Aggregation of Multiple Knockoffs

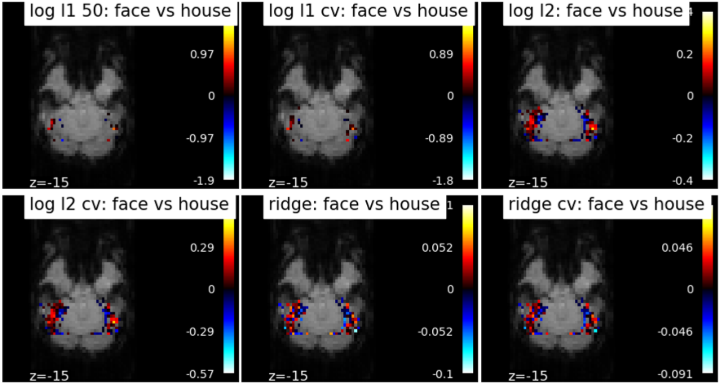
B. Nguyen [1,2,4] J.-A. Chevalier [1,2,3] S. Arlot [1,4] B. Thirion [1,2]

[1]INRIA [2]CEA/Neurospin [3]Telecom ParisTech

[4]Laboratoire de Mathématiques d'Orsay, CNRS, Université Paris-Saclay

June 03, 2020

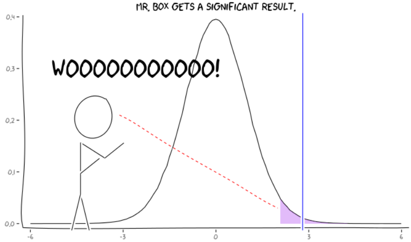
Motivation

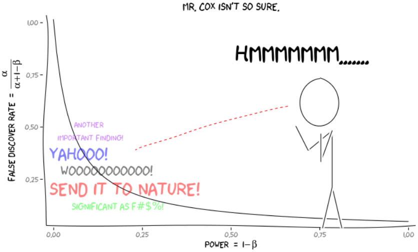


Source: nilearn.github.io

Introduction to False Discovery Rate

- Introduced by Benjamini and Hochberg (1995).
- False Discovery Proportion (FDP): How many False Discoveries made among all discoveries?
- False Discovery Rate (FDR): the expected value of FDP.





Source: stats.stackexchange.com

Problem settings

- $\mathbf{X} \in \mathbb{R}^{n \times p}$, $\mathbf{y} \in \mathbb{R}^n$. Example: $\mathbf{X}$ is MRI data, $\mathbf{y}$ outcome.
- Linear model assumption $\mathbf{y} = \mathbf{X}\boldsymbol{\beta}^* + \sigma\boldsymbol{\epsilon}$ with $\epsilon_i \sim \mathcal{N}(0, 1)$.
- Support set $\mathcal{S} := \{i : \beta_i^* \neq 0\}$ and its estimate $\hat{\mathcal{S}}$.

False Discovery Proportion – FDP

$$\text{FDP} = \frac{\mathbf{card}(\hat{\mathcal{S}} \cap \mathcal{S}^c)}{\mathbf{card}(\hat{\mathcal{S}}) \vee 1}$$

False Discovery Rate – FDR

$$\text{FDR} = \mathbb{E}[\text{FDP}]$$

Multivariate Statistical Inference

$n > p$: Ordinary Least Square (OLS)

$$\hat{\boldsymbol{\beta}}^{OLS} = \underset{\boldsymbol{\beta} \in \mathbb{R}^p}{\operatorname{argmin}} \frac{1}{2} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2$$

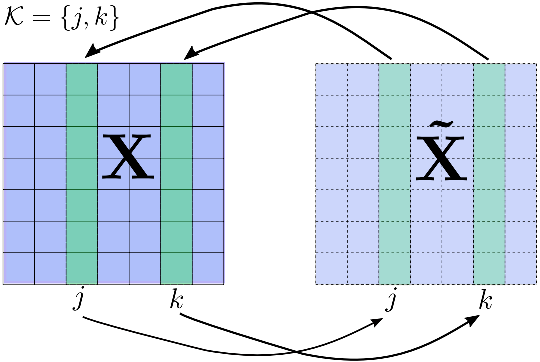
$\longrightarrow \hat{\boldsymbol{\beta}}^{OLS} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \rightarrow$ test score, p-value, confidence interval.

$n < p$: No closed form solution – $\mathbf{X}^T\mathbf{X}$ not invertible

Solution: : Lasso, Ridge, Elastic Net $\rightarrow$ p-value, confidence interval for variable selection?

$\longrightarrow$ Knockoff Inference: Multivariate Variable Selection for High-dimensional settings.

Knockoff Variables: Definition



Definition (?)

$\tilde{\mathbf{X}} = (\mathbf{x}_1, \ldots, \mathbf{x}_p)$ is model-X knockoffs of $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_p)$ if and only if:

1. $\forall$ subset $\mathcal{K} \subset \{1, \ldots, p\}$: $(\mathbf{X}, \tilde{\mathbf{X}})_{\text{swap}(\mathcal{K})} \overset{d}{=} (\mathbf{X}, \tilde{\mathbf{X}})$
2. $\tilde{\mathbf{X}} \perp \mathbf{y} \mid \mathbf{X}$

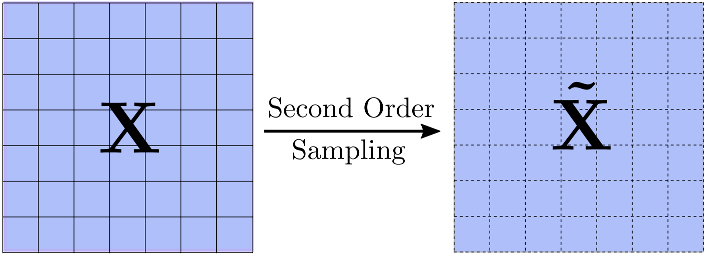
Knockoff Variables: Intuition

$\tilde{X}$ is **noisy copies** of original variables X:

- Share same ‘structure‘ with original design matrix, but
- Has to be null variables.

Knockoff Sampling: Second-order Knockoffs

$$\mathrm{cov}(\mathbf{X}, \tilde{\mathbf{X}}) = \begin{bmatrix} \mathbf{\Sigma} & \mathbf{\Sigma} - \mathrm{diags}\{s\} \\ \mathbf{\Sigma} - \mathrm{diags}\{s\} & \mathbf{\Sigma} \end{bmatrix}$$



Shares the same first 2 moments - mean and covariance:

$$\mathbb{E}[\tilde{\mathbf{X}}] = \mathbb{E}[\mathbf{X}] = \boldsymbol{\mu}, \quad \mathbb{E}[\tilde{\mathbf{X}}^T \tilde{\mathbf{X}}] = \mathbf{\Sigma} \quad \text{and} \quad \mathbb{E}[\tilde{\mathbf{X}}^T \mathbf{X}] = \mathbf{\Sigma} - \mathrm{diag}\{\mathbf{s}\}$$

Knockoff Sampling: Second-order Knockoffs

- $\mathrm{diag}\{\mathbf{s}\}$ is perturbation matrix that makes the joint covariance cov$(\mathbf{X}, \tilde{\mathbf{X}})$ matrix positive definite:
  $\longrightarrow$ **Equi-correlated formula**: $s_j = 2\lambda_{\mathsf{min}}(\mathbf{\Sigma}) \wedge 1$ (Candès et al., 2018)
- **In practice:** Estimation of $\mathbf{\Sigma}$ is required.

Assumption

$$(X_{i1}, \dots X_{ip}, y_i) \overset{i.i.d}{\sim} F_{XY}, \forall i = 1, \dots, n, \text{ and } F_X \text{ is known.}$$

Assumption: $\mathbf{X}$ has Gaussian design

$$\tilde{\mathbf{x}_j} \mid \mathbf{x}_j \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{V})$$

$$\mathbf{V} = 2\mathrm{diag}\{\mathbf{s}\} - \mathrm{diag}\{\mathbf{s}\}\mathbf{\Sigma}^{-1}\mathrm{diag}\{\mathbf{s}\}$$

Knockoff Statistic

Definition (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. Depends only on $\mathbf{X}, \tilde{\mathbf{X}}$ and $\mathbf{y}$

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

2. Swapping the original variable column $\mathbf{x}_j$ and its knockoff column $\tilde{\mathbf{x}}_j$ will switch the sign of W_j iff j is in the support set $\mathcal{S}$:

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

Knockoff Statistic

Assumption (Null Distribution of Knockoff Statistic)

Under the Null hypothesis, the Knockoff Statistics defined above, i.e. $\{W_j\}_{j \in \mathcal{S}^c}$, *follow the same distribution.*

Remark

From Barber and Candès (2015): this Null distribution is symmetric around 0.

Knockoff Inference (Barber and Candès, 2015)

Step 1

Construct knockoff variables, concatenate $[\mathbf{X}, \tilde{\mathbf{X}}] \in \mathbb{R}^{n \times 2p}$

Step 2

Calculate knockoff test-statistics: *Lasso coefficient-difference*, obtain

$$\hat{\boldsymbol{\beta}} = \min_{\mathbf{w} \in \mathbb{R}^{2p}} \frac{1}{2} \|\mathbf{y} - [\mathbf{X}, \tilde{\mathbf{X}}]\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_1$$

then take the difference: $W_j = \left|\hat{\beta}_j(\lambda)\right| - \left|\hat{\beta}_{j+p}(\lambda)\right|$ for each j

Knockoff Inference (Barber and Candès, 2015)

Step 3 – FDR controlling threshold

For given $t > 0$, False Discoveries Proportion can be estimated as:

$$\widehat{\mathsf{FDP}}(t) = \frac{1 + \#\{j : W_j \leq -t\}}{\#\{j : W_j \geq t\}}$$

then, for FDR level $\alpha \in (0, 1)$, calculate the threshold $\tau > 0$

$$\tau = \min \left\{ t > 0 : \widehat{\mathsf{FDP}}(t) \leq \alpha \right\}$$

Step 4

Select the variables: $\hat{S}(\tau) = \{j : W_j \geq \tau \mid j = 1, \dots, p\}$

FDP estimation with Knockoff Statistic

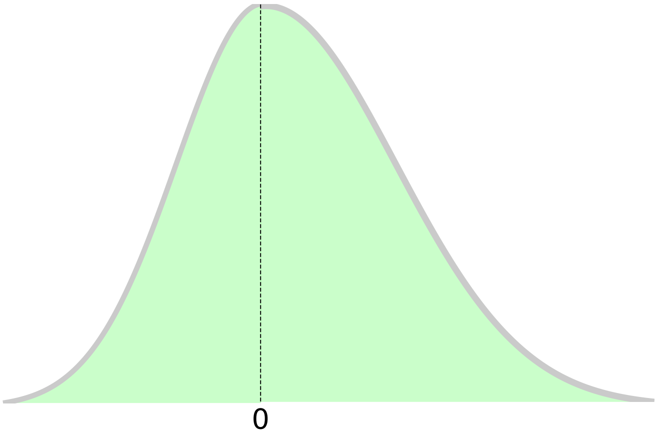


Figure: Distribution of Knockoff Statistic $\{W_j\}_{j=1}^{p}$

FDP estimation with Knockoff Statistic

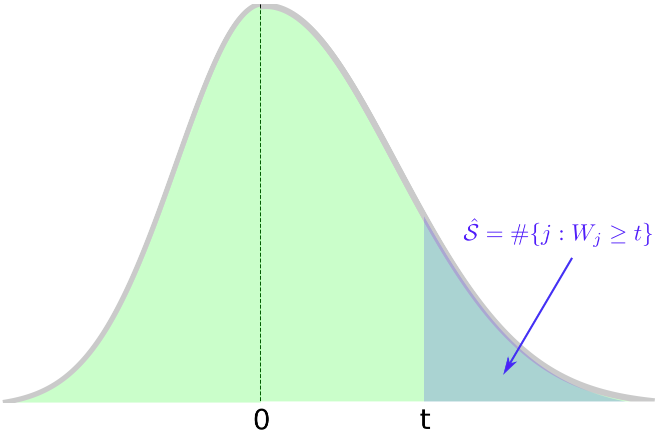


Figure: Distribution of Knockoff Statistic $\{W_j\}_{j=1}^{p}$

FDP estimation with Knockoff Statistic

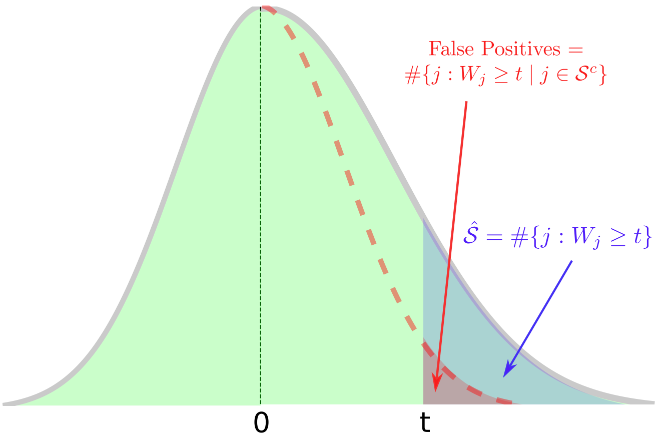


Figure: Distribution of Knockoff Statistic $\{W_j\}_{j=1}^p$

FDP estimation with Knockoff Statistic

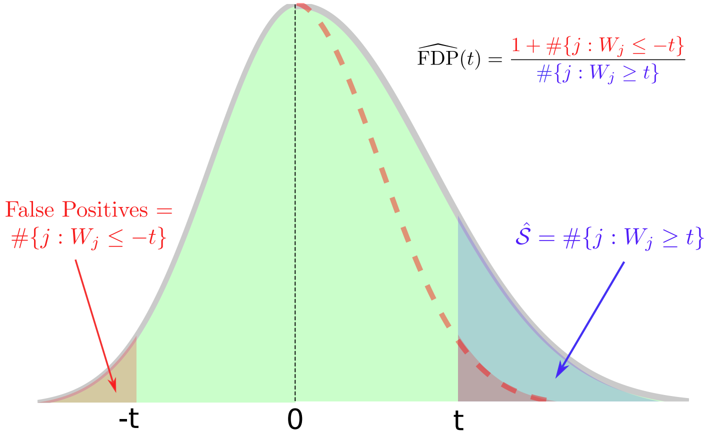


Figure: Distribution of Knockoff Statistic $\{W_j\}_{j=1}^p$

Knockoff Inference: Theoretical Guarantee FDR control

Theorem (Barber and Candès, 2015; Candès et al., 2018)

$$\mathsf{FDR}(\tau) = \mathbb{E}\left[\frac{\mathbf{card}(\hat{S}(\tau) \cap \mathcal{S}^c)}{\mathbf{card}(\hat{S}(\tau)) \vee 1}\right] \leq \alpha$$

Proof: Using martingale theory (optional stopping time theorem).

Instability of knockoff procedure

Settings for Simple Scenario Simulation: 3 simulation parameters: ρ, snr and sparsity.

- $n = 500$, $p = 1000$.
- $\mathbf{X} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$ symmetric Toeplitz matrix.
- $\rho \in [0, 1) : \mathbf{\Sigma} = \begin{bmatrix} \rho^0 & \rho^1 & \rho^2 & \dots & \rho^{p-1} \\ \rho^1 & \rho^0 & \rho^1 & \dots & \rho^{p-2} \\ \vdots & \dots & \ddots & \dots & \vdots \\ \rho^{p-2} & \rho^{p-3} & \dots & \rho^0 & \rho^1 \\ \rho^{p-1} & \rho^{p-2} & \rho^{p-3} & \dots & \rho^0 \end{bmatrix}$
- $\mathbf{y} = \mathbf{X}\boldsymbol{\beta}^* + \sigma\boldsymbol{\epsilon}$, $\epsilon_i \sim \mathcal{N}(0, 1)$.
- $\sigma = \dfrac{\|\mathbf{X}\boldsymbol{\beta}^*\|_2}{\text{snr} \times \|\boldsymbol{\epsilon}\|_2}$
- sparsity $= \dfrac{\textbf{card}(\mathcal{S})}{p} \in [0, 1]$

Instability of knockoff procedure

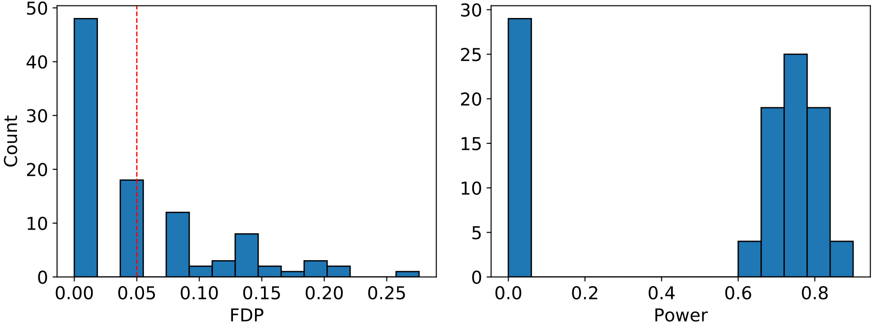
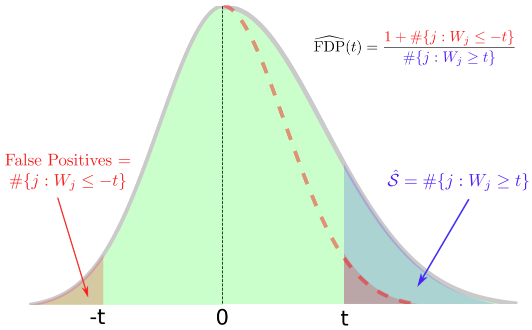


Figure: 100 runs of knockoff inference on the **same simulation**
n=500, p=1000, snr=3.0, $\rho = 0.7$, sparsity = 0.06

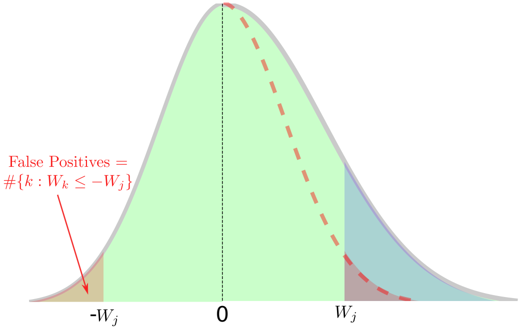
Solution: Knockoffs Statistic conversion



Introduce the intermediate p-values: convert Knockoff statistic W_j to π_j:

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

Solution: Knockoffs Statistic conversion



Introduce the intermediate p-values: convert Knockoff statistic W_j to π_j:

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

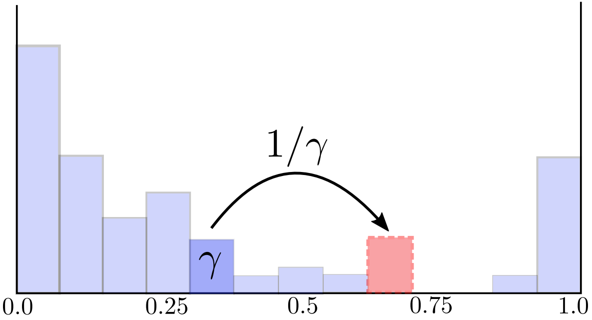
Aggregation of Multiple Knockoffs

Step 1: For $b = 1, 2, \ldots, B$ the number of bootstraps:

- Run knockoff sampling, calculate test statistic $\left\{W_j^{(b)}\right\}_{j \in [p]}$
- Convert the test statistic $W_j^{(b)}$ to $\pi_j^{(b)}$:

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

Aggregation of Multiple Knockoffs



Step 2 – Quantile Aggregation of p-values (Meinshausen et al., 2009)

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

For $\gamma \in (0,1)$ with $q_\gamma(\cdot)$ the empirical γ-quantile function.

Aggregation of Multiple Knockoffs

Step 3 – FDR control with $\bar{\pi}$

- Order $\bar{\pi}_j$ ascendingly: $\bar{\pi}_{(1)} < \bar{\pi}_{(2)} \cdots < \bar{\pi}_{(p)}$
- Given FDR control level $\alpha \in (0, 1)$, find largest k such that:
  - $\bar{\pi}_{(k)} \leq k\alpha/p$ (Benjamini and Hochberg, 1995), or
  - $\bar{\pi}_{(k)} \leq \dfrac{k\alpha}{p \sum_{i=1}^{p} 1/i}$ (Benjamini and Yekutieli, 2001)

  $\longrightarrow$ FDR threshold: $\tau = \bar{\pi}_{(k)}$
- $\hat{\mathcal{S}}_{AKO} = \{j : \bar{\pi}_j \leq \tau \mid j \in [p]\}$

Theoretical Results

Assumption (Null Distribution of Knockoff Statistic)

Under the Null hypothesis, the Knockoff Statistics defined above, i.e. $\{W_j\}_{j \in \mathcal{S}^c}$, *follow the same distribution.*

Lemma (Non-asymtotical validity of Intermediate p-Values)

Under the above assumption , and furthermore assume $|\mathcal{S}^c| \geq 2$, *the empirical p-value* π_j *satisfies*

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

for all $j \in \mathcal{S}^c = \{j = 1, \ldots, p : \beta^*_j = 0\}$ *and where* $\kappa = \dfrac{\sqrt{22} - 2}{7\sqrt{22} - 32} \leq 3.24$

Theoretical Results - Main theorem

Theorem (Non-asymtotic guarantee for FDR control with AKO)

If the above assumption holds, and if $|\mathcal{S}^c| \geq 2$, *then for an arbitrary number of bootstraps* B, *the output* $\hat{\mathcal{S}}_{AKO}$ *of Aggregation of Multiple Knockoff (AKO) controls FDR under predefined level* $\alpha \in (0, 1)$ *in asymptotic regime:*

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

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

Experimental Results - Synthetic Data

Same settings: Simple Scenario with

- $n = 500$, $p = 1000$.
- $\mathbf{X} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$
- $\mathbf{y} = \mathbf{X}\boldsymbol{\beta}^* + \sigma\boldsymbol{\epsilon}$, $\epsilon_i \sim \mathcal{N}(0, 1)$. with $\mathbf{\Sigma}$ symmetric Toeplitz matrix.
- 3 simulation parameters: ρ, snr and sparsity.

Experimental Results - Synthetic Data

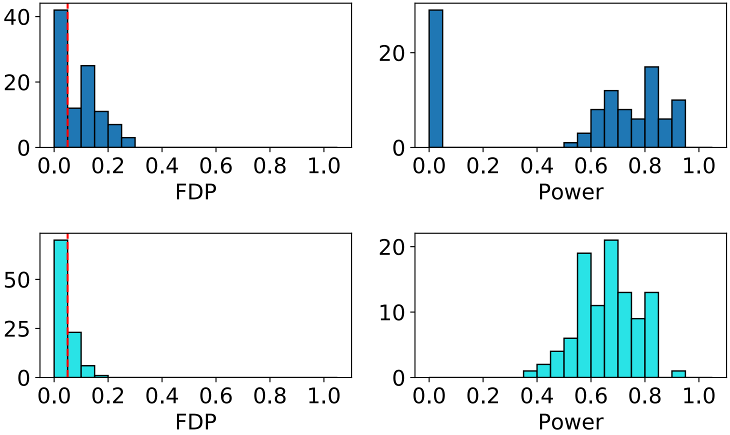


Figure: **Histogram of FDP & Power for 100 runs of Original Knockoff (top) vs. Aggregated Knockoff (bottom) under <u>the same simulation</u>.**
SNR $= 3.0, \rho = 0.5,$ sparsity $= 0.06.$

Empirical Analysis for the choice of B and γ

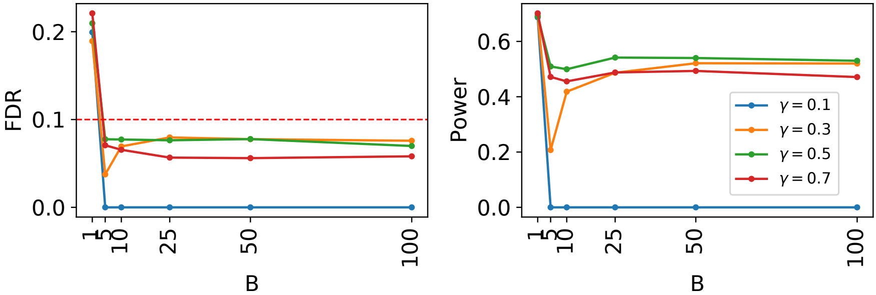
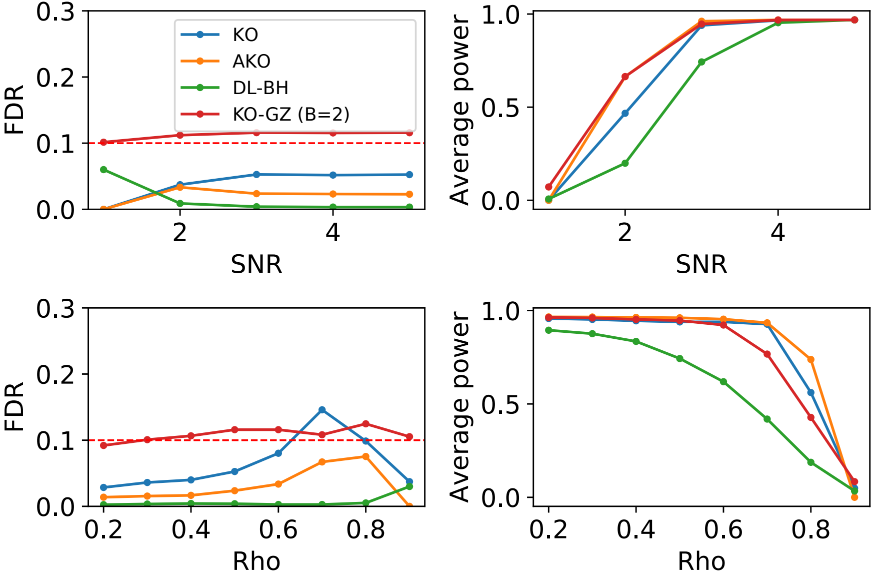


Figure: FDR and Averaging Power for 30 simulations with fixed SNR $= 3.0, \rho = 0.7,$ sparsity $= 0.06$ while varying B and γ

Experimental Results - Synthetic Data

- Vary each of the three simulation parameters while keeping the others unchanged at default value: SNR $= 3.0, \rho = 0.5,$ sparsity $= 0.06$
- Benchmarking methods:
  - **Aggregation of Multiple Knockoffs (AKO)**
  - Vanilla Knockoff (KO) (Candès et al., 2018)
  - Debiased Lasso (DL-BH) (Javanmard and Javadi, 2019)
  - Simultaneous Knockoff (KO-GZ) (Gimenez and Zou, 2019)

Experimental Results - Synthetic Data



Experimental Results - Genome Wide Association Study

- Data: Flowering Phenotype of Arabidopsis Thaliana – $n = 166, p = 9938$
- Objective: detect association of 174 candidate genes with phenotype FT_GH that dictates flowering time (Atwell et al., 2010).
- Preprocessing: dimension reduction following Slim et al. (2019)

$$p = 9938 \longrightarrow p = 1500.$$

Experimental Results - Genome Wide Association Study

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

Figure: **List of detected genes associated with phenotype FT_GH**. Empty line (—) signifies no detection.

Confirmation from previous studies:

- AT2G21070 (Kim et al., 2008)
- AT4G02780 (Silverstone et al., 1998)
- AT5G47640 (Cai et al., 2007)

Experimental Results - Brain Imaging

- Data: Human Connectome Project
- Objective: predict the experimental condition per task given brain activity
- $n = 900$ subjects, $p \approx 212000$
- Preprocessing: dimension reduction by clustering
$$p = 212000 \longrightarrow p = 1000$$

Experimental Results - Brain Imaging

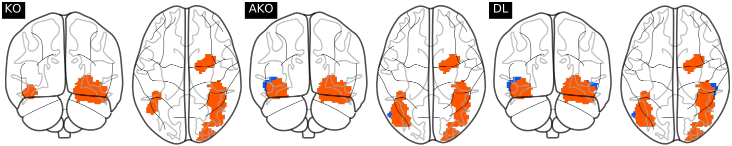


Figure: Detection of significant brain regions for HCP data (900 subjects). Selected regions in a reaction with Emotion images task.

Orange: brain areas with positive sign activation.

Blue: brain areas with negative sign activation

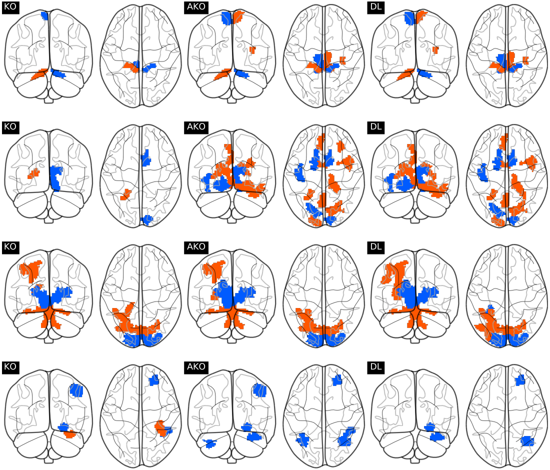
Experimental Results - Brain Imaging

Motor Foot

Gambling

Relational

Working memory



Experimental Results - Brain Imaging

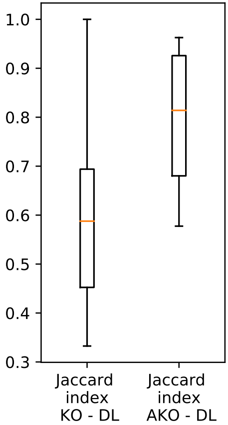


Figure: Jaccard index measuring the Jaccard similarity between the KO/AKO solutions and the DL solution over 7 tasks of HCP900

Conclusion

Conclusion:

- Knockoff:
  - Versatile (different loss functions, different test statistics)
  - But unstable, depends on quality of knockoff variables.
- Aggregation of Multiple Knockoffs
  $\longrightarrow$ increases stability
  $\longrightarrow$ theoretically control FDR
  $\longrightarrow$ higher power

Perspective

Future work:

- Knockoff Sampling Scheme for scaling to very large dimension: promising methods include Deep Knockoffs Machine (Romano et al., 2018) – Generative adversarial knockoff networks (Jordon and Yoon, 2019)
- Further analysis on Theoretical Properties of AKO: relax assumptions about knockoff statistics

Acknowledgement

Sylvain Arlot

Bertrand Thirion

- Coauthors: Jerome-Alexis Chevalier, Sylvain Arlot, Bertrand Thirion
- Lotfi Slim & Chloe-Agathe Azencott for helping with preprocessing genomic dataset.
- ANR project FAST-BIG.

Questions?

Main Reference: Nguyen, T.B., Chevalier, J-A, Arlot, S., and Thirion, B. (2020) *Aggregation of Multiple Knockoffs*. To appear at the 37th International Conference on Machine Learning (ICML 2020).

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

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

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.

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

Cai, X., Ballif, J., Endo, S., Davis, E., Liang, M., Chen, D., DeWald, D., Kreps, J., Zhu, T., and Wu, Y. (2007). A putative ccaat-binding transcription factor is a regulator of flowering timing in arabidopsis. *Plant Physiology*, 145(1):98–105.

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.

Gimenez, J. R. and Zou, J. (2019). Improving the stability of the knockoff procedure: Multiple simultaneous knockoffs and entropy maximization. In Chaudhuri, K. and Sugiyama, M., editors, *Proceedings of Machine Learning Research*, volume 89 of *Proceedings of Machine Learning Research*, pages 2184–2192. PMLR.

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

Jordon, J. and Yoon, J. (2019). KnockoffGAN: Generating Knockoffs for Feature Selection using Generative Adversarial Networks. *International Conference on Learning Representations*, page 25.

Kim, J., Kim, Y., Yeom, M., Kim, J.-H., and Nam, H. G. (2008). Fiona1 is essential for regulating period length in the arabidopsis circadian clock. *The Plant Cell*, 20(2):307–319.

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

Romano, Y., Sesia, M., and Candès, E. J. (2018). Deep Knockoffs. *arXiv:1811.06687 [math, stat]*. arXiv: 1811.06687.

Silverstone, A. L., Ciampaglio, C. N., and Sun, T.-p. (1998). The arabidopsis rga gene encodes a transcriptional regulator repressing the gibberellin signal transduction pathway. *The Plant Cell*, 10(2):155–169.

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