \end{aligned}$$

Then,

$$\mathbb{E}\left[\frac{|\widehat{S} \cap \mathcal{N}|}{|\widehat{S}| \vee 1} \right] \leqslant \sum_{j=1}^{p-1} \left(\frac{1}{j} - \frac{1}{j+1} \right) F(j) + \frac{F(p)}{p} \tag{A.9}$$

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

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

$$\forall t \geqslant 0, \forall b \in [B]\,, \forall i \in \mathcal{N}\,, \qquad \mathbb{P}\left(\pi_i^{(b)} \leqslant t \right) \leqslant Ct\,, \tag{A.10}$$

then we have

$$\mathbb{E}\left[\frac{|\widehat{S} \cap \mathcal{N}|}{|\widehat{S}| \vee 1} \right] \leqslant \frac{|\mathcal{N}|C}{p} \left(\sum_{j=1}^{p} \frac{1}{j} \right) \bar{\alpha}\,. \tag{A.11}$$

Let us emphasize that Lemma A.2 can be useful in general, well beyond knockoff aggregation. To the best of our knowledge, Lemma A.2 is new. In particular, the recent preprint by Romano and DiCiccio (2019), that studies p-values aggregation procedures, focuses on FWER controlling procedures, whereas Lemma A.2 provides an FDR control for a less conservative procedure.

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

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

Then,

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

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

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

we get —following the computations of Meinshausen et al. (2009, proof of Theorems 3.3), which themselves rely on the ones of Benjamini and Yekutieli (2001)—,

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

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

Since the above upper bound is equal to

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

$$\text{we get that}\qquad\mathbb{E}\left[\frac{|\widehat{S}\cap\mathcal{N}|}{|\widehat{S}|\vee 1}\right]\leqslant\sum_{j=1}^{p}\left(\frac{1}{j}-\frac{1}{j+1}\right)\overline{F}(j)+\frac{\overline{F}(p)}{p}\,. \tag{A.12}$$

Notice also that

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

Now, as done by Meinshausen et al. (2009, proof of Theorems 3.1), we remark that $Q_i\leqslant j\bar{\alpha}$ is equivalent to

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

so that

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

Therefore,

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

so that Eq. (A.12) implies Eq. (A.9).

If condition (A.10) holds true, then, for every $j \in [p]$,

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

hence Eq. (A.9) shows that

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

which is the desired result. □

A.5. Proof of Theorem 1

We can now prove Theorem 1. We apply Lemma A.2 with $\bar{\alpha} = \beta(p)\alpha$, $\mathcal{N} = \mathcal{S}^c$, so that $\widehat{S} = \widehat{\mathcal{S}}_{AKO+BY}$. Since the $\pi_j^{(b)}$, $b = 1, \ldots, B$, have the same distribution as π_j, by Lemma 2, condition (A.10) holds true with $C = \kappa p / |\mathcal{S}^c|$, and Eq. (A.11) yields the desired result. □

Note that an FDR control for AKO such as Theorem 1 cannot be obtained straightforwardly from the arguments of Barber and Candès (2015) and Candès et al. (2018). One key reason for this is that their proof relies on a reordering of the features according to the values of $(|W_j|)_{j\in[p]}$ (Barber and Candès, 2015, Section 5.2), such a reordering being permitted since the signs of the W_j are iid coin flips *conditionally to* the $(|W_j|)_{j\in[p]}$ (Candès et al., 2018, 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 Candès et al. (2018, 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 Candès et al. (2018) breaks down in the case of AKO with $B > 1$.

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.1 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.2: 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 (Meinshausen et al., 2009), hence justifies our use of BH step-up procedure for selecting FDR controlling threshold in the AKO algorithm.

Remark B.2. 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.

B.3. Brain Maps Obtained For Seven Classification Tasks

The decoding maps returned by the KO, AKO and DL inference procedures are presented in Figure B.3. 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.

References

Rina Foygel Barber and Emmanuel J. Candès. Controlling the false discovery rate via knockoffs. *The Annals of Statistics*, 43(5):2055–2085, October 2015. ISSN 0090-5364. doi: 10.1214/15-AOS1337. URL `http://arxiv.org/abs/1404.5609`. arXiv: 1404.5609.

Yoav Benjamini and Daniel Yekutieli. The control of the false discovery rate in multiple testing under dependency. *Ann. Statist.*, 29(4):1165–1188, 08 2001. doi: 10.1214/aos/1013699998. URL `https://doi.org/10.1214/aos/1013699998`.

Emmanuel Candès, Yingying Fan, Lucas Janson, and Jinchi 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. doi: 10.1111/rssb.12265. URL `https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/rssb.12265`.

Jaime Roquero Gimenez and James Zou. Improving the stability of the knockoff procedure: Multiple simultaneous knockoffs and entropy maximization. In Kamalika Chaudhuri and Masashi Sugiyama, editors, *Proceedings of Machine Learning Research*, volume 89 of *Proceedings of Machine Learning Research*, pages 2184–2192. PMLR, 16–18 Apr 2019. URL `http://proceedings.mlr.press/v89/gimenez19b.html`.

Christophe Giraud. *Introduction to high-dimensional statistics*. Chapman and Hall/CRC, 2014.

Nicolai Meinshausen, Lukas Meier, and Peter Bühlmann. p-values for high-dimensional regression. *Journal of the American Statistical Association*, 104(488):1671–1681, 2009. doi: 10.1198/jasa.2009.tm08647. URL `https://doi.org/10.1198/jasa.2009.tm08647`.

Joseph P. Romano and Cyrus 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.

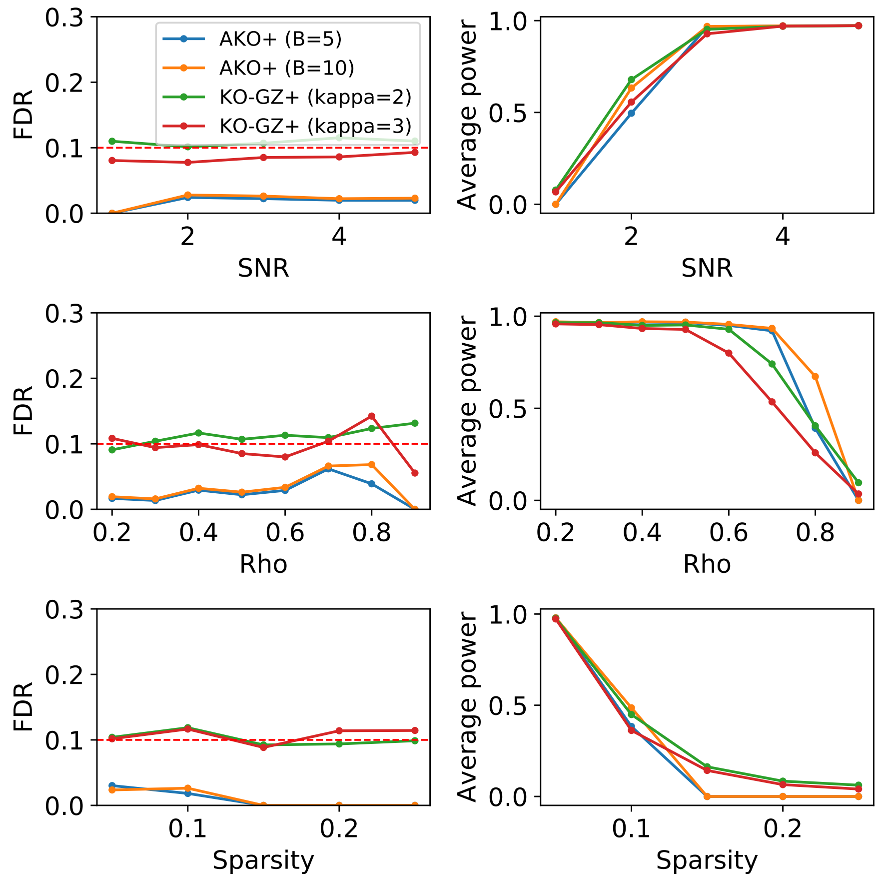

Figure B.1. **Aggregation of multiple knockoffs** ($B = 5$ **and** $B = 10$) **vs. simultaneous knockoffs** ($\kappa = 2$ **and** $\kappa = 3$)**.** A clear loss in statistical power is demonstrated in the latter method when increasing the number of bootstraps κ, while the former (AKO) shows the opposite: with $B = 10$ bootstraps there are small, yet consistent gains in the number of true detections compared to using only $B = 5$ bootstraps.

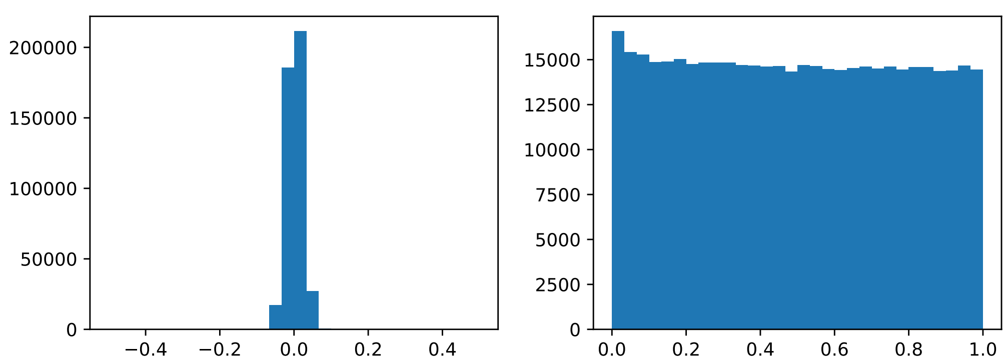

Figure B.2. **Left: Histogram of Spearman correlation values for 100 samples of null aggregated p-values $\bar{\pi}_j$. Right: Histogram of corresponding p-values for the Spearman correlation**

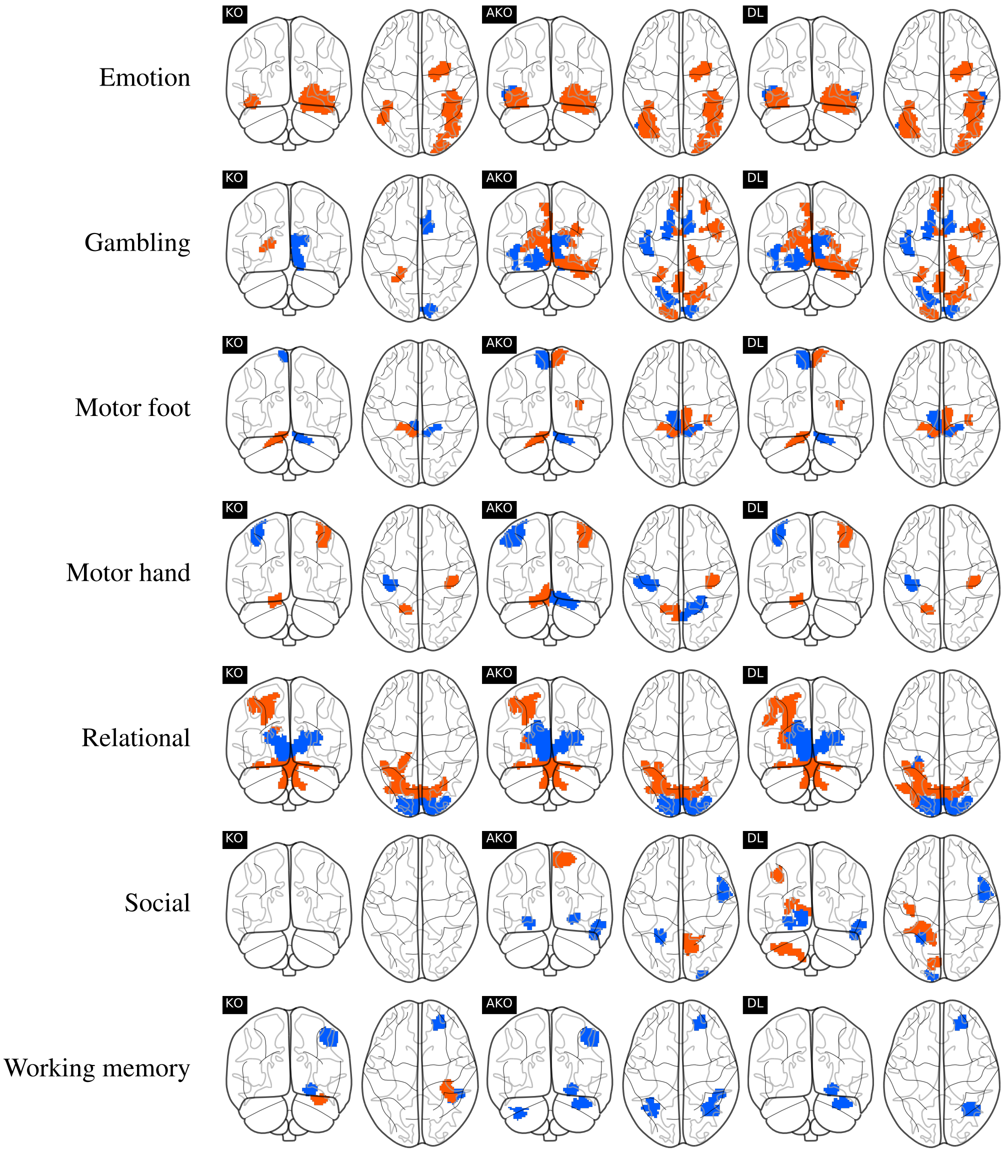

Figure B.3. **Decoding maps obtained for seven classification tasks.** Emotion, gambling, motor foot, motor hand, relational, social and working memory refer to 7 binary tasks that were considered based on the HCP900 dataset. We observe that AKO is typically more sensitive than KO, and yields solution closer to that of an independent solution based on a desparsified-Lasso (DL) estimator.