Aggregation of Multiple Knockoffs

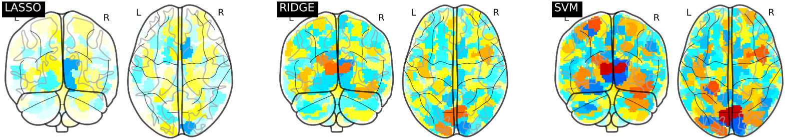
T.-B. Nguyen [1,2,4] J.-A. Chevalier [1,2,3] S. Arlot [1,4] B. Thirion [1,2]

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

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

Motivation

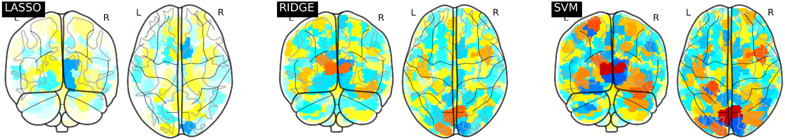
Data: Human Connectome Project (`humanconnectomeproject.org`)



- Brain decoding: different inference results depending on the estimators used.

Motivation

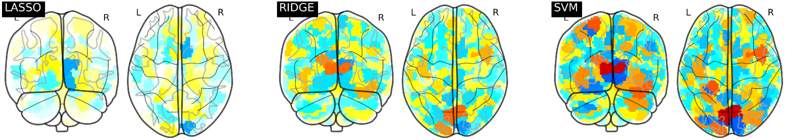
Data: Human Connectome Project (`humanconnectomeproject.org`)



- Brain decoding: different inference results depending on the estimators used.
- Limit the number of falsely detected variables is crucial: introduced the False Discovery Rate – FDR (Benjamini and Hochberg, 1995)

Motivation

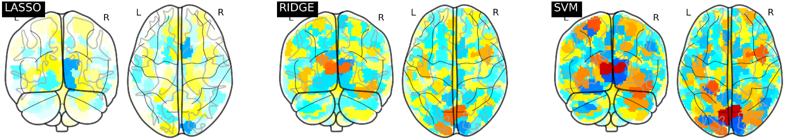
Data: Human Connectome Project (`humanconnectomeproject.org`)



- Brain decoding: different inference results depending on the estimators used.
- Limit the number of falsely detected variables is crucial: introduced the False Discovery Rate – FDR (Benjamini and Hochberg, 1995)
- **However: controlling FDR is still a challenging problem in high-dimensional settings**

Motivation

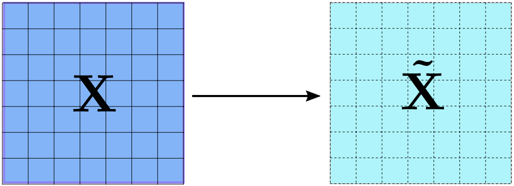
Data: Human Connectome Project (`humanconnectomeproject.org`)



- Brain decoding: different inference results depending on the estimators used.
- Limit the number of falsely detected variables is crucial: introduced the False Discovery Rate – FDR (Benjamini and Hochberg, 1995)
- **However: controlling FDR is still a challenging problem in high-dimensional settings**
  $\longrightarrow$ Knockoff Inference (Barber and Candès, 2015; Candès et al., 2018): recent advance in Multivariate Inference with guaranteed FDR controlling.

Summary

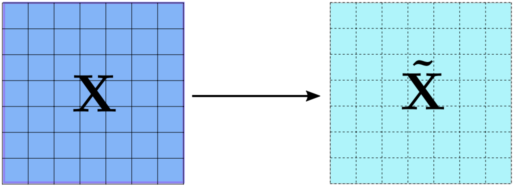
Knockoff Inference (Barber and Candès, 2015; Candès et al., 2018)



- Create $\tilde{\mathbf{X}}$ – the *random noisy copies* of original variables $\mathbf{X}$

Summary

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



- Create $\tilde{\mathbf{X}}$ – the *random noisy copies* of original variables $\mathbf{X}$
- Calculate Knockoff Statistics $\mathbf{W}$: a measure of feature importance $\rightarrow$ FDR threshold calculation $\rightarrow$ Feature Selection

Summary

Knockoff Inference (Barber and Candès, 2015; Candès et al., 2018)
$\longrightarrow$ **Major issue: unstable**

Summary

Knockoff Inference (Barber and Candès, 2015; Candès et al., 2018)
$\longrightarrow$ **Major issue: unstable**

Our contribution: Aggregation of Multiple Knockoffs

- Fix the instability problem
- Theoretically control FDR
- Increasing statistical power empirically

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 $\mathcal{S} := \{i : \beta_i^* \neq 0\}$; null index set $\mathcal{S}^c := \{i : \beta_i^* = 0\}$
- **Objective:** find $\hat{\mathcal{S}}$ – estimation of $\mathcal{S}$

False Discovery Rate – FDR

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

$\longrightarrow$ FDR: the average proportion of false discoveries made amongst all discoveries

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 $\mathbf{W}$: *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: $\hat{S}(\tau) = \{j : W_j \geq \tau \mid j = 1, \ldots, p\}$

FDP estimation with Knockoff Statistic

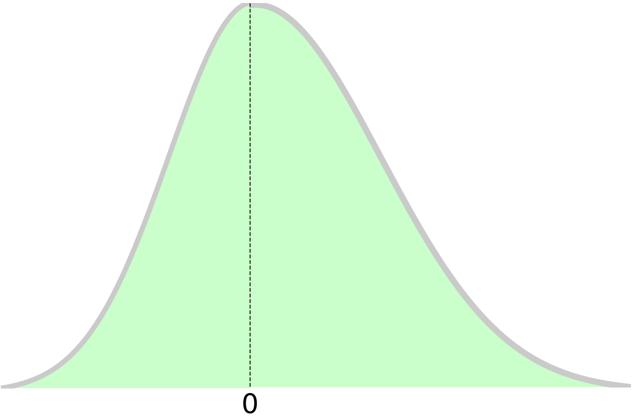


Figure: Distribution of Knockoff Statistic $\{W_j\}_{j=1}^{p}$

FDP estimation with Knockoff Statistic

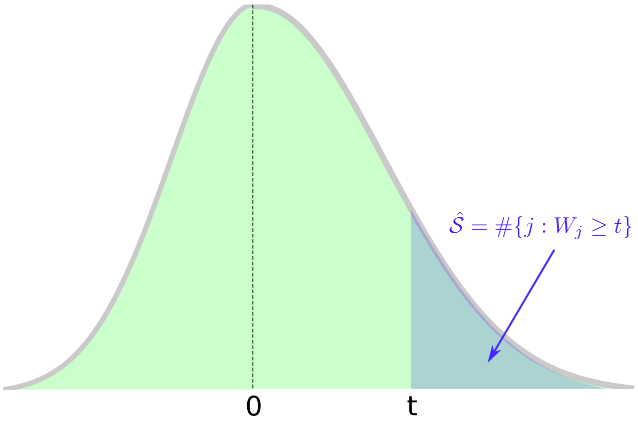


Figure: Distribution of Knockoff Statistic $\{W_j\}_{j=1}^{p}$

FDP estimation with Knockoff Statistic

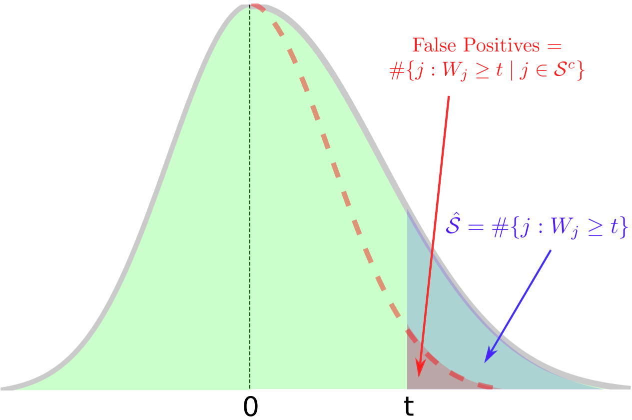


Figure: Distribution of Knockoff Statistic $\{W_j\}_{j=1}^{p}$

FDP estimation with Knockoff Statistic

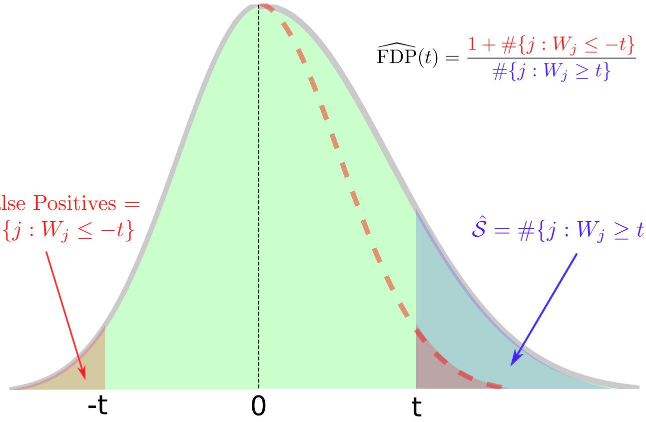


Figure: Distribution of Knockoff Statistic $\{W_j\}_{j=1}^{p}$

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).

Demonstration: Instability of Knockoff Procedure

Settings: Gaussian design with 3 simulation parameters

- ρ: correlation between variables,
- snr: signal to noise ratio
- sparsity: how sparse the signals are.
- $\mathbf{X} \in \mathbb{R}^{n \times p}$, $\mathbf{y} \in \mathbb{R}^{n}$: $n = 500$, $p = 1000$.

Demonstration: Instability of Knockoff Procedure

Settings: Gaussian design with 3 simulation parameters

- ρ: correlation between variables,
- snr: signal to noise ratio
- sparsity: how sparse the signals are.
- $\mathbf{X} \in \mathbb{R}^{n \times p}$, $\mathbf{y} \in \mathbb{R}^{n}$: $n = 500$, $p = 1000$.

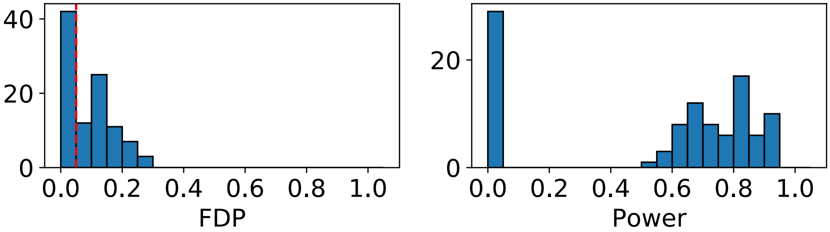
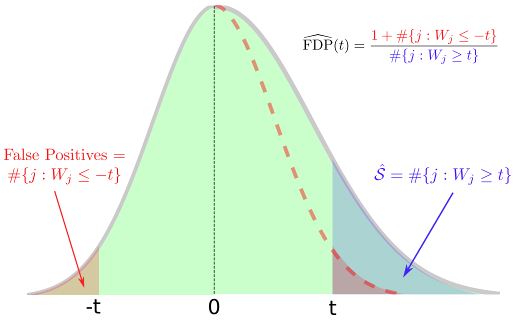
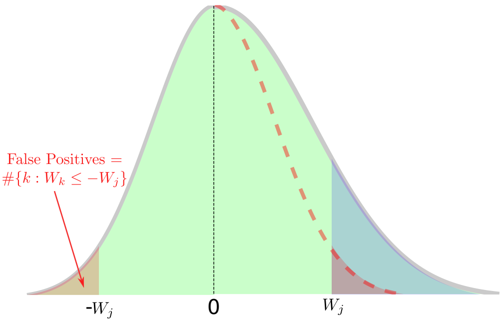


Figure: 100 runs of knockoff inference on the **same simulation**
n=500, p=1000, snr=3.0, $\rho = 0.7$, sparsity = 0.06

Solution: Knockoff Statistics conversion



Solution: Knockoff Statistics conversion



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

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

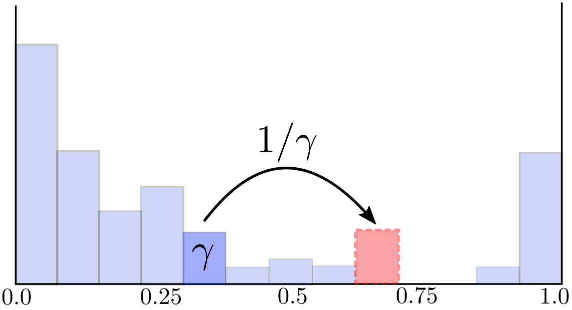
Aggregation of Multiple Knockoffs

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)

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

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

Aggregation of Multiple Knockoffs

Step 3 – FDR control with $\bar{\boldsymbol{\pi}}$

- 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: $\tau = \bar{\pi}_{(k)}$

Step 4 – Estimate $\hat{\mathcal{S}}$

- $\hat{\mathcal{S}}_{AKO} = \{j : \bar{\pi}_j \leq \tau \mid j \in [p]\}$

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-asymptotic 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-asymptotic 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: Gaussian design with $n = 500$, $p = 1000$, 3 simulation parameters: ρ (correlation), snr (signal-to-noise ratio), and sparsity.

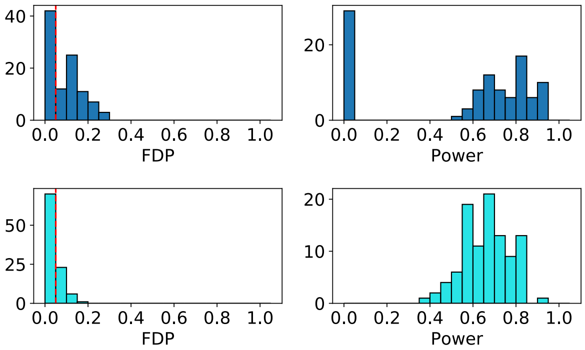
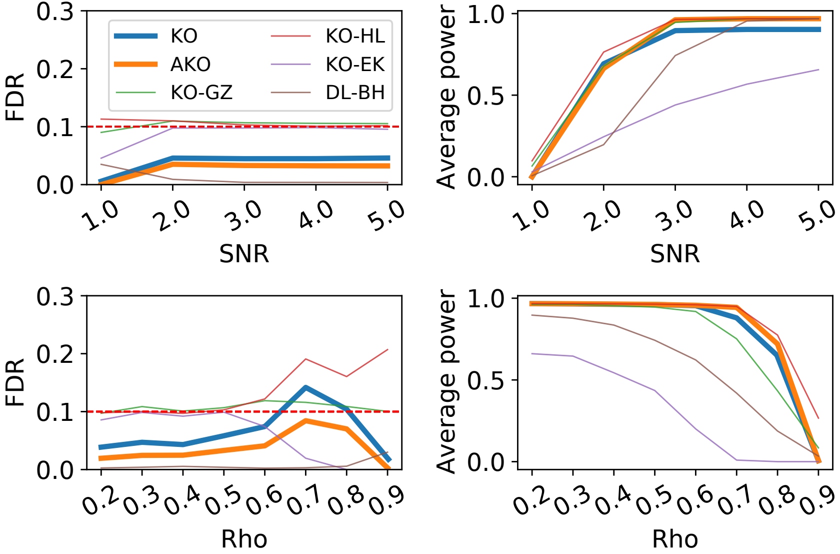


Figure: **Histogram of FDP & Power for 100 runs of Original Knockoff (KO – top) vs. Aggregated Knockoff (AKO – bottom) under <u>the same simulation</u>**. SNR $= 3.0, \rho = 0.5,$ sparsity $= 0.06$.

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:
  - Ours: Aggregation of Multiple Knockoffs **(AKO)**
  - Vanilla Knockoff **(KO)** (Candès et al., 2018)
  - Related knockoff aggregation methods: Holden and Helton (2018) **(KO-HL)**, Emery and Keich (2019) **(KO-EK)**, Gimenez and Zou (2019) **(KO-GZ)**
  - Debiased Lasso **(DL-BH)** (Javanmard and Javadi, 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

- 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)

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

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

Figure: **List of detected genes associated with phenotype FT_GH. Genes detected are confirmed from previous studies**. Empty line (—) signifies no detection.

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

- 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

$$p = 212000 \longrightarrow p = 1000$$

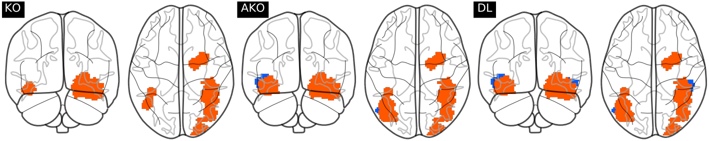


Figure: Detection of significant brain regions for HCP data (900 subjects). Selected regions in a reaction with Emotion images task.
Orange: brain areas with positive sign activation.
Blue: brain areas with negative sign activation

Experimental Results - Brain Imaging

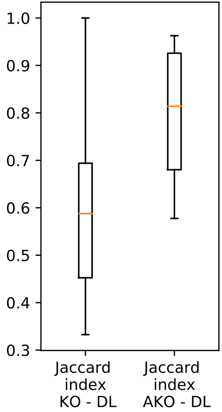


Figure: Jaccard index measuring the Jaccard similarity between the KO/AKO solutions and the DL solution over 7 tasks of HCP900

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 demonstrated empirically

Thank you for listening!

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.

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.

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

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.

Holden, L. and Helton, K. H. (2018). Multiple model-free knockoffs. *arXiv preprint arXiv:1812.04928*.

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

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

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.