
Aggregation of Multiple Knockoffs

Tuan-Binh Nguyen^{1,2} Jérôme-Alexis Chevalier² Bertrand Thirion² Sylvain Arlot¹

Abstract

We develop an extension of the knockoff inference procedure, introduced by Barber & 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 (Barber & Candès, 2015; Candès et al., 2018) 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 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, Benjamini & Hochberg 1995), 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 Gimenez & Zou (2019) and by Emery & Keich (2019), 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 (Meinshausen et al., 2009), 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} , $|\mathcal{A}| \triangleq \text{card}(\mathcal{A})$; 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 Barber & Candès (2015).

¹Université Paris-Saclay, CNRS, Inria, Laboratoire de mathématiques d’Orsay, 91405, Orsay, France ²Inria, CEA, Université Paris-Saclay, France. Correspondence to: Tuan-Binh Nguyen <tuan-binh.nguyen@inria.fr>.

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 Barber & Candès (2015) and Candès et al. (2018):

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

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 $\hat{\mathcal{S}}$ 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

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

Specifically, we want to bound the proportion of false positives among selected variables, that is, control the false discovery rate (FDR, Benjamini & Hochberg 1995) under certain predefined level α :

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

Knockoff Inference. Introduced originally by Barber & Candès (2015), the knockoff filter is a variable selection method for multivariate models with theoretical control of FDR. Candès et al. (2018) expanded the method to work in the case of (mildly) high-dimensional data, with the assumption that $\mathbf{x} = (x_1, \dots, x_p) \sim P_X$ 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, Candès et al. 2018). *The model-X knockoffs for the family of random variables $\mathbf{x} = (x_1, \dots, x_p)$ are a new family of random variables $\tilde{\mathbf{x}} = (\tilde{x}_1, \dots, \tilde{x}_p)$ constructed to satisfy the two properties:*

1. For any subset $\mathcal{K} \subset \{1, \dots, p\}$, $(\mathbf{x}, \tilde{\mathbf{x}})_{\text{swap}(\mathcal{K})} \stackrel{d}{=} (\mathbf{x}, \tilde{\mathbf{x}})$, 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 $\stackrel{d}{=}$ denotes equality in distribution.
2. $\tilde{\mathbf{x}} \perp \mathbf{y} \mid \mathbf{x}$.

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, Candès et al. 2018). *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{x}_j switches the sign of W_j :

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

Following previous works on the analysis of the knockoff properties (Arias-Castro & Chen, 2017; Rabinovich et al., 2020), 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 Candès et al. (2018, 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 (Tibshirani, 1996):

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

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

$$\forall j \in [p], \quad W_j = |\hat{\beta}_j| - |\hat{\beta}_{j+p}|. \quad (3)$$

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:

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

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

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 π_j is defined as:*

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

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 Meinshausen et al. (2009):

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

where $q_\gamma(\cdot)$ is the γ -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, Benjamini & Hochberg 1995) to control the FDR.

Definition 4 (BH step-up, Benjamini & Hochberg 1995). *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:

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

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

This procedure controls the FDR, but only under independence or positive-dependence between p-values (Benjamini & Yekutieli, 2001). 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 Benjamini & Yekutieli (2001). 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 Equation (5) does not deviate significantly from independence (details in supplementary material).

Definition 5 (BY step-up, Benjamini & Yekutieli 2001). *Given an ordered list of p-values as in step 1 of BH step-up $\bar{\pi}_{(1)} \leq \bar{\pi}_{(2)} \leq \dots \leq \bar{\pi}_{(p)}$ and predefined level $\alpha \in (0, 1)$, the Benjamini-Yekutieli step-up procedure first finds:*

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

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

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

Blanchard & Roquain (2009) 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: $\hat{\mathcal{S}}_{AKO}$ – Set of selected variables index

for $b = 1$ **to** B **do**

$\tilde{\mathbf{X}}^{(b)} \leftarrow \text{SAMPLING_KNOCKOFF}(\mathbf{X})$

$\mathbf{W}^{(b)} \leftarrow \text{KNOCKOFF_STATISTIC}(\mathbf{X}, \tilde{\mathbf{X}}^{(b)}, \mathbf{y})$

$\boldsymbol{\pi}^{(b)} \leftarrow \text{CONVERT_STATISTIC}(\mathbf{W}^{(b)})$ // Using

Eq. (5)

end for

for $j = 1$ **to** p **do**

$\bar{\pi}_j \leftarrow \text{QUANTILE_AGGREGATION}(\{\pi_j^{(b)}\}_{b=1}^B)$ //

Using Eq. (6)

end for

$\hat{k} \leftarrow \text{FDR_THRESHOLD}(\alpha, (\bar{\pi}_1, \bar{\pi}_2, \dots, \bar{\pi}_p))$ //

Using either Eq. (7) or Eq. (8)

Return: $\hat{\mathcal{S}}_{AKO} \leftarrow \{j \in [p] : \bar{\pi}_j \leq \bar{\pi}_{\hat{k}}\}$

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 Su et al. (2015) 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 Barber & Candès 2015). 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 Holden & Helton (2018) 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, Gimenez & Zou (2019) and Emery & Keich (2019) 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 $\mathcal{O}(B^3 p^3)$, while for AKO, runtime is only of $\mathcal{O}(B p^3)$ 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 Barber & Candès (2015) is used. We have set up additional experiments in supplementary material to illustrate this phenomenon. In addition, the threshold introduced by Emery & Keich (2019) 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 ($B = 1, \gamma = 1$) and Vanilla Knockoff

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

Proposition 1 (Proof in supplementary material). *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 Barber & Candès (2015).

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 supplementary material). *Define the equi-correlation set as:*

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

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

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

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 π_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 π_j defined by Eq. (5) satisfies:*

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

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

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 π_j , when $W_j > 0$ we have:

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

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 $\{W_k\}_{k \in \mathcal{S}^c \setminus \{j\}}$. Therefore, for every $t \in [0, 1)$,

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

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 $F_0^{-1}(U), F_0^{-1}(U_1), \dots, F_0^{-1}(U_m)$ where U, U_1, \dots, 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. (12) can be rewritten as

$$\mathbb{P}(\pi_j \leq t) \leq \mathbb{P}\left(\frac{m}{p} \widetilde{F}_m(F_0^{-1}(U)) + \frac{1}{p} \leq t\right) \quad (13)$$

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

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. (13) shows

that

$$\begin{aligned} \mathbb{P}(\pi_j \leq t) &\leq \mathbb{P}\left(m \widehat{G}_m(U) \leq tp - 1\right) \\ &= \int_0^1 \mathbb{P}\left(m \widehat{G}_m(u) \leq tp - 1\right) du. \end{aligned} \quad (14)$$

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 (Boucheron et al., 2013, Eq. 2.10) shows that for every $0 \leq x \leq u \leq 1$,

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

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

$$\forall w \geq \frac{1-\lambda}{1-7\lambda} \geq 1, \quad \frac{w-1}{7w-1} \geq \lambda$$

hence $\forall u \geq x \frac{1-\lambda}{1-7\lambda}$,

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

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

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

Taking $x = (tp - 1)/m$, we obtain from Eq. (14) that $\forall \lambda \in (0, 1/7)$

$$\begin{aligned} \mathbb{P}(\pi_j \leq t) &\leq \frac{1-\lambda}{1-7\lambda} \frac{tp-1}{m} + \frac{1}{3m\lambda} \\ &= \frac{1-\lambda}{1-7\lambda} \frac{tp}{m} + \left(\frac{1}{3\lambda} - \frac{1-\lambda}{1-7\lambda}\right) \frac{1}{m}. \end{aligned} \quad (15)$$

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 π_j is replaced by

$$\pi_{j,c} \triangleq \begin{cases} \frac{c + \#\{k : W_k \leq -W_j\}}{p} & \text{if } W_j > 0 \\ 1 & \text{if } W_j \leq 0 \end{cases} \quad (16)$$

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

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

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 supplementary material that if $p \rightarrow +\infty$ with $|\mathcal{S}| \ll p$, then for every $j \geq 1$ such that $\beta_j^* = 0$, π_j is an asymptotically valid p-value, that is,

$$\forall t \in [0, 1], \quad \limsup_{p \rightarrow +\infty} \mathbb{P}(\pi_j \leq t) \leq t. \quad (17)$$

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{S}_{AKO+BY} of aggregation of multiple knockoff (Algorithm 1), with the BY step-up procedure, has a FDR controlled as follows:*

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

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

Sketch of the proof. The proof of Meinshausen et al. (2009, Theorem 3.3), which itself relies partly on Benjamini & Yekutieli (2001), 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 supplementary material. \square

Note that Theorem 1 loses a factor κ compared to the nominal FDR level α . This can be solved by changing α into α/κ 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 α .

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 Gimenez & Zou (2019) (KO-GZ), along with its variant of Emery & Keich (2019) (KO-EK); the *knockoff statistics aggregation* by Holden & Helton (2018) (KO-HH); and *debiased Lasso* (DL-BH) (Javanmard & Javadi, 2019).

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 $\Sigma \in \mathbb{R}^{p \times p}$. We generate Σ as a symmetric Toeplitz matrix that has the structure:

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

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 β^* vector is picked with a sparsity parameter that controls the proportion of non-zero elements with amplitude 1. The noise ϵ is generated to follow $\mathcal{N}(\mu, \mathbf{I}_n)$ with its magnitude $\sigma = \|\mathbf{X}\beta^*\|_2 / (\text{SNR} \|\epsilon\|_2)$ 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 ρ , 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.

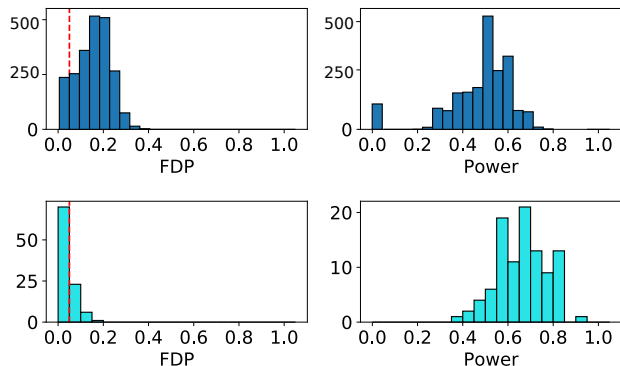


Figure 1. Histogram of FDP and power for 2500 runs of KO (blue, top row) vs. 100 runs of AKO with $B = 25$ (teal, bottom row) under the same simulation. Simulation parameter: $\text{SNR} = 3.0, \rho = 0.5, \text{sparsity} = 0.06$. FDR is controlled at level $\alpha = 0.05$.

Inference Results on Different Simulation Settings. To observe how each algorithm performs under various scenarios, 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 ρ) 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 ρ gets bigger than 0.6. Debiased Lasso (DL-BH) and KO-EK control FDR well in all scenarios, but are the two most conservative procedures.

Choice of B and γ for AKO. Figure 3 shows an experiment when varying γ and B . FDR and power are averaged across 30 simulations of fixed parameters: $\text{SNR}=3.0, \rho = 0.7, \text{sparsity}=0.06$. Notably, it seems that there is no further gain in statistical power when $B > 25$. Similarly, the power is essentially equal for γ values greater than 0.1 when $B \geq 25$. Based on the results of this experiment we set the default value of $B = 25, \gamma = 0.3$.

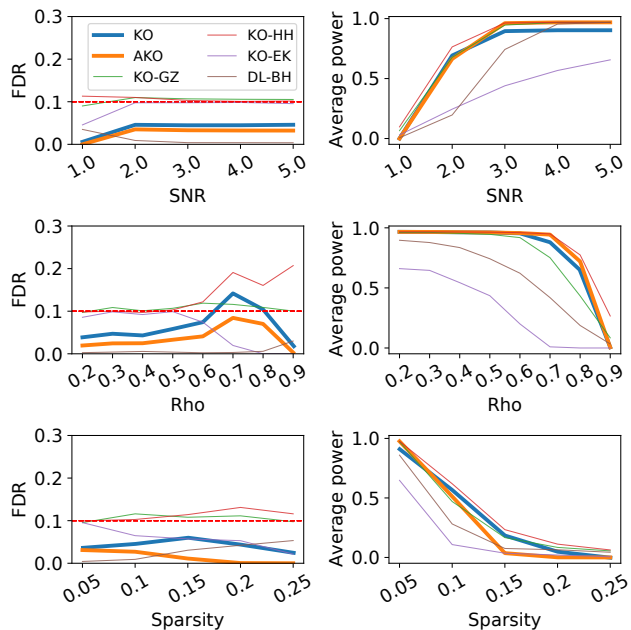


Figure 2. FDR (left) and average power (right) of several methods for 100 runs with varying simulation parameters. For each varying parameter, we keep the other ones at default value: $\text{SNR} = 3.0, \rho = 0.5, \text{sparsity} = 0.06$. FDR is controlled at level $\alpha = 0.1$. The benchmarked methods are: aggregation of multiple knockoffs (AKO – ours); vanilla knockoff (KO); simultaneous knockoff by Gimenez & Zou 2019 (KO-GZ) and by Emery & Keich 2019 (KO-EK); knockoff statistics aggregation (KO-HH); debiased-Lasso (DL-BH).

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 Atwell et al. (2010). Preprocessing is done similarly to Azencott et al. (2013): 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 Slim et al. (2019) to make the final design matrix of size $n = 166 \text{ samples} \times p = 1560 \text{ features}$. 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 Kim et al. (2008) to be vital for regulating period length in the *Arabidopsis* circadian clock. FIONA1 also appears to be involved in photoperiod-dependent flowering and in daylength-dependent seedling

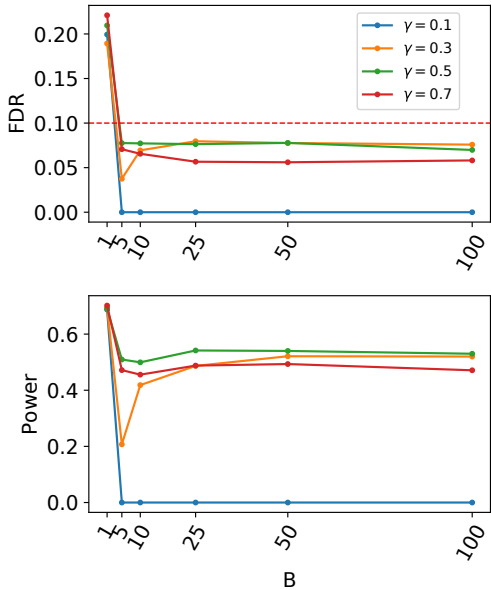


Figure 3. FDR and average power for 30 simulations of fixed parameters: SNR=3.0, $\rho = 0.7$, sparsity=0.06. There is virtually no gain in statistical power when $B > 25$ and when $\gamma \geq 0.1$.

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	—

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 (Silverstone et al., 1998) that AT4G02780 encodes a mutation for late flowering. Meanwhile, AT5G47640 mutant delay flowering in long-day but not in short-day experiments (Cai et al., 2007).

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 supplementary material. 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.

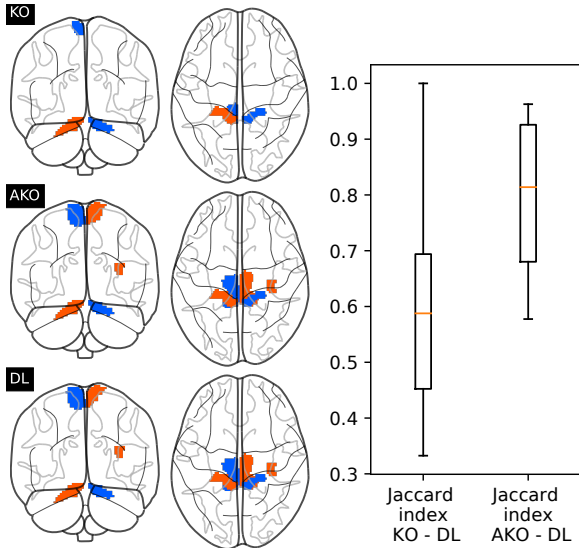


Figure 4. Detection of significant brain regions for HCP data (900 subjects). (left) Selected regions in a left or right foot movement task. **Orange**: brain areas with positive sign activation. **Blue**: brain areas with negative sign activation. Here the AKO solution recovers an anatomically correct 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 Barber & Candès (2015) no longer holds (see supplementary material); *ii*) a factor κ 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 κ . 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 Meinshausen et al. (2009) 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 Candès et al. 2018) 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 (Samson, 2000; Merlevède et al., 2009; Hang & Steinwart, 2017, among others). Similarly, the proof of Eq. (17) 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, Abdesselam, 2018, 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 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 \hat{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.2 in supplementary material) 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 (Benjamini & Yekutieli, 2001) 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 α (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 γ 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 γ , adaptive solutions have been proposed (Meinshausen et al., 2009), 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

- Abdesselam, A. The weakly dependent strong law of large numbers revisited. *arXiv e-prints*, art. arXiv:1801.09265, January 2018.
- Arias-Castro, E. and Chen, S. Distribution-free multiple testing. *Electron. J. Statist.*, 11(1):1983–2001, 2017. doi: 10.1214/17-EJS1277. URL <https://doi.org/10.1214/17-EJS1277>.

- Atwell, S., Huang, Y. S., Vilhjálmsón, B. J., Willems, G., Horton, M., Li, Y., Meng, D., Platt, A., Tarone, A. M., Hu, T. T., et al. Genome-wide association study of 107 phenotypes in arabidopsis thaliana inbred lines. *Nature*, 465(7298):627, 2010.
- Azencott, C.-A., Grimm, D., Sugiyama, M., Kawahara, Y., and Borgwardt, K. M. Efficient network-guided multi-locus association mapping with graph cuts. *Bioinformatics*, 29(13):i171–i179, 2013.
- Barber, R. F. and Candès, E. J. 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.
- Benjamini, Y. and Hochberg, Y. 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. ISSN 0035-9246. URL <https://www.jstor.org/stable/2346101>.
- Benjamini, Y. and Yekutieli, D. 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>.
- Blanchard, G. and Roquain, É. Adaptive false discovery rate control under independence and dependence. *Journal of Machine Learning Research*, 10(Dec):2837–2871, 2009.
- Boucheron, S., Lugosi, G., and Massart, P. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, Oxford, 2013. ISBN 978-0-19-953525-5.
- Cai, X., Ballif, J., Endo, S., Davis, E., Liang, M., Chen, D., DeWald, D., Kreps, J., Zhu, T., and Wu, Y. A putative ccaat-binding transcription factor is a regulator of flowering timing in arabidopsis. *Plant Physiology*, 145(1): 98–105, 2007. ISSN 0032-0889. doi: 10.1104/pp.107.102079. URL <http://www.plantphysiol.org/content/145/1/98>.
- Candès, E., Fan, Y., Janson, L., and Lv, J. 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>.
- Emery, K. and Keich, U. Controlling the FDR in variable selection via multiple knockoffs. *arXiv e-prints*, art. arXiv:1911.09442, November 2019.
- Gimenez, J. R. and Zou, J. Improving the stability of the knockoff procedure: Multiple simultaneous knockoffs and entropy maximization. In Chaudhuri, K. and Sugiyama, M. (eds.), *Proceedings of Machine Learning Research*, volume 89 of *Proceedings of Machine Learning Research*, pp. 2184–2192. PMLR, 16–18 Apr 2019. URL <http://proceedings.mlr.press/v89/gimenez19b.html>.
- Hang, H. and Steinwart, I. A Bernstein-type inequality for some mixing processes and dynamical systems with an application to learning. *Ann. Statist.*, 45(2):708–743, 2017. ISSN 0090-5364. doi: 10.1214/16-AOS1465. URL <https://doi.org/10.1214/16-AOS1465>.
- Holden, L. and Helton, K. H. Multiple model-free knockoffs. *arXiv preprint arXiv:1812.04928*, 2018.
- Javanmard, A. and Javadi, H. False discovery rate control via debiased lasso. *Electron. J. Statist.*, 13(1):1212–1253, 2019. doi: 10.1214/19-EJS1554. URL <https://doi.org/10.1214/19-EJS1554>.
- Kim, J., Kim, Y., Yeom, M., Kim, J.-H., and Nam, H. G. Fion1 is essential for regulating period length in the arabidopsis circadian clock. *The Plant Cell*, 20(2): 307–319, 2008. ISSN 1040-4651. doi: 10.1105/tpc.107.055715. URL <http://www.plantcell.org/content/20/2/307>.
- Meinshausen, N., Meier, L., and Bühlmann, P. 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>.
- Merlevède, F., Peligrad, M., and Rio, E. 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.*, pp. 273–292. Inst. Math. Statist., Beachwood, OH, 2009. doi: 10.1214/09-IMSCOLL518. URL <https://doi.org/10.1214/09-IMSCOLL518>.
- Rabinovich, M., Ramdas, A., Jordan, M. I., and Wainwright, M. J. Optimal Rates and Tradeoffs in Multiple Testing. *Statistica Sinica*, 2020. ISSN 10170405. doi: 10.5705/ss.202017.0468. URL http://www3.stat.sinica.edu.tw/ss_newpaper/SS-2017-0468_na.pdf.
- Samson, P.-M. Concentration of measure inequalities for Markov chains and Φ -mixing processes. *Ann. Probab.*, 28(1):416–461, 2000. ISSN 0091-1798.

Silverstone, A. L., Ciampaglio, C. N., and Sun, T.-p. The arabidopsis rga gene encodes a transcriptional regulator repressing the gibberellin signal transduction pathway. *The Plant Cell*, 10(2):155–169, 1998. ISSN 1040-4651. doi: 10.1105/tpc.10.2.155. URL <http://www.plantcell.org/content/10/2/155>.

Slim, L., Chatelain, C., Azencott, C.-A., and Vert, J.-P. kernelpsi: a post-selection inference framework for non-linear variable selection. In *International Conference on Machine Learning*, pp. 5857–5865, 2019.

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

Tibshirani, R. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58, 1996. URL <http://www.jstor.org/stable/2346178>.