Some Contributions to Modern Multiple Hypothesis Testing in High-dimension

Binh Nguyen
Inria Parietal, CEA, Laboratoire de Mathématiques d'Orsay,
EDMH, Université Paris-Saclay
December 10, 2021

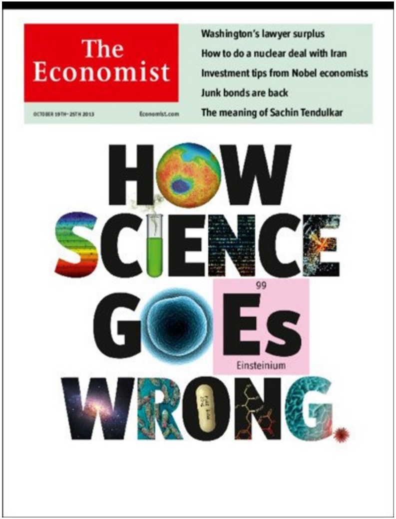
Ph.D. Advisors: Sylvain Arlot *(Université Paris-Saclay)*
Bertrand Thirion *(INRIA Parietal, CEA)*

Ph.D. Committee: Jelle Goeman *(Leiden University – Reviewer)*
Etienne Roquain *(LPSM, Sorbonne Université – Reviewer)*
Jeanette Mumford *(Stanford University)*
Claire Boyer *(LPSM, Sorbonne Université)*
Christophe Giraud *(Université Paris-Saclay)*

Outline

Outline

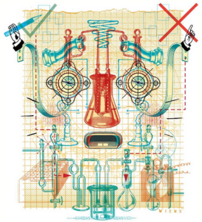
Reproducibility Crisis: on Popular Media...



The New York Times

RAW DATA

New Truths That Only One Can See



Carl Wiens

By George Johnson
Jan. 20, 2014

Since 1955, The Journal of Irreproducible Results has offered "spoofs, parodies, whimsies, burlesques, lampoons and satires"

https://www.economist.com/leaders/2013/10/21/how-science-goes-wrong
https://www.nytimes.com/2014/01/21/science/new-truths-that-only-one-can-see.html

Reproducibility Crisis: ...and Scientific Essay/Paper

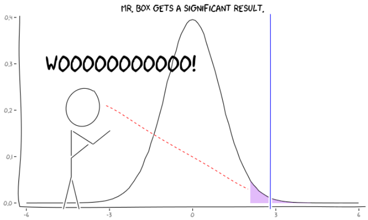
Statistical "Discoveries" and Effect-Size Estimation

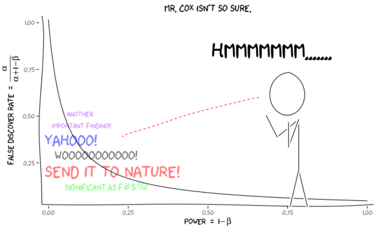
BRANKO SORIĆ*

Essay

Why Most Published Research Findings Are False

John P. A. Ioannidis





Soric (1989); Ioannidis (2005)

Most Discoveries Might Be False (Ioannidis, 2005)

Naive Hypothesis Testing

- $p = 100,000$ hypotheses (brain voxels), only $2,000$ are important.
- Testing at 5% significant level, assume all important variables are selected:

$$\text{False Discovery Proportion} = \frac{5\% \times 98,000}{2000 + 5\% \times 98,000} \approx 70\%$$

Most Discoveries Might Be False (Ioannidis, 2005)

Naive Hypothesis Testing

- $p = 100,000$ hypotheses (brain voxels), only $2,000$ are important.
- Testing at 5% significant level, assume all important variables are selected:

$$\text{False Discovery Proportion} = \frac{5\% \times 98,000}{2000 + 5\% \times 98,000} \approx 70\%$$

False Discovery Rate (Benjamini and Hochberg, 1995)

- False Discovery Rate: the average number of *false discoveries* made among all discoveries.
- FDR control is less conservative than Family-Wise Error Rate control

Marginal Inference

- $\mathrm{X} \in \mathbb{R}^{n \times p}$, $\mathrm{y} \in \mathbb{R}^n$. Example: X is MRI data, y outcome
- Linear Model
$$\mathrm{y} = \mathrm{X}\boldsymbol{\beta}^0 + \sigma\boldsymbol{\xi},$$
with $\sigma > 0$, $\boldsymbol{\xi} \sim \mathcal{N}(0, \mathrm{I}_n)$
- Support set $\mathcal{S} \stackrel{\Delta}{=} \{j \in [p] \,|\, \beta_i^0 \neq 0\}$;
- Objective: find $\hat{\mathcal{S}} \subset \mathcal{S}$ as large as possible

Marginal Inference

- $\mathrm{X} \in \mathbb{R}^{n \times p}$, $\mathrm{y} \in \mathbb{R}^n$. Example: X is MRI data, y outcome
- Linear Model

$$\mathrm{y} = \mathrm{X}\boldsymbol{\beta}^0 + \sigma\boldsymbol{\xi},$$

  with $\sigma > 0$, $\boldsymbol{\xi} \sim \mathcal{N}(0, \mathrm{I}_n)$
- Support set $\mathcal{S} \triangleq \{j \in [p] \,|\, \beta_i^0 \neq 0\}$;
- Objective: find $\hat{\mathcal{S}} \subset \mathcal{S}$ as large as possible

Marginal Testing

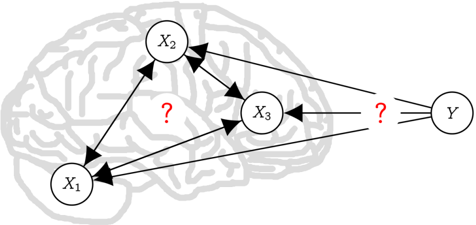
For each $j = 1, \ldots p$:

$$\text{(null) } \mathcal{H}_0^j : X_{*,j} \perp y \quad \text{vs.} \quad \text{(alternative) } \mathcal{H}_\alpha^j : X_{*,j} \not\perp y$$

$\longrightarrow$ FDR control: easy, solvable problem (Poldrack et al., 2012)

Conditional Inference

Source: Weichwald et al. (2015)



Conditional Independence Testing

Generalized Linear Model (GLM): $\mathrm{y} = g(\mathrm{X}\boldsymbol{\beta}^0) + \sigma\boldsymbol{\xi}$

Testing variable j but also taking interaction with other variables X_{-j}

$$\text{(null) } \mathcal{H}_0^j : X_{*,j} \perp y \mid \mathrm{X}_{-j} \quad \text{vs.} \quad \text{(alternative) } \mathcal{H}_\alpha^j : X_{*,j} \not\perp y \mid \mathrm{X}_{-j},$$

or, equivalently

$$\text{(null) } \mathcal{H}_0^j : \beta_j^0 = 0 \quad \text{vs.} \quad \text{(alternative) } \mathcal{H}_\alpha^j : \beta_j^0 \neq 0.$$

FDR control with Conditional Inference

Conditional inference is challenging in high-dimensional settings: how to obtain statistical guarantee: p-value, confidence interval?

$\longrightarrow$ FDR controlling?

[1] Barber and Candès (2015); Candès et al. (2018)

FDR control with Conditional Inference

Conditional inference is challenging in high-dimensional settings: how to obtain statistical guarantee: p-value, confidence interval?

$\longrightarrow$ FDR controlling?

Knockoff Inference [1]

State-of-the-art in high-dimension conditional inference with guaranteed FDR control

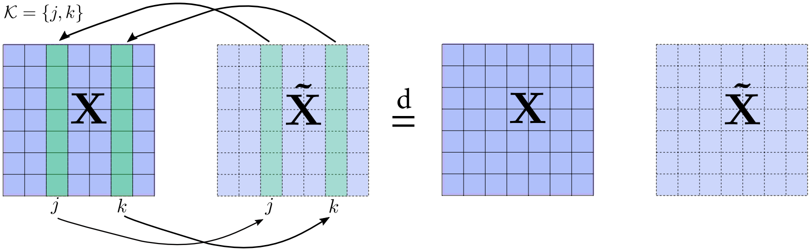
[1] Barber and Candès (2015); Candès et al. (2018)

Knockoff Inference

Knockoff variables (Candès et al., 2018)

$\tilde{\mathrm{X}} = (\tilde{\mathrm{x}}_1, \ldots, \tilde{\mathrm{x}}_p)$ is model-X knockoff variables of $\mathrm{X} = (\mathrm{x}_1, \ldots, \mathrm{x}_p)$ iff:

1. For all subset $\mathcal{K} \subset \{1, \ldots, p\}$: $(\mathrm{X}, \tilde{\mathrm{X}})_{\mathtt{swap}(\mathcal{K})} \stackrel{d}{=} (\mathrm{X}, \tilde{\mathrm{X}})$
2. $\tilde{\mathrm{X}} \perp \mathrm{y} \mid \mathrm{X}$



Knockoff variables: *noisy copies* of original variables

Knockoff Inference

Step 1 – Model-X Knockoff

Assuming distribution of X is known, construct knockoff variables, concatenate $[\mathrm{X}, \tilde{\mathrm{X}}] \in \mathbb{R}^{n \times 2p}$

Step 2

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

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

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

Knockoff Inference

Step 3 – FDR control threshold

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

$$\widehat{\mathrm{FDP}}(t) = \frac{1 + \#\{j \in [p] \mid W_j \leq -t\}}{\#\{j \in [p] \mid W_j \geq t\} \vee 1}$$

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

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

Step 4

Select the variables: $\hat{S}(\tau) = \{j \in [p] \mid W_j \geq \tau\}$

FDP estimation with Knockoff Statistic

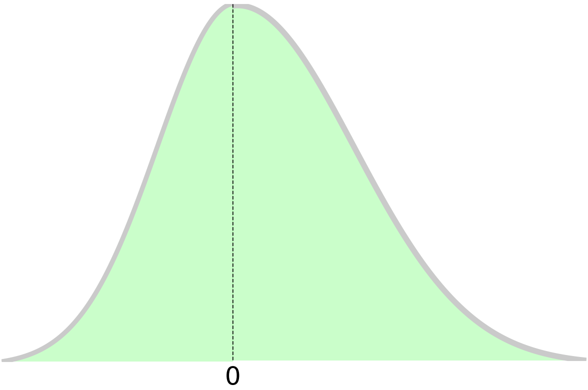


Figure: Knockoff Statistic Distribution

FDP estimation with Knockoff Statistic

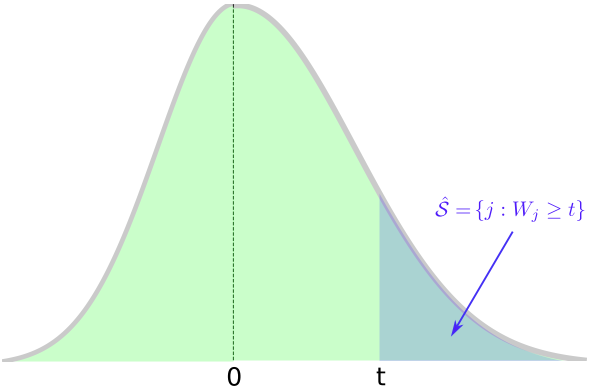


Figure: Knockoff Statistic Distribution

FDP estimation with Knockoff Statistic

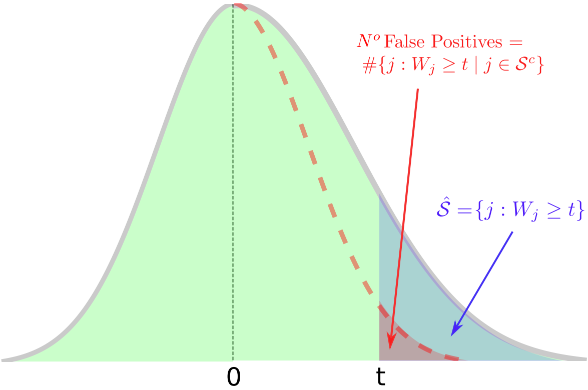


Figure: Knockoff Statistic Distribution

FDP estimation with Knockoff Statistic

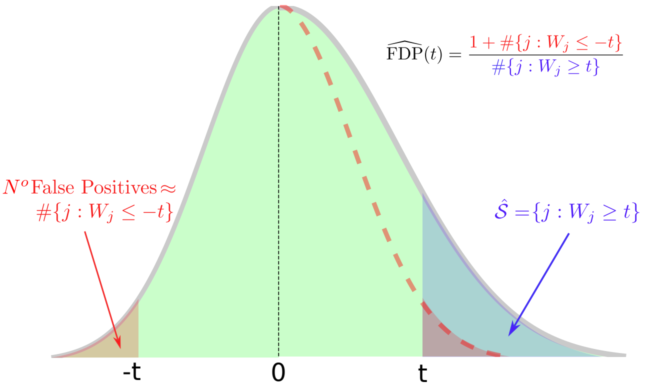


Figure: Knockoff Statistic Distribution

Candès et al. (2018, Lemma 3.3): Under $\mathcal{H}_0^j : \beta_j^0 = 0$, the distribution of W_j is symmetric around 0, *i.e.* $(W_j, -W_k)$ are exchangeable.

Knockoff Inference: Theoretical Guarantee

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

$$\mathrm{FDR}(\tau) = \mathbb{E}\left[\frac{|\hat{S}(\tau) \cap \mathcal{S}^c|}{|\hat{S}(\tau)| \vee 1}\right] \le \alpha,$$

where $\mathcal{S}^c = [p] \backslash \mathcal{S}$: set of null index.

- Result is non-asymptotic.
- Model-X assumption: distribution of X is known.
- Proof: using martingale theory (optional stopping time theorem).

Knockoff Inference: Theoretical Guarantee

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

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

where $\mathcal{S}^c = [p] \backslash \mathcal{S}$: set of null index.

- Result is non-asymptotic.
- Model-X assumption: distribution of X is known.
- Proof: using martingale theory (optional stopping time theorem).

⚠ Major issue: inference results are random.

Demonstration: Instability of Knockoff Procedure

$$\mathbf{y} = \mathbf{X}\beta^{\mathbf{0}} + \sigma\xi$$

ρ sparsity snr

- $n = 500$, $p = 1000$
- $\mathrm{X} \sim \mathcal{N}(0, \mathbf{\Sigma})$
- $\mathbf{\Sigma} = \begin{bmatrix} 1 & \rho & \rho^2 & \dots & \rho^{p-1} \\ \rho^1 & 1 & \rho & \dots & \rho^{p-2} \\ \vdots & \dots & \ddots & \dots & \vdots \\ \rho^{p-2} & \rho^{p-3} & \dots & 1 & \rho \\ \rho^{p-1} & \rho^{p-2} & \rho^{p-3} & \dots & 1 \end{bmatrix}$, with $\rho \in [0, 1)$
- $\xi \sim \mathcal{N}(0, \mathrm{I}_n)$
- $\text{sparsity} = \dfrac{|\mathcal{S}|}{p}$

Demonstration: Instability of Knockoff Procedure

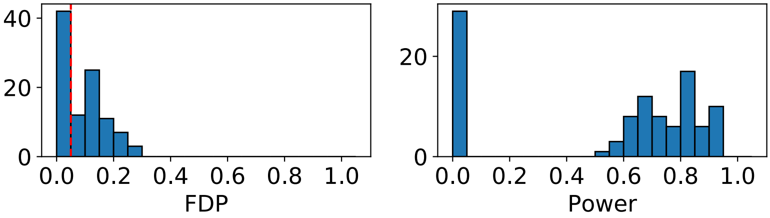
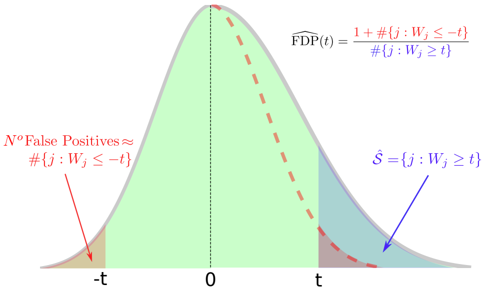


Figure: 100 runs of knockoff inference on the <u>same simulated dataset</u>
n=500, p=1000, snr=3.0, $\rho = 0.7$, sparsity $= 0.06$

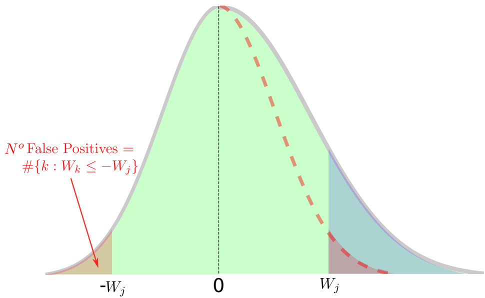
⚠ Large variance on both FDP and Power

Outline

Proposed Solution: Knockoff Statistics conversion



Proposed Solution: Knockoff Statistics conversion



Introduce the intermediate p-values: convert Knockoff statistic W_j to $\hat{p}_j$:

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

AKO – Aggregation of Multiple Knockoffs

- Running multiple sampling of knockoffs, find knockoff statistics
- Convert knockoff statistics to intermediate p-values
- Quantile-aggregation of p-values (Meinshausen et al., 2009)

N., Chevalier, Thirion & Arlot (2020)

AKO – Aggregation of Multiple Knockoffs

- Running multiple sampling of knockoffs, find knockoff statistics
- Convert knockoff statistics to intermediate p-values
- Quantile-aggregation of p-values (Meinshausen et al., 2009)

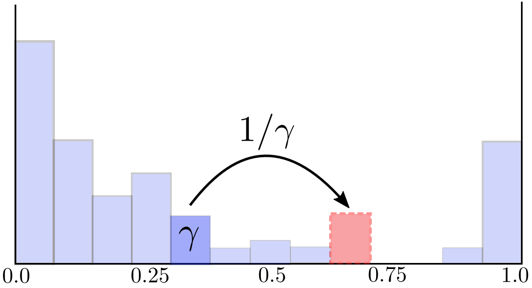
Step 1: For $b = 1, 2, \ldots, B$:

- Run knockoff sampling, calculate test statistic $\{W_j^{(b)}\}_{j=1}^p$
- Convert the test statistic $W_j^{(b)}$ to $\hat{p}_j^{(b)}$:

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

N., Chevalier, Thirion & Arlot (2020)

AKO – Aggregation of Multiple Knockoffs



Step 2 – P-values Aggregation (Meinshausen et al., 2009)

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

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

N., Chevalier, Thirion & Arlot (2020)

AKO – Aggregation of Multiple Knockoffs

Step 3 – FDR control with $\{\bar{\mathrm{p}}_j\}_{j=1}^{p}$

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

$\longrightarrow$ FDR threshold: $\tau = \bar{p}_{(k)}$

Step 4 – Estimate $\hat{S}$

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

N., Chevalier, Thirion & Arlot (2020)

Theoretical Results for AKO

Assumption (Null Distribution of Knockoff Statistic)

The null knockoff statistics $(W_j)_{j \in S^c}$ *are* i.i.d.

Lemma

Under the above assumption, and furthermore assume $|\mathcal{S}^c| \geq 2$*, for all* $j \in \mathcal{S}^c$ *the intermediate p-value* $\hat{p}_j$ *satisfies*

$$\forall t \in (0,1): \quad \mathbb{P}(\hat{p}_j \leq t) \leq \frac{p}{|\mathcal{S}^c|} t$$

Remark

An improved version of Lemma 2, N., Chevalier, Thirion & Arlot (2020).

Theoretical Results for AKO

Theorem (Finite-sample guarantee of FDR control)

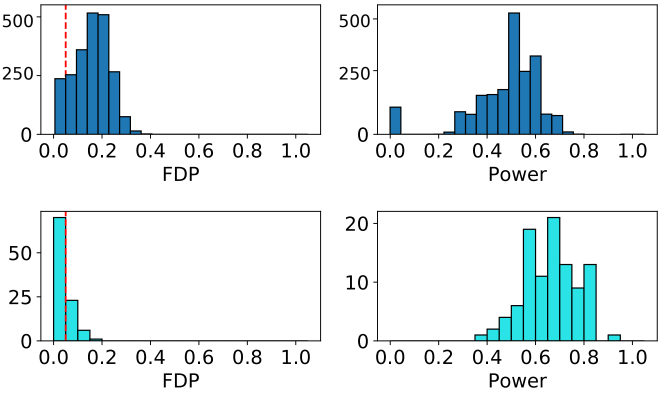
Assuming the null knockoff statistics $(W_j)_{j \in \mathcal{S}^c}$ *are* i.i.d. , *and* $|\mathcal{S}^c| \geq 2$, *then for an arbitrary number of samplings* B, *the output* $\hat{\mathcal{S}}_{AKO}$ *of Aggregation of Multiple Knockoff (AKO) controls FDR under predefined level* $\alpha \in (0, 1)$, i.e.

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

Remark

- An improved version of Theorem 1, N., Chevalier, Thirion & Arlot (2020).
- AKO with $B = 1$ is equivalent to KO.

Experimental Results - Synthetic Data



Histogram of FDP & Power under the same simulated dataset:

- 2500 runs of Original Knockoff (KO – top)
- 100 runs of Aggregated Knockoff (AKO, $B = 25$ – bottom)

Experimental Results - Synthetic Data

- Vary each of the three simulation parameters while keeping the others fixed
- Benchmarking methods:
  - *Ours: Aggregation of Multiple Knockoffs (AKO)*
  - Vanilla Knockoff (KO) (Barber and Candès, 2015; 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

- Vary each of the three simulation parameters while keeping the others fixed

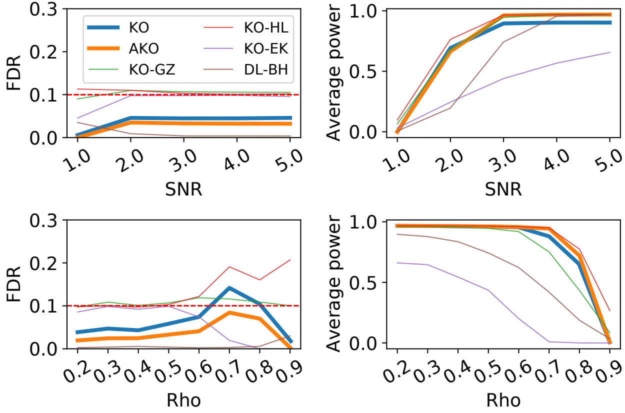


Figure: 100 runs with varying simulation parameters. Default: SNR $= 3.0, \rho = 0.5,$ sparsity $= 0.06$. FDR is controlled at level $\alpha = 0.1$.

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

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

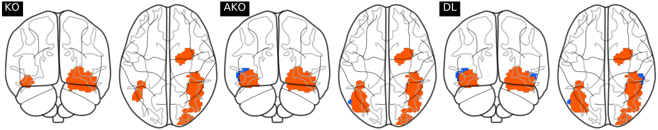


Figure: Detection of significant brain regions for HCP data – Emotion task (face vs. shape) (900 subjects)

- FDR control at $\alpha = 0.1$.
- Orange: brain areas with positive weight.
- Blue: brain areas with negative weight.

Experimental Results - Brain Imaging

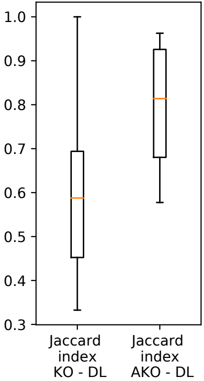


Figure: Jaccard index measuring the Jaccard similarity between the KO/AKO solutions and the Debiased Lasso (DL) solution over 7 tasks of HCP900.

Outline

Binary classification with logistic relationship

- *Binary* response vector $\mathrm{y} \in \{0,1\}^n$.
- Logistic relationship

$$\mathbb{P}(\mathrm{y}_i = 1 \mid \mathrm{X}_{i,*}) = \frac{1}{1 + \exp(-\mathrm{X}_{i,*}^T \boldsymbol{\beta}^0)}.$$

- Estimate β^0 with Penalized Logistic Regression:

$$\hat{\boldsymbol{\beta}}^{\mathrm{PEN}} = \mathrm{argmin}_{\boldsymbol{\beta} \in \mathbb{R}^p} \sum_{i=1}^{n} \log\left[1 + \exp(-y_i(\mathrm{X}_{i,*}^T \boldsymbol{\beta}))\right] + \lambda \left\|\boldsymbol{\beta}\right\|_1 .$$

Penalized Logistic Regression

$$\hat{\boldsymbol{\beta}}^{\mathrm{PEN}} = \mathrm{argmin}_{\boldsymbol{\beta} \in \mathbb{R}^p} \sum_{i=1}^{n} \log\left[1 + \exp(-y_i(\mathrm{X}_{i,*}^T \boldsymbol{\beta}))\right] + \lambda \left\|\boldsymbol{\beta}\right\|_1 .$$

- When $n < p$: hard problem (Sur and Candès, 2019; Zhao et al., 2020) $\longrightarrow$ P-value? Confidence interval? Conditional Independence Testing?
- Original Knockoff Inference: possible with ℓ_1-logistic loss.

Penalized Logistic Regression

$$\hat{\boldsymbol{\beta}}^{\mathrm{PEN}} = \operatorname{argmin}_{\boldsymbol{\beta} \in \mathbb{R}^p} \sum_{i=1}^{n} \log\left[1 + \exp(-y_i(\mathrm{X}_{i,*}^T \boldsymbol{\beta}))\right] + \lambda \left\|\boldsymbol{\beta}\right\|_1 .$$

- When $n < p$: hard problem (Sur and Candès, 2019; Zhao et al., 2020) $\longrightarrow$ P-value? Confidence interval? Conditional Independence Testing?
- Original Knockoff Inference: possible with ℓ_1-logistic loss.

Conditional Randomization Test (CRT)

Candès et al. (2018): An alternative, more straight-forward method to knockoff inference.

Conditional Randomization Test (CRT)

Algorithm 1: Conditional Randomization Test

INPUT dataset (X, y), with $\mathrm{X} \in \mathbb{R}^{n \times p}$, $\mathrm{y} \in \mathbb{R}^{n}$, number of sampling runs B, test statistic T_j, conditional distribution $P_{j|-j}$ for each $j = 1, \ldots, p$;

OUTPUT vector of p-values $\{\hat{p}_j\}_{j=1}^{p}$;

for $j = 1, 2, \ldots, p$ **do**

  for $b = 1, 2, \ldots, B$ **do**

    1. Generate $\tilde{\mathrm{X}}_{*,j}^{(b)}$, a noisy variable from $P_{j|-j}$;
    2. Compute test statistics T_j for original variable and $\tilde{T}_j^{(b)}$ for noisy variables;

  end

  Compute the empirical p-value

$$\hat{p}_j = \frac{1 + \sum_{b=1}^{B} \mathbb{1}_{\{\tilde{T}_j^{(b)} \geq T_j\}}}{1 + B}$$

end

Conditional Randomization Test (CRT)

⚠ Huge computational cost: B inferences for *each* variable j
$\longrightarrow \mathcal{O}(Bp^4)$ with Lasso program to compute T_j

Distillation Conditional Randomization Test (Liu et al., 2020): analytical formula for p-values

- Remove the multiple sampling of noisy variables.
- Pre-screening step: estimate $\hat{\mathcal{S}}^{\text{SCREENING}} \subset [p]$, only calculate test-statistics inside this set.

Distillation Conditional Randomization Test (dCRT)

Algorithm 2: Lasso-dCRT (Liu et al., 2020)

INPUT dataset (X, y), $X \in \mathbb{R}^{n \times p}$, $\mathrm{y} \in \mathbb{R}^n$;
OUTPUT vector of p-values $\{p_j\}_{j=1}^{p}$;
$\hat{\mathcal{S}}^{\mathrm{SCREENING}} = \{j \in [p] \mid \hat{\boldsymbol{\beta}}_j^{\mathrm{PEN}} \neq 0\}$;
for $j \notin \hat{\mathcal{S}}^{SCREENING}$ do
    $p_j = 1$
end
for $j \in \hat{\mathcal{S}}^{SCREENING}$ do
    1. Distill info. of X_{-j} to $\mathrm{X}_{*,j}$ and y, obtain $\hat{\boldsymbol{\beta}}^{d_{\mathrm{X}_{*,j}}}$ and $\hat{\boldsymbol{\beta}}^{d_y,j}$
    2. Obtain test statistic:

$$T_j = \sqrt{n}\, \frac{(\mathrm{y} - \mathrm{X}_{-j}\hat{\boldsymbol{\beta}}^{d_y,j})^T(\mathrm{x}_j - \mathrm{X}_{-j}\hat{\boldsymbol{\beta}}^{d_{\mathrm{X}_{*,j}}})}{\|\mathrm{y} - \mathrm{X}_{-j}\hat{\boldsymbol{\beta}}^{d_y,j}\|_2 \|\mathrm{X}_{*,j} - \mathrm{X}_{-j}\hat{\boldsymbol{\beta}}^{d_{\mathrm{X}_{*,j}}}\|_2}$$

    3. Compute (two-sided) p-value $p_j = 2[1 - \Phi(|T_j|)]$
end

Distillation Operation

For each variable j, *remove* all the conditional information of the remaining variables X_{-j} to $\mathrm{X}_{*,j}$ and to y

Lasso-Distillation

- $\hat{\boldsymbol{\beta}}^{d_y,j} = \mathrm{argmin}_{\boldsymbol{\beta} \in \mathbb{R}^{p-1}} \sum_{i=1}^{n} \log \left[1 + \exp(-y_i(\mathrm{X}_{i,-j}^T \boldsymbol{\beta})) \right] + \lambda \|\boldsymbol{\beta}\|_1$
- $\hat{\boldsymbol{\beta}}^{d_{\mathrm{X}_{*,j}}}(\lambda) = \mathrm{argmin}_{\boldsymbol{\beta} \in \mathbb{R}^{p-1}} \frac{1}{2} \|\mathrm{X}_{*,j} - \mathrm{X}_{-j}\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_1$

dCRT test statistics

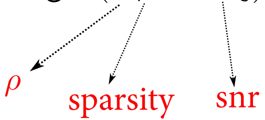
$$T_j = \sqrt{n} \; \frac{(\mathrm{y} - \mathrm{X}_{-j}\hat{\boldsymbol{\beta}}^{d_y,j})^T(\mathrm{x}_j - \mathrm{X}_{-j}\hat{\boldsymbol{\beta}}^{d_{\mathrm{X}_{*,j}}})}{\|\mathrm{y} - \mathrm{X}_{-j}\hat{\boldsymbol{\beta}}^{d_y,j}\|_2 \|\mathrm{X}_{*,j} - \mathrm{X}_{-j}\hat{\boldsymbol{\beta}}^{d_{\mathrm{X}_{*,j}}}\|_2} \xrightarrow[n \to +\infty]{\mathcal{H}_0^j} \mathcal{N}(0,1) \, .$$

conditional to y and X_{-j}

Distillation Operator for Logistic Regression?

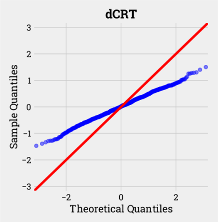
- Lasso-distillation in Liu et al. (2020): model misspecification with logistic relationship
- Demo:

$$\mathbf{y} = \texttt{logit}(\mathbf{X}\boldsymbol{\beta}^0 + \sigma\boldsymbol{\xi})$$

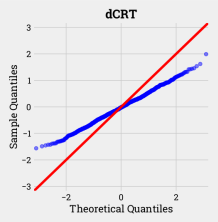


- 100 simulations, $p = 400$, $\mathrm{X} \sim \mathcal{N}(0, \mathbf{\Sigma})$ with $\mathbf{\Sigma}$ a Toeplitz matrix.

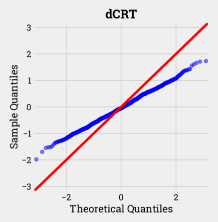
Null distribution of dCRT test statistic



(a) $\mathbf{n = 200}$



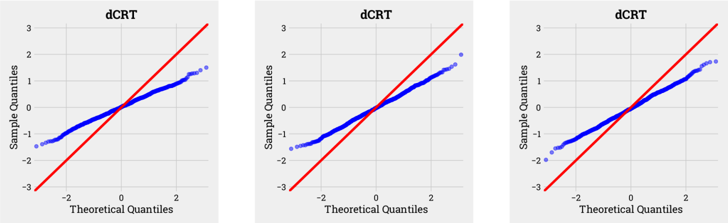
(b) $\mathbf{n = 400}$



(c) $\mathbf{n = 800}$

- QQ-Plot for one null dCRT statistic, 1000 samplings
- Fixed $p = 400$ varying, $n \in \{200, 400, 800\}$
- Theoretical quantile is of a standard Gaussian distribution

Null distribution of dCRT test statistic



(a) $\mathbf{n = 200}$ (b) $\mathbf{n = 400}$ (c) $\mathbf{n = 800}$

- QQ-Plot for one null dCRT statistic, 1000 samplings
- Fixed $p = 400$ varying, $n \in \{200, 400, 800\}$
- Theoretical quantile is of a standard Gaussian distribution

⚠ Null distribution is far from standard normal

Adaptation of CRT to high-dim logistic regresssion

- Ning and Liu (2017): T_j^{decorr} – decorrelating test-statistic T_j
- Finding $\hat{\boldsymbol{\beta}}^{d_y,j}$: find $\hat{\boldsymbol{\beta}}^{\text{PEN}}$, then omitting the jth coefficient
- Finding $\hat{\boldsymbol{\beta}}^{d_{\mathrm{X}_{*,j}}}$: using weighted Lasso instead of standard Lasso.

Adaptation of CRT to high-dim logistic regresssion

- Ning and Liu (2017): T_j^{decorr} – decorrelating test-statistic T_j
- Finding $\hat{\boldsymbol{\beta}}^{d_y,j}$: find $\hat{\boldsymbol{\beta}}^{\text{PEN}}$, then omitting the jth coefficient
- Finding $\hat{\boldsymbol{\beta}}^{d_{\mathrm{X}_{*,j}}}$: using weighted Lasso instead of standard Lasso.

Intuition: based on classical Rao's test score

$$\hat{\boldsymbol{\beta}}^{\text{PEN}} = \text{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^p} \underbrace{\sum_{i=1}^{n} \log\left[1+\exp(-y_i(\mathrm{X}_{i,*}^T\boldsymbol{\beta}))\right]}_{\ell(\boldsymbol{\beta})} + \lambda \|\boldsymbol{\beta}\|_1$$

$$T_j^{\text{Rao}} = n^{1/2}\nabla_{\beta_j}\ell(\boldsymbol{\beta})\hat{\mathrm{I}}_{j|-j}^{-1/2}$$

- In high-dimension, T_j^{Rao} is biased.
- The general formula of decorrelated test score T_j^{decorr} is a debiased version of T_j^{Rao}.

Proposed Solution: CRT-Logit

Algorithm 3: CRT-logit

INPUT dataset (X, y), $\mathrm{X} \in \mathbb{R}^{n \times p}, \mathrm{y} \in \mathbb{R}^{n}$;

OUTPUT vector of p-values $\{p_j\}_{j=1}^{p}$;

$\hat{\boldsymbol{\beta}} \leftarrow \texttt{penalized\_MLE}(\mathrm{X}, \mathrm{y})$; $\hat{\mathcal{S}}^{\texttt{screening}} \leftarrow \{j \in [p] \mid \hat{\beta}_j^{\texttt{MLE}} \neq 0\}$;

for $j \notin \hat{\mathcal{S}}^{screening}$ **do**

  $p_j = 1$

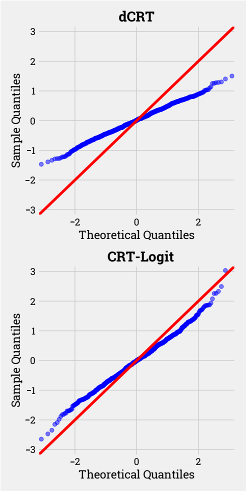
end

for $j \in \hat{\mathcal{S}}^{screening}$ **do**

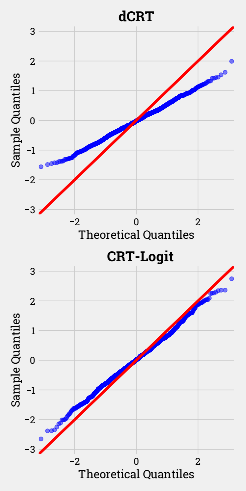
  1. $\hat{\boldsymbol{\beta}}^{d_{\mathrm{X}_{*,j}}} \leftarrow \texttt{scaled\_lasso}(\mathrm{X}_{*,j}, \mathrm{X}_{*,-j})$
  2. $\hat{\boldsymbol{\beta}}^{d_y, j} \leftarrow (\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_{j-1}, \hat{\beta}_{j+1}, \ldots, \hat{\beta}_p)$
  3. $T_j^{\texttt{decorr}} \leftarrow \texttt{decorrelated\_test\_score}(\mathrm{X}, \mathrm{y})$
  4. $p_j \leftarrow 2[1 - \Phi(|T_j^{\texttt{decorr}}|)]$

end

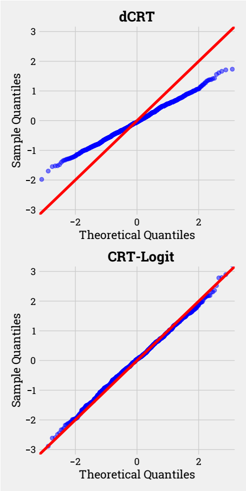
Effectiveness of decorrelation on test statistics



(a) n = 200

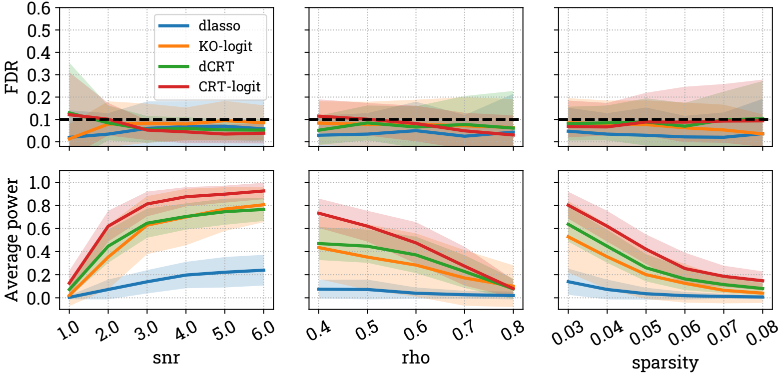


(b) n = 400



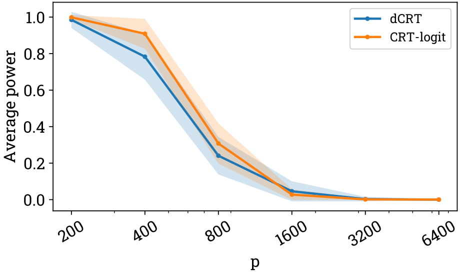
(c) n = 800

Simulation: Mildly High-dimensional Scenario



- 100 runs of simulations across varying parameters; FDR controlled $\alpha = 0.1$.
- Methods: Debiased Lasso (`dlasso`), model-X Knockoff (`KO-logit`), original dCRT (`dCRT`), our version of CRT (`CRT-logit`).

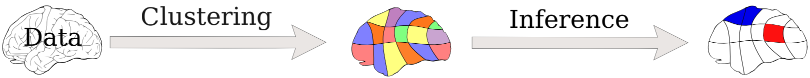
Problem: Curse of Dimensionality



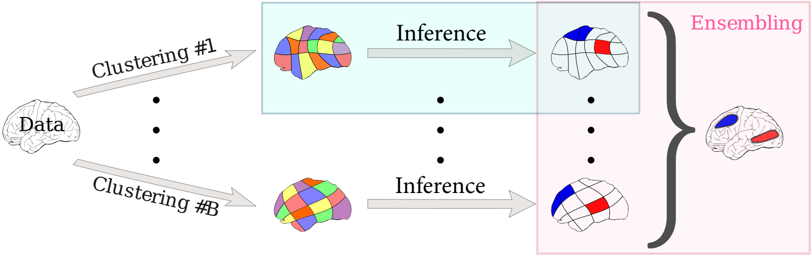
⚠ Failure of detecting variables when dimension grows large.

Inference with Clusters of Variables

- Solution: Dimension reduction via spatially constrained clustering $p \longrightarrow C$ such that $C \ll p$: `cCRT-logit`

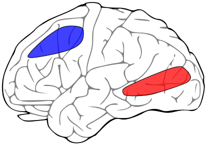


- Stabilize inference results with multiple clusterings + p-values aggregation (`cCRT-logit-agg`):



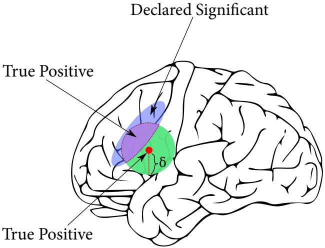
Statistical inference with spatial tolerance

- Brain spatial organization: "close" voxels $\leftrightarrow$ "close" weights

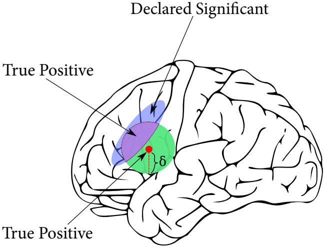


- Null weight voxels
- Positive weight voxels
- Negative weight voxels

- Spatial tolerance δ for false discoveries: FDR^{δ}



False Discovery Rate with spatial tolerance



- Distance between voxels: $d(j,k)$ for $(j,k) \in [p]^2$
- δ-null region: $N^\delta = \{j \in [p] \mid \forall k \in [p], d(j,k) \leq \delta \implies \beta^0_k = 0\}$

FDP^δ and FDR^δ

Given an estimation of the support $\hat{\mathcal{S}}$:

$$\text{FDP}^\delta = \frac{|\{N^\delta \cap \hat{\mathcal{S}}\}|}{|\hat{\mathcal{S}}| \vee 1}$$

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

Theoretical Results for CRT-logit

Estimate support, for $\alpha \in (0,1)$:

- $\hat{\mathcal{S}}_{\texttt{cCRT-logit}} = \texttt{FDR\_control}(\{\hat{p}_j^{\texttt{cCRT-logit}}\}_{j=1}^{p}, \alpha)$
- $\hat{\mathcal{S}}_{\texttt{cCRT-logit-agg}} = \texttt{FDR\_control}(\{\hat{p}_j^{\texttt{cCRT-logit}}\}_{j=1}^{p}, \alpha)$

Conjecture

If the clusters are independent, and all the clusters from all partitions considered have a diameter smaller than δ, and the variables located between clusters are positively correlated, then, the output $\hat{\mathcal{S}}_{\texttt{cCRT-logit}}$ and $\hat{\mathcal{S}}_{\texttt{cCRT-logit-agg}}$ control FDR^{δ} under predefined level $\alpha \in (0,1)$, *i.e.*

$$\limsup_{n\to\infty} \mathbb{E}\left[\frac{|\hat{\mathcal{S}}_{\texttt{cCRT-logit}} \cap N^{\delta}|}{|\hat{\mathcal{S}}_{AKO}| \vee 1}\right] \le \alpha$$

and

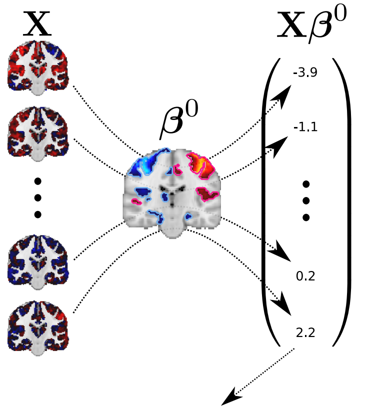
$$\limsup_{n\to\infty} \mathbb{E}\left[\frac{|\hat{\mathcal{S}}_{\texttt{cCRT-logit-agg}} \cap N^{\delta}|}{|\hat{\mathcal{S}}_{AKO}| \vee 1}\right] \le \alpha$$

where N^{δ} is the δ-null region defined above.

Semi-simulated dataset (HCP 900)

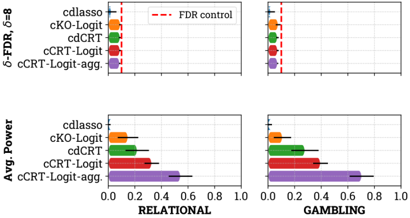
Semi-simulated dataset:

- Use real data X (*e.g.* emotion task).
- build β^0 independently from data of different task, *e.g.* $X_{\texttt{motor\_foot}}$.
- Generate synthetic responses y from $X_{\texttt{emotion}}$ and β^0.



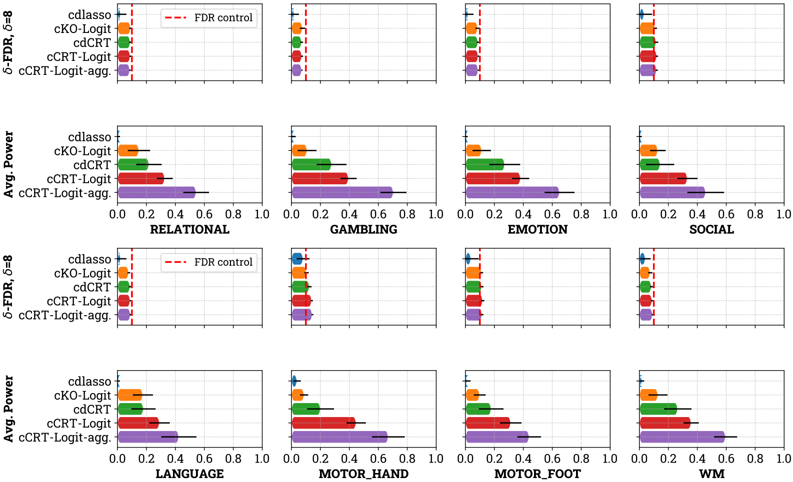
$$\mathbb{P}(\mathbf{y}_i = 1 \mid \mathbf{X}_{i,*}) = \frac{1}{1 + \exp(-\mathbf{X}_{i,*}\boldsymbol{\beta}^0 + \sigma\xi_i)}$$

Semi-simulated dataset (HCP 900)

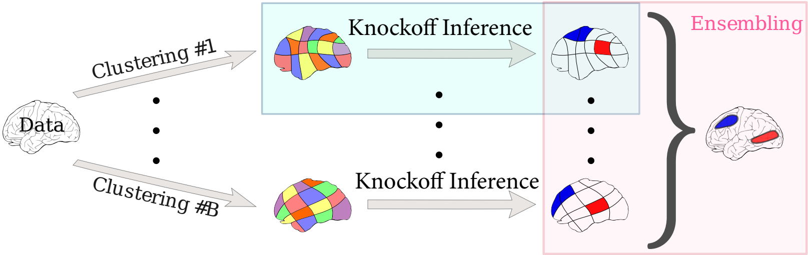


- FDR/Average Power of 50 runs of simulations on Human Brain Connectome dataset.
- Parameters: $n = 800$ (taken from 400 subjects), SNR $= 1.5$. FDR^{δ} is controlled at level $\alpha = 0.1$ and $\delta = 8$.
- Methods (clustering versions): Desparsified Lasso (`cdlasso`), model-X Knockoff (`cKO-logit`), original dCRT (`cdCRT`), our version of CRT (`cCRT-logit`) and the aggregation of CRT-logit across clusterings (`cCRT-logit-agg.`)

Semi-simulated dataset (HCP 900)



Related: Ensemble of Clustered Knockoffs



Nguyen et al. (2019), journal version in progress

Outline

Summary

New procedures for statistical inference with high-dimensional data

Aggregation of Multiple Knockoffs

- FDR control guarantee.
- Demonstrated empirically: more stable in inference results and higher statistical power.

Conditional Randomization Test for high-dimensional logistic regression (`CRT-logit`)

- Reduce computational cost of original CRT.
- Ensemble of clusterings version works well in very high-dimension.

Remark

Clustered version involves additional assumptions for statistical guarantee.

Perspectives

- Formal statement and proof of the Conjecture on FDR control with CRT-logit
- Theoretical analysis of clustering inference with Knockoffs and CRT-logit: relaxing the assumption on independence of clusters.
- Applications for genomics data.
- Generative networks for knockoff variables generation.

Perspectives

- Formal statement and proof of the Conjecture on FDR control with CRT-logit
- Theoretical analysis of clustering inference with Knockoffs and CRT-logit: relaxing the assumption on independence of clusters.
- Applications for genomics data.
- Generative networks for knockoff variables generation.

Thank you for listening!

Second-order Model-X Knockoffs

Shares the first two moments - mean and covariance, *i.e.* :

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

Additional assumption: X has Gaussian design

$$\longrightarrow \tilde{x}_j \mid x_j \overset{d}{=} \mathcal{N}(\boldsymbol{\mu}, \mathrm{V})$$

$\longrightarrow$ Finding diag{s} by:

- Semi-definite Programming (SDP)
- Approximate Semi-definite program (ASDP)
- Equi-correlated

Knockoff Statistic

Definition (Candès et al. (2018))

A knockoff statistic $\mathrm{W} = \{W_j\}_{j\in[p]}$ is a measure of feature importance that satisfies the two following properties:

1. Depends only on $\mathrm{X}, \tilde{\mathrm{X}}$ and y

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

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

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

Theoretical Results for AKO

Assumption (Null Distribution of Knockoff Statistic)

Under the null hypothesis $H_{0,j} : \beta_j^0 = 0$, *the Knockoff Statistics follow the same null distribution.*

Lemma (Lemma 2 – N., Chevalier, Thirion, Arlot, 2020)

Under the above assumption, and furthermore assume $|\mathcal{S}^c| \geq 2$, *for all* $j \in \mathcal{S}^c$ *the intermediate p-value* $\hat{p}_j$ *satisfies*

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

where

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

Theoretical Results for AKO

Theorem (Theorem 1 – N., Chevalier, Thirion, Arlot, 2020)

(Finite-sample guarantee of FDR control)
If, under the null hypothesis $H_{0,j} : \beta_j^0 = 0$*, the Knockoff Statistics follow the same distribution, and if* $|\mathcal{S}^c| \geq 2$*, then for an arbitrary number of samplings* B*, the output* $\hat{\mathcal{S}}_{AKO}$ *of Aggregation of Multiple Knockoff (AKO) controls FDR under predefined level* $\alpha \in (0, 1)$, i.e.

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

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

AKO extra results - Genome Wide Association Study

- Data: Flowering Phenotype of Arabidopsis Thaliana (FT_GH) – $n = 166, p = 9938$
- 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	—

Table: List of detected genes associated with phenotype FT_GH.

From previous studies: AT2G21070 (Kim et al., 2008), AT4G02780 (Silverstone et al., 1998), AT5G47640 (Cai et al., 2007)

Adaptation of CRT to high-dim logistic regresssion

- $T_j^{\texttt{decorr}}$: Decorrelating test-statistic T_j (Ning and Liu, 2017)
- Finding $\hat{\boldsymbol{\beta}}^{d_y,j}$: find $\hat{\boldsymbol{\beta}}^{\mathrm{PEN}}$, then omitting the jth coefficient, *i.e.*

$$\hat{\boldsymbol{\beta}}_j^{d_y,j} = (\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_{j-1}, \hat{\beta}_{j+1}, \ldots, \hat{\beta}_p)$$

- Finding $\hat{\boldsymbol{\beta}}^{d_{\mathrm{X}_{*,j}}}$: using weighted Lasso instead of standard Lasso.

$$\hat{\boldsymbol{\beta}}^{d_{\mathrm{X}_{*,j}}} = \mathrm{argmin}_{\boldsymbol{\beta} \in \mathbb{R}^{p-1}} \frac{1}{n} \sum_{i=1}^{n} \frac{\exp(\hat{\boldsymbol{\beta}}^T \mathbf{x}_i)}{[1 + \exp(\hat{\boldsymbol{\beta}}^T \mathbf{x}_i)]^2} (x_{i,j} - \boldsymbol{\beta}^T \mathrm{X}_{-j})^2 + \lambda \|\boldsymbol{\beta}\|_1$$

Adaptation of CRT to high-dim logistic regresssion

Decorrelated test statistic

$$T_j^{\texttt{decorr}} =$$

$$-(n\,\hat{\mathrm{I}}_{j|-j})^{-1/2}\sum_{i=1}^{n}\left[y_i - \frac{1}{1+\exp\left(-\mathrm{X}_{i,-j}\hat{\boldsymbol{\beta}}^{d_y,j}\right)}\right]\left[\mathrm{X}_{i,j} - \mathrm{X}_{i,-j}^{T}\hat{\boldsymbol{\beta}}^{d\mathrm{X}_{*,j}}\right],$$

where $\hat{\mathrm{I}}_{j|-j}$ is the estimated partial Fisher information:

$$\hat{\mathrm{I}}_{j|-j} = \frac{1}{n}\sum_{i=1}^{n}\frac{\exp\left(\hat{\boldsymbol{\beta}}\mathrm{X}_{i,*}\right)}{\left[1+\exp\left(\hat{\boldsymbol{\beta}}\mathrm{X}_{i,*}\right)\right]^2}\left(\mathrm{X}_{i,j} - \hat{\boldsymbol{\beta}}^{d_x\,T}\mathrm{X}_{i,-j}\right)^2\,\mathrm{X}_{i,j}\,.$$

Asymptotic distribution (Ning and Liu, 2017)

$$T_j^{\texttt{decorr}} \xrightarrow[n\to+\infty]{\mathcal{H}_0^j} \mathcal{N}(0,1)$$

Assumptions in Clustered Inference

Spatial homogeneity with distance δ

For all $(j, k) \in [p] \times [p]$, $d(j, k) \leq \delta$ implies that $\mathbf{\Sigma}_{j,k} \geq 0$, where $\mathbf{\Sigma}_{j,k} \triangleq \mathrm{Cov}(\mathrm{x}_j, \mathrm{x}_k)$.

Sparse-smooth with distance δ

For all $(j, k) \in [p] \times [p]$, $d(j, k) \leq \delta$ implies that $\mathrm{sign}(\beta^0_j) = \mathrm{sign}(\beta^0_k)$.