Contrôle du taux de Fausses Découvertes pour les Knockoffs agrégés

Alexandre Blain [1] & Bertrand Thirion [2] & Olivier Grisel [3] & Pierre Neuvial [4]

[1] *INRIA, Université Paris-Saclay,* alexandre.blain@inria.fr
[2] *INRIA, CEA,* bertrand.thirion@inria.fr
[3] *INRIA,* olivier.grisel@inria.fr
[4] *Institut de Mathématiques de Toulouse, Université de Toulouse,*
pierre.neuvial@math.univ-toulouse.fr

Résumé. La sélection de variables contrôlée est une étape importante dans divers domaines scientifiques, tels que l'imagerie cérébrale ou la génomique. Dans ces contextes de données de haute dimension, considérer trop de variables conduit à des modèles médiocres et à des coûts élevés, d'où la nécessité de garanties statistiques sur les faux positifs. Les Knockoffs sont un outil statistique populaire pour la sélection conditionnelle de variables en haute dimension. Cependant, ils contrôlent l'espérance de la proportion de fausses découvertes (FDR) et non leur proportion réelle (FDP). Nous présentons une nouvelle méthode, KOPI, qui exploite la notion d'inférence post hoc pour contrôler les quantiles du FDP pour l'inférence basée sur les Knockoffs. La méthode proposée repose également sur un nouveau type d'agrégation pour contrer le caractère aléatoire indésirable associé à l'inférence Knockoff classique. Nous démontrons le contrôle du FDP et des gains de puissance substantiels par rapport aux méthodes basées sur les Knockoffs existantes dans divers contextes de simulation et obtenons de bons compromis sensibilité/spécificité sur des données d'imagerie cérébrale et génomique. Ce travail a fait l'objet d'un poster à la conférence NeurIPS 2023: `https://arxiv.org/abs/2310.10373`.

Mots-clés. Sélection de variable contrôlée, inférence Knockoffs, IRMf, génomique

Abstract. Controlled variable selection is an important analytical step in various scientific fields, such as brain imaging or genomics. In these high-dimensional data settings, considering too many variables leads to poor models and high costs, hence the need for statistical guarantees on false positives. Knockoffs are a popular statistical tool for conditional variable selection in high dimension. However, they control for the expected proportion of false discoveries (FDR) and not their actual proportion (FDP). We present a new method, KOPI, that controls the proportion of false discoveries for Knockoff-based inference. The proposed method also relies on a new type of aggregation to address the undesirable randomness associated with classical Knockoff inference. We demonstrate FDP control and substantial power gains over existing Knockoff-based methods in various simulation settings and achieve good sensitivity/specificity tradeoffs on brain imaging and genomic data. This work was pubished at NeurIPS 2023: `https://arxiv.org/abs/2310.10373`.

Keywords. Controlled variable selection, Knockoffs inference, fMRI, genomics

1 Inférence Knockoffs

Notation. Nous désignons les vecteurs par des lettres minuscules en gras. Un vecteur $\mathbf{x} = \{x_1, \ldots, x_p\}$ duquel nous avons retiré la j^{me} coordonnée est désigné par $\mathbf{x}_{-j}$, c'est-à-dire $\mathbf{x} \setminus \{x_j\}$. L'indépendance entre deux vecteurs aléatoires $\mathbf{x}$ et $\mathbf{y}$ est notée par $\mathbf{x} \perp \mathbf{y}$. Pour deux vecteurs $\mathbf{x}$ et $\tilde{\mathbf{x}}$ et un sous-ensemble S d'indices, $(\mathbf{x}, \tilde{\mathbf{x}})_{swap(S)}$ désigne le vecteur obtenu à partir de $(\mathbf{x}, \tilde{\mathbf{x}})$ en échangeant les entrées x_j et $\tilde{x}_j$ pour chaque $j \in S$. Les matrices sont notées par des lettres majuscules en gras, la seule exception étant le vecteur des statistiques Knockoff que nous notons par $\mathbf{W}$ comme dans [1, 6]. Pour tout ensemble S, $|S|$ désigne le cardinal de S. Pour un vecteur $\mathbf{z} = (z_j)_{1 \leq j \leq p}$ et $S \subset [\![p]\!]$, nous notons par $z_{(j:S)}$ (ou $z_{(j)}$ lorsqu'il n'y a pas d'ambiguïté) la j^{me} plus petite valeur dans le sous-vecteur $(\mathbf{z}_s)_{s \in S}$. Pour un entier k, $[\![k]\!]$ désigne l'ensemble $\{1, \ldots, k\}$. L'égalité en distribution est notée par $\overset{d}{=}$.

Le problème de sélection de variable conditionnelle. Les données d'entrée sont notées par $\mathbf{X} \in \mathbb{R}^{n \times p}$, où n est le nombre d'observations et p le nombre de variables. La variable d'intérêt est noté par $\mathbf{y} \in \mathbb{R}^n$. L'objectif est de sélectionner les variables qui sont associées à la variable d'intérêt *conditionnellement à toutes les autres*. Formellement, nous testons simultanément pour tout $j \in [\![p]\!]$:

$$H_{0,j} : y \perp x_j | \mathbf{x}_{-j} \quad \text{contre} \quad H_{1,j} : y \not\perp x_j | \mathbf{x}_{-j}.$$

La sortie d'une méthode de sélection de variables est un ensemble de rejet $\hat{S} \subset [\![p]\!]$ qui estime le support inconnu vrai $\mathcal{H}_1 = \{j : y \not\perp x_j | \mathbf{x}_{-j}\}$. Son complément est l'ensemble des vraies hypothèses nulles $\mathcal{H}_0 = \{j : y \perp x_j | \mathbf{x}_{-j}\}$. On note $p_0 = |\mathcal{H}_0|$. Pour assurer une inférence fiable, notre objectif est de fournir une garantie statistique sur la proportion de fausses découvertes dans $\hat{S}$. La Proportion de Fausses Découvertes (FDP) et le Taux de Fausse Découvertes (FDR) [2] sont définis comme suit :

$$\mathrm{FDP}(\hat{S}) = \frac{|\hat{S} \cap \mathcal{H}_0|}{|\hat{S}| \vee 1}, \quad \mathrm{FDR}(\hat{S}) = \mathbb{E}[\mathrm{FDP}(\hat{S})] = \mathbb{E}\left[\frac{|\hat{S} \cap \mathcal{H}_0|}{|\hat{S}| \vee 1}\right].$$

Pour contrôler le FDP, nous utilisons la notion d'inférence post hoc introduite par [7]. Une borne supérieure post hoc de niveau α pour le FDP est une fonction V qui vérifie :

$$\mathbb{P}\left(\forall S \subset [\![p]\!], \mathrm{FDP}(S) \leq V(S)/|S|\right) \geq 1 - \alpha\,.$$

Knockoffs. Le filtre Knockoff est une technique de sélection de variables introduite par [1] et affinée par [6] qui contrôle le FDR. Cette procédure repose sur la construction de copies bruitées des variables originales appelées variables Knockoff, qui sont conçues pour servir de contrôles pour la sélection de variables.

Definition 1 (Model-X Knockoffs, 6)**.** Pour la famille de variables aléatoires $\mathbf{x} = (x_1, \ldots, x_p)$, les Knockoffs sont une nouvelle famille de variables aléatoires $\tilde{\mathbf{x}} = (\tilde{x}_1, \ldots, \tilde{x}_p)$ satisfaisant :

1. pour tout $S \subset [\![p]\!]$, $(\mathbf{x}, \tilde{\mathbf{x}})_{swap(S)} \overset{d}{=} (\mathbf{x}, \tilde{\mathbf{x}})$

2. $\tilde{\mathbf{x}} \perp \mathbf{y} | \mathbf{x}$.

Une fois que nous disposons de telles variables, nous quantifions leur importance par rapport aux originales. Cela est fait en calculant les statistiques Knockoff $\mathbf{W} = (W_1, \ldots, W_p)$ qui sont définies comme suit.

Definition 2 (Statistique Knockoff, 6)**.** Une statistique Knockoff $\mathbf{W} = (W_1, \ldots, W_p)$ est une mesure de l'importance des caractéristiques qui satisfait :

1. $\mathbf{W}$ dépend uniquement de $\mathbf{X}, \tilde{\mathbf{X}}$ et $\mathbf{y}$: $\mathbf{W} = g(\mathbf{X}, \tilde{\mathbf{X}}, \mathbf{y})$.
2. Échanger la colonne $\mathbf{x}_j$ et sa colonne knockoff $\tilde{\mathbf{x}}_j$ inverse le signe de W_j:

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

La statistique Knockoff la plus couramment utilisée est la différence des coefficients Lasso (LCD) [18]. Cette statistique est obtenue en ajustant un estimateur Lasso [15] sur $[\mathbf{X}, \tilde{\mathbf{X}}] \in \mathbb{R}^{n \times 2p}$, ce qui donne $\widehat{\boldsymbol{\beta}} \in \mathbb{R}^{2p}$. Ensuite, la statistique Knockoff peut être calculée en utilisant $\widehat{\boldsymbol{\beta}}$:

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

Ce coefficient résume l'importance de la j^{me} variable originale par rapport à sa propre Knockoff : $W_j > 0$ indique que la variable originale est plus importante pour ajuster y que la variable Knockoff, signifiant que la j^{me} variable est probablement pertinente. Inversement, $W_j < 0$ indique que la j^{me} variable est probablement non pertinente. Nous souhaitons donc sélectionner les variables correspondant aux W_j grands et positifs. Formellement, l'ensemble de rejet $\hat{S}$ peut être écrit $\hat{S} = \{j : W_j > T_q\}$, où T_q est choisi pour contrôler de manière prouvée le FDR au niveau q [6].

Schémas d'agrégation. En raison de l'aléa inhérent au processus de génération des knockoffs, différentes variables peuvent être sélectionnées pour deux exécutions différentes de la méthode. Pour atténuer cela, l'agrégation de plusieurs tirages de Knockoffs est nécessaire. Ren and Barber [14] a introduit un schéma d'agrégation qui repose sur la définition des e-values Knockoffs.

$$e_j = \frac{p}{1 + |\{k : W_k \leq -T_q\}|} 1_{\{W_j \geq T_q\}}.$$

Ces e-values peuvent être moyennées sur D tirages et la procédure e-BH [16] est effectuée pour la sélection de variables. Alternativement, [13] définit la π-statistique suivante, qui quantifie les preuves contre une variable :

$$\pi_j = \begin{cases} \frac{1+|\{k:W_k \leq -W_j\}|}{p} \text{ si} & W_j > 0 \\ 1 \quad \text{si} & W_j \leq 0. \end{cases} \tag{1}$$

Dans [13] les π-statistiques sont traitées comme des p-valeurs et agrégées en utilisant l'agrégation quantile [10]. Cependant, elles ne peuvent être considérées comme des p-valeurs que sous des hypothèses restrictives difficiles à vérifier. Dans la section suivante, ces statistiques sont utilisées comme un bloc de construction pour atteindre le contrôle du FDP. Le cadre KOPI ne nécessite pas que les π-statistiques soient des p-valeurs valides.

2 KOPI: Contrôle du taux de Fausses Découvertes pour les Knockoffs agrégés

La méthode que nous proposons vise à résoudre les deux problèmes principaux des méthodes Knockoffs existantes: *i)* le contrôle du FDR et non du FDP, qui ne sont pas des quantités équivalentes [11] et *ii)* le caractère aléatoire de l'inférence Knockoffs dû au processus de génération des Knockoffs.

2.1 Contrôle post hoc du FDP pour les π-statistiques

Pour obtenir un contrôle du FDP, nous nous appuyons sur le contrôle du Joint Error Rate (JER) tel qu'introduit dans [5]. Pour $k_{max} \in [\![p]\!]$, nous définissons une *famille de seuils* de taille k_{max} comme un vecteur $\mathbf{t} = (t_j)_{j \in [\![k_{max}]\!]}$ tel que $0 \leq t_1 \leq \cdots \leq t_{k_{max}} \leq 1$.

Definition 3 (Joint Error Rate, 5)**.** Soit $\pi_{(j:\mathcal{H}_0)}$ la j^{me} plus petite valeur π_j parmi toutes les hypothèses nulles. Le JER associé à $\mathbf{t} = (t_j)_{j \in [\![k_{max}]\!]}$ est :

$$\mathrm{JER}(\mathbf{t}) = \mathbb{P}\left(\exists j \in [\![k_{max} \wedge p_0]\!] : \pi_{(j:\mathcal{H}_0)} < t_j\right). \tag{2}$$

On dit que la famille de seuils $\mathbf{t}$ contrôle le JER au niveau α ssi $\mathrm{JER}(\mathbf{t}) \leq \alpha$.

Une borne supérieure de niveau α pour le FDP peut être déduite du contrôle du JER via le résultat suivant :

Proposition 1 (Contrôle du FDP via contrôle du JER 5)**.** *Si* $\mathbf{t}$ *est une famille de seuils de taille* k_{max} *qui contrôle le* JER *au niveau* α*, alors,* $V^{\mathbf{t}}(S)/|S|$ *est une borne supérieure de niveau* α *pour le FDP, avec :*

$$V^{\mathbf{t}}(S) = \min_{1 \leq k \leq k_{max}} (k-1) + \sum_{i \in S} 1_{\{\pi_i > t_k\}}. \tag{3}$$

Dans la suite de cette section, nous montrons comment contrôler le JER dans le cas des π-statistiques.

2.2 Distribution conjointe des π-statistiques sous l'hypothèse nulle

Par la Définition 3, le $\mathrm{JER}(\mathbf{t})$ d'une famille de seuils donnée dépend uniquement de la distribution conjointe nulle des π-statistiques. Comme pour les résultats antérieurs de contrôle

du FDR [1] ou de contrôle du k-FWER [8], l'idée clé pour obtenir un contrôle du JER pour les π-statistiques est de prouver que la partie pertinente de cette distribution est en fait connue, grâce aux propriétés des statistiques knockoff. Nous utilisons la même notation que dans [8]. Soit $Z_j = |\{k \in [\![p]\!] : W_k \leq -W_j\}|$ et $\chi_j = sign(W_j)$, les π-statistiques $(\pi_j)_{j=[\![p]\!]}$ sont alors données par :

$$\pi_j = \frac{1+Z_j}{p} 1_{\{\chi_j=1\}} + 1_{\{\chi_j=-1\}}.$$

Pour un $\mathbf{W}$ donné, soit $\sigma(\mathbf{W})$ être une permutation de $[\![p]\!]$ qui trie $\mathbf{W}$ par module décroissant : $\sigma(\mathbf{W}) = (\sigma_1, \ldots, \sigma_p)$ tel que $|W_{\sigma_1}| \geq |W_{\sigma_2}| \cdots \geq |W_{\sigma_p}|$. Nous commençons par prouver que les statistiques Z peuvent être exprimées comme une fonction du vecteur de statistiques χ :

Lemma 1. *Pour $j \in [\![p]\!]$ tel que $\chi_{\sigma_j} = 1$, $Z_{\sigma_j} = \sum_{k=1}^{j-1} 1_{\{\chi_{\sigma_k}=-1\}}$.*

Preuve du Lemme 1. Puisque $\chi_{\sigma_j} = 1$, nous avons :

$$\begin{aligned}
Z_{\sigma_j} &= |\left\{k \in [\![p]\!] : W_{\sigma_k} \leq -W_{\sigma_j}\right\}| \\
&= |\left\{k \in [\![p]\!] : W_{\sigma_k} < 0 \text{ et } W_{\sigma_k} \leq -W_{\sigma_j}\right\}| \\
&= |\left\{k \in [\![p]\!] : W_{\sigma_k} < 0 \text{ et } |W_{\sigma_k}| \leq |W_{\sigma_j}|\right\}| \\
&= |\{k \in [\![p]\!] : W_{\sigma_k} < 0 \text{ et } k \leq j\}| \\
&= \sum_{k=1}^{j-1} 1_{\{\chi_{\sigma_k}=-1\}}.
\end{aligned}$$

□

Le Lemme 1 implique que la distribution des statistiques d'ordre de $\pi|\sigma(\mathbf{W})$ est entièrement déterminée par celle de $\chi|\sigma(\mathbf{W})$. Pour formaliser cela, nous introduisons les statistiques π^0.

Definition 4 (Statistiques π^0)**.** Soit $\chi^0 = (\chi_j^0)_{1\leq j\leq p}$ une collection de p variables aléatoires de Rademacher i.i.d., c'est-à-dire, pour tout j, $\mathbb{P}(\chi_j^0 = 1) = \mathbb{P}(\chi_j^0 = -1) = 1/2$. Les statistiques π^0 associées sont définies pour $j \in [\![p]\!]$ par

$$\pi_j^0 = \frac{1+Z_j^0}{p} 1_{\{\chi_j^0=1\}} + 1_{\{\chi_j^0=-1\}}, \text{ où } Z_j^0 = \sum_{k=1}^{j-1} 1_{\{\chi_k^0=-1\}}. \tag{4}$$

Theorem 1. *Soit $\mathbf{t}$ une famille de seuils de taille k_{max}. Alors, pour $\pi^0 = (\pi_j^0)_{j\in[\![p]\!]}$ comme dans* (4)*,*

$$\mathrm{JER}(\mathbf{t}) \leq \mathrm{JER}^0(\mathbf{t}) := \mathbb{P}\left(\exists k \in [\![k_{max}]\!] : \pi_{(k)}^0 < t_k\right). \tag{5}$$

Le Théorème 1 – prouvé dans le papier [3] – est lié au Lemme 3.1 de Janson and Su [8] et au Lemme 3.1 de Li et al. [9], qui s'appuient sur la propriété de changement de signe des statistiques Knockoff sous l'hypothèse nulle [1]. L'intérêt du Théorème 1 est que la borne supérieure $\mathrm{JER}^0(\mathbf{t})$ dépend uniquement des statistiques π^0 et de la famille de seuils $\mathbf{t}$, et non des données originales. Par conséquent, elle peut être estimée avec une précision arbitraire pour tout $\mathbf{t}$ donné en utilisant une simulation Monte-Carlo, comme expliqué dans la section suivante et décrit dans dans le Supp. Mat du papier [3].

2.3 Contrôle du Joint Error Rate pour les π-statistiques par calibration

Pour approcher la borne supérieure du JER dérivée dans le Théorème 1, nous tirons B échantillons Monte-Carlo en utilisant l'Algorithme 1 du papier [3]. On obtient ainsi un ensemble de B vecteurs de statistiques π^0 notés par $\pi^0_b \in \mathbb{R}^p$ pour chaque $b \in [\![B]\!]$. Cela nous permet d'évaluer le JER empirique, qui estime la borne supérieure d'intérêt.

Definition 5 (JER Empirique)**.** Pour B vecteurs de statistiques π^0 et une famille de seuils $\mathbf{t}$, le JER empirique est défini comme :

$$\widehat{\mathrm{JER}}^0_B(\mathbf{t}) = \frac{1}{B}\sum_{b=1}^{B} 1\left\{\exists k \in [\![k_{max}]\!] : \pi^0_{b(k)} < t_k\right\}, \tag{6}$$

où pour chaque $b \in [\![B]\!]$, $\pi^0_{b(1)} \leq \cdots \leq \pi^0_{b(p)}$.

Puisque $\widehat{\mathrm{JER}}^0_B(\mathbf{t})$ peut être rendu arbitrairement proche (en choisissant B assez grand) de $\widehat{\mathrm{JER}}^0(\mathbf{t})$ pour toute famille de seuils $\mathbf{t}$ donnée, il reste à choisir $\mathbf{t}$ tel que $\widehat{\mathrm{JER}}^0(\mathbf{t}) \leq \alpha$ afin d'assurer un contrôle du JER. À cette fin, nous considérons un ensemble trié de familles de seuils candidates appelé un *template* :

Definition 6 (Template [5])**.** Un template est une fonction non décroissante composant par composant $\mathbf{T} : [0,1] \mapsto \mathbb{R}^p$ qui associe un paramètre $\lambda \in [0,1]$ à une famille de seuils $\mathbf{T}(\lambda) \in \mathbb{R}^p$.

Cette définition peut naturellement s'étendre au cas de templates contenant un nombre fini de familles de seuils. Le template correspondant à B' familles de seuils est alors noté par $(\mathbf{T}(b'/B'))_{b' \in [\![B']\!]}$.

Une fois un template spécifié, la procédure de *calibration* [5] peut être appli ; cela consiste à trouver la famille de seuils la moins conservatrice $\mathbf{t}$ du template, parmi celles qui contrôlent le JER empirique au niveau α. Formellement, nous considérons la famille de seuils définie $\mathbf{t}^B_\alpha = \mathbf{T}(\lambda_B(\alpha))$, où

$$\lambda_B(\alpha) = \frac{1}{B'}\max\left\{b' \in [\![B']\!] \quad s.t. \quad \widehat{\mathrm{JER}}^0_B\left(\mathbf{T}\left(\frac{b'}{B'}\right)\right) \leq \alpha\right\}.$$

Comme observé par Blain et al. [4], une puissance optimale est atteinte lorsque les familles candidates correspondent à la forme de la distribution des statistiques nulles. Nous définissons un template basé sur la distribution des statistiques π^0 apparaissant dans le Théorème 1. En pratique, nous tirons B' échantillons de cette distribution indépendamment des B échantillons Monte Carlo pour éviter les biais de circularité. Puisqu'un template doit être non décroissant composant par composant, c'est-à-dire que l'ensemble des familles de seuils candidates doit être trié, nous extrayons des quantiles empiriques de ces B' vecteurs triés. Cela produit un template $\mathbf{T}^0$ composé de B' courbes candidates qui correspondent aux quantiles de la distribution des statistiques π^0. La courbe de quantile $\frac{b'}{B'}$ définit la famille de seuils $\mathbf{T}^0(b'/B')$. Nous obtenons le résultat suivant :

Theorem 2 (Contrôle du JER pour les π-statistiques)**.** *Considérons la famille de seuils définie par* $\mathbf{t}_\alpha^B = \mathbf{T}^0(\lambda_B(\alpha))$*. Alors, lorsque* $B \to +\infty$*,*

$$\mathrm{JER}(\mathbf{t}_\alpha^B) \leq \alpha + O_P(1/\sqrt{B}).$$

Le nombre B d'échantillons Monte-Carlo dans le Théorème 2 peut être choisi arbitrairement grand pour obtenir un contrôle du JER, conduisant à des bornes FDP valides via l'Équation 3. Ce résultat est prouvé dans le papier [3].

Cette approche s'étend naturellement au cas agrégé comme montré dans le papier [3].

3 Quelques résultats expérimentaux

Méthodes considérées. Dans notre mise en œuvre de KOPI, nous nous appuyons sur la moyenne harmonique [19] comme schéma d'agrégation f. De plus, nous fixons $k_{max} = \lfloor p/50 \rfloor$ suivant l'approche de [4]. Nous considérons également les schémas d'agrégation de Knockoffs de l'état de l'art: AKO (Aggregation of Multiple Knockoffs, 13) et l'agrégation basée sur les e-values [14]. En outre, nous considérons les Knockoffs standard ("Vanilla knockoffs)", c'est-à-dire [6] et le contrôle du FDP via le Closed Testing [9]. Dans les expériences sur des données simulées, nous générons des Knockoffs en supposant une distribution gaussienne pour $\mathbf{X}$, avec toutes les variables centrées. Pour les méthodes qui permettent l'agrégation, nous utilisons $D = 50$ tirages de Knockoffs.

3.1 Données simulées

Cadre de simulation. À chaque simulation, nous générons des données gaussiennes $\mathbf{X} \in \mathbb{R}^{n \times p}$ avec une matrice de corrélation de Toeplitz correspondant à un modèle auto-régressif d'ordre 1, de paramètre ρ, c'est-à-dire $\mathbf{\Sigma}_{i,j} = \rho^{|i-j|}$.

Ensuite, nous tirons le vrai support $\boldsymbol{\beta}^* \in \{0,1\}^p$. Le nombre de coefficients non nuls de $\boldsymbol{\beta}^*$ est contrôlé par le paramètre de parcimonie s_p, c'est-à-dire $s_p = \|\boldsymbol{\beta}^*\|_0/p$. La variable cible $\mathbf{y}$ est construite à l'aide d'un modèle linéaire :

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

avec σ contrôlant l'amplitude du bruit : $\sigma = \|\mathbf{X}\boldsymbol{\beta}^*\|_2/(\mathrm{SNR}\|\boldsymbol{\epsilon}\|_2)$, SNR étant le rapport signal à bruit. Nous choisissons le paramètre central $n = 500, p = 500, \rho = 0.5, s_p = 0.1, \mathrm{SNR} = 2$. Pour chaque paramètre, nous explorons une gamme de valeurs possibles pour comparer les méthodes dans divers paramètres.

Pour sélectionner les variables en utilisant les bornes supérieures du FDP, nous retenons l'ensemble le plus grand possible de variables S tel que $V(S) \leq q|S|$. Pour chacune des N simulations et chaque méthode, nous calculons le FDP empirique et la Proportion de Vrais Positifs (TPP) :

$$\widehat{FDP}(S) = \frac{|S \cap \mathcal{H}_0|}{|S|} \quad \text{et} \quad \widehat{TPP}(S) = \frac{|S \cap \mathcal{H}_1|}{|\mathcal{H}_1|}.$$

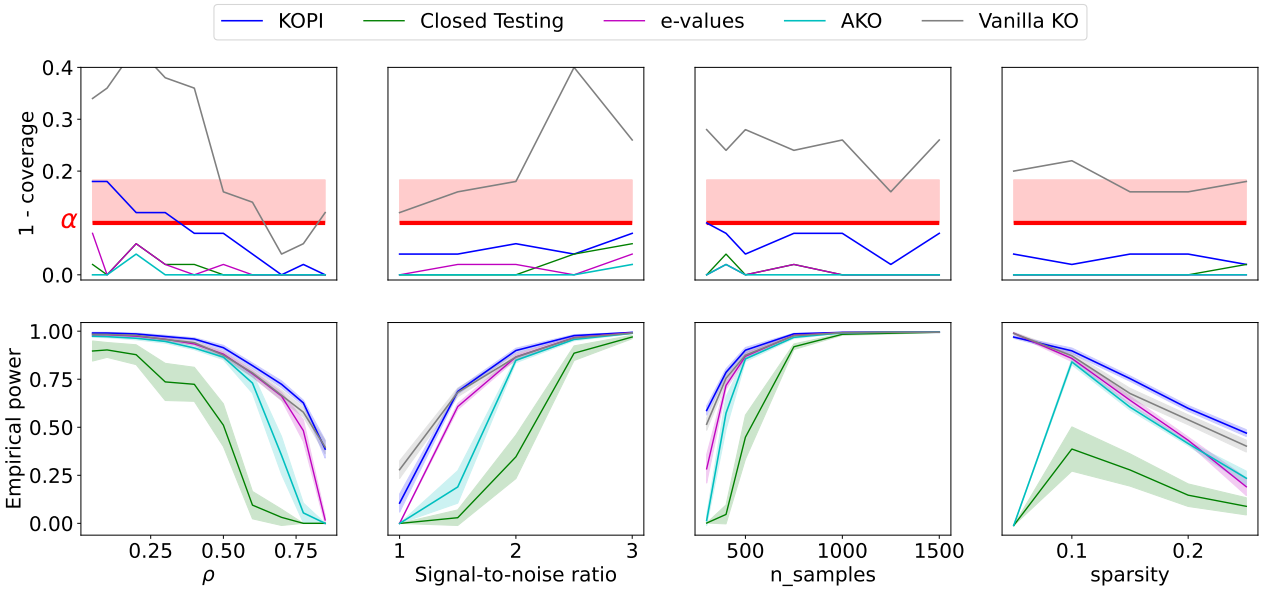

Figure 1: **Couverture de la borne du FDP au niveau α et puissance empirique pour 50 exécutions de simulation et cinq méthodes différentes :** Knockoffs standard, Knockoffs agrégés utilisant les e-values, Knockoffs agrégés utilisant l'agrégation quantile, KOPI et l'inférence Knockoff via le Closed Testing. Nous utilisons $D = 50$ tirages de Knockoffs et les paramètres de simulation suivants : $\alpha = 0.1, q = 0.1, p = 500$. Chaque colonne correspond à une expérience ou l'on fait varier le paramètre indiqué en abscisse, la première ligne affichant la couverture du FDP et la deuxième ligne la puissance. La ligne rouge et les bandes d'erreur associées représentent les limites acceptables (au niveau α) pour la couverture de la borne du FDP. KOPI surpasse constamment toutes les autres méthodes tout en conservant le contrôle du FDP.

Si le FDP est contrôlé au niveau α, $|\{k \in [\![N]\!] : \widehat{FDP}(S_k) > q\}| \sim \mathcal{B}(N, \alpha)$. Ensuite, nous pouvons calculer les bandes d'erreur sur le niveau α en utilisant $\text{std}\left(\mathcal{B}(N, \alpha)/N\right) = \sqrt{\alpha(1-\alpha)/N}$. La deuxième ligne de la Fig. 1 représente la puissance empirique atteinte par chaque méthode, qui correspond à la moyenne des TPPs définis ci-dessus pour N simulations, c'est-à-dire $\text{Puissance} = \sum_{k=1}^{N} \widehat{TPP}(S_k)/N$. La Fig. 1 montre que dans tous les paramètres différents, KOPI conserve le contrôle du FDP. Nous pouvons également voir que le contrôle du FDR n'implique pas le contrôle du FDP, car les Knockoffs standard sont systématiquement en dehors des intervalles de couverture de la borne du FDP. Cependant, les deux schémas d'agrégation existants (AKO et e-valeurs) qui garantissent formellement le contrôle du FDR sont généralement conservateurs et contrôlent empiriquement le FDP. Ceci est cohérent avec les conclusions de [14]. La procédure de Closed Testing de [9] contrôle le FDP comme annoncé mais souffre d'un manque de puissance.

Nous constatons ici que KOPI contrôle le FDP tout en offrant des gains de puissance par rapport aux méthodes d'agrégation de Knockoffs contrôlant le FDR. Pourtant, le contrôle du FDP est une garantie plus stricte que le contrôle du FDR, comme discuté précédemment. Ces gains sont particulièrement notables dans les cadres de simulations les plus difficiles, où la plupart des autres méthodes montrent une diminution claire de la puissance ou même un comportement catastrophique (c'est-à-dire une puissance nulle).

3.2 Application aux données d'IRMf

L'objectif de la cartographie du cerveau humain est d'associer des tâches cognitives à des régions cérébrales pertinentes. Ce problème est abordé à l'aide de l'Imagerie par résonance magnétique fonctionnelle (IRMf), qui consiste à enregistrer le niveau d'oxygénation du sang dans le cerveau à l'aide d'un scanner IRM.

L'importance de l'inférence conditionnelle pour ce problème a été soulignée dans [17]. Nous utilisons l'ensemble de données du Projet Connectome Humain (HCP900) qui contient des images cérébrales de jeunes adultes en bonne santé effectuant différentes tâches tout en étant dans un scanner IRM. Les détails sur cet ensemble de données et les résultats empiriques peuvent être trouvés dans l'Annexe D du papier [3].

Le contrôle du FDP et la puissance ne peuvent pas être évalués dans l'application précédente, puisque la vérité terrain n'est pas connue. Par conséquent, suivant l'approche de [12], nous avons effectué une expérience supplémentaire qui consiste à utiliser des données semi-simulées. Nous considérons un premier ensemble de données fMRI $(\mathbf{X}_1, \mathbf{y}_1)$ sur lequel nous effectuons une inférence en utilisant un estimateur Lasso ; cela produit $\boldsymbol{\beta}_1^* \in \mathbb{R}^p$ que nous utiliserons comme notre vérité terrain. Ensuite, nous considérons un ensemble de données fMRI différent $(\mathbf{X}_2, \mathbf{y}_2)$ pour la génération de données. L'intérêt d'utiliser un ensemble de données différent est d'éviter un biais de circularité entre la définition de la vérité terrain et la procédure d'inférence. Concrètement, nous écartons le vecteur de réponse original $\mathbf{y}_2$ pour cet ensemble de données et construisons une réponse simulée $\mathbf{y}_2^{sim}$ en utilisant un modèle linéaire, avec la même notation que précédemment (nous fixons σ de sorte que SNR $= 4$) : $\mathbf{y}_2^{sim} = \mathbf{X}_2\boldsymbol{\beta}_1^* + \sigma\boldsymbol{\epsilon}$.

Ensuite, l'inférence knockoff est réalisée à partir des données $(\mathbf{X}_2, \mathbf{y}_2^{sim})$. Puisque nous considérons $\boldsymbol{\beta}_1^*$ comme la vérité terrain, le FDP et le TPP peuvent être calculés pour chaque méthode. Comme on peut le voir dans la Fig. 2, KOPI est la méthode la plus puissante parmi celles qui contrôlent le FDP.

4 Discussion

Nous avons proposé une nouvelle méthode qui contrôle du FDP sur les Knockoffs agrégés. Elle combine les avantages de l'agrégation, c'est-à-dire l'amélioration de la stabilité de l'inférence, en plus de fournir un contrôle probabiliste du FDP, plutôt que de contrôler seulement son espérance, le FDR.

Les résultats de simulation confirment que KOPI contrôle effectivement le FDP. De plus, bien que le contrôle du FDP soit une garantie plus stricte que le contrôle du FDR, KOPI offre en réalité des gains de puissance par rapport à l'état de l'art pour l'agrégation de Knockoffs.

Un package Python contenant le code pour KOPI est disponible à l'adresse: `https://github.com/alexblnn/KOPI`.

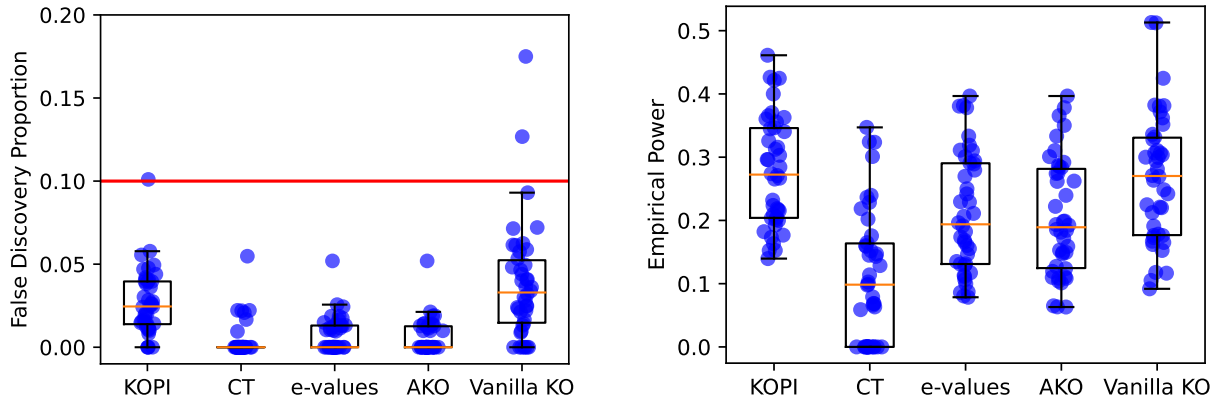

Figure 2: **FDP empirique et puissance sur des données semi-simulées pour 42 paires de contrastes.** Nous utilisons 7 contrastes HCP C0 : "Main motrice", C1 : "Pied motrice", C2 : "Jeu", C3 : "Relationnel", C4 : "Émotion", C5 : "Social", C6 : "Mémoire de travail". Nous considérons toutes les 42 paires d'entraînement/test possibles : le contraste d'entraînement est utilisé pour obtenir une vérité terrain, tandis que le contraste de test est utilisé pour générer la réponse. L'inférence est effectuée en utilisant les 5 méthodes considérées dans l'article et le FDP empirique est rapporté dans le diagramme à boîte de gauche, tandis que la puissance est rapportée dans le diagramme à boîte de droite. Remarquez (figure de droite) que KOPI offre une puissance supérieure par rapport à toutes les autres méthodes basées sur les Knockoffs tout en contrôlant le FDP (fig. de gauche).

References

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

[2] Benjamini, Y. and Hochberg, Y. (1995). 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.

[3] Blain, A., Thirion, B., Grisel, O., and Neuvial, P. (2023). False discovery proportion control for aggregated knockoffs. *NeurIPS 2023*.

[4] Blain, A., Thirion, B., and Neuvial, P. (2022). Notip: Non-parametric true discovery proportion control for brain imaging. *NeuroImage*, 260:119492.

[5] Blanchard, G., Neuvial, P., Roquain, E., et al. (2020). Post hoc confidence bounds on false positives using reference families. *Annals of Statistics*, 48(3):1281–1303.

[6] Candès, E., Fan, Y., Janson, L., and Lv, J. (2018). 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.

[7] Goeman, J. J. and Solari, A. (2011). Multiple testing for exploratory research. *Statistical Science*, 26(4):584–597.

[8] Janson, L. and Su, W. (2016). Familywise error rate control via knockoffs. *Electronic Journal of Statistics*, 10(1):960–975.

[9] Li, J., Maathuis, M. H., and Goeman, J. J. (2022). Simultaneous false discovery proportion bounds via knockoffs and closed testing. *arXiv preprint arXiv:2212.12822*.

[10] Meinshausen, N., Meier, L., and Bühlmann, P. (2009). P-values for high-dimensional regression. *Journal of the American Statistical Association*, 104(488):1671–1681.

[11] Neuvial, P. (2020). *Contributions to statistical inference from genomic data*. Habilitation thesis, Université Toulouse III Paul Sabatier.

[12] Nguyen, B. T., Thirion, B., and Arlot, S. (2022). A Conditional Randomization Test for Sparse Logistic Regression in High-Dimension. In *NeurIPS 2022*, volume 35 of *Advances in Neural Information Processing Systems*, New Orleans, United States.

[13] Nguyen, T.-B., Chevalier, J.-A., Thirion, B., and Arlot, S. (2020). Aggregation of multiple knockoffs. In *International Conference on Machine Learning*, pages 7283–7293. PMLR.

[14] Ren, Z. and Barber, R. F. (2022). Derandomized knockoffs: leveraging e-values for false discovery rate control. *arXiv preprint arXiv:2205.15461*.

[15] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288.

[16] Wang, R. and Ramdas, A. (2022). False discovery rate control with e-values. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(3):822–852.

[17] Weichwald, S., Meyer, T., Özdenizci, O., Schölkopf, B., Ball, T., and Grosse-Wentrup, M. (2015). Causal interpretation rules for encoding and decoding models in neuroimaging. *NeuroImage*, 110:48–59.

[18] Weinstein, A., Su, W. J., Bogdan, M., Barber, R. F., and Candès, E. J. (2020). A power analysis for knockoffs with the lasso coefficient-difference statistic. *arXiv preprint arXiv:2007.15346*.

[19] Wilson, D. J. (2019). The harmonic mean p-value for combining dependent tests. *Proceedings of the National Academy of Sciences*, 116(4):1195–1200.