

Semi-knockoffs: a model-agnostic Conditional Independence Testing method with finite-sample guarantees

Angel REYERO LOBO

Big Stats seminar
IMT & Inria Paris-Saclay
Joint work with:
Pierre NEUVIAL & Bertrand THIRION.

June 24, 2026



1 Introduction

2 Semi-knockoffs

- Model-X assumption and Exchangeability under the Null
- Oracle Semi-knockoffs
- Practical Semi-knockoffs
- False Discovery Rate (FDR) control

3 Experiments

4 References

1 Introduction

2 Semi-knockoffs

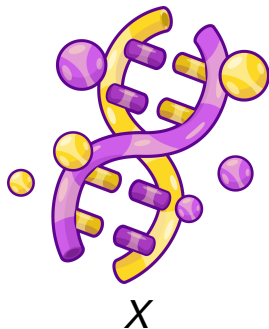
- Model-X assumption and Exchangeability under the Null
- Oracle Semi-knockoffs
- Practical Semi-knockoffs
- False Discovery Rate (FDR) control

3 Experiments

4 References

Scientific Discovery with Machine Learning (ML)

Which covariates X^j are important with respect to an outcome y ?



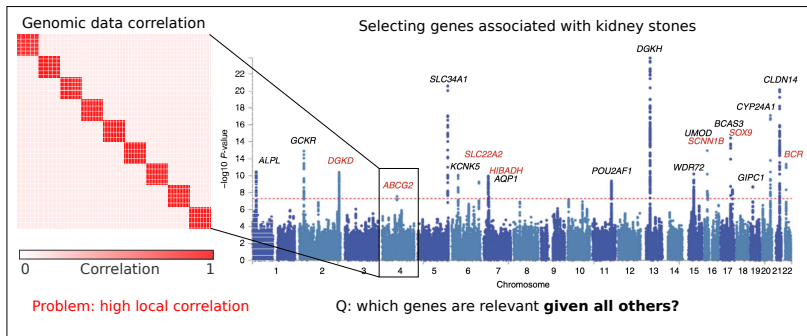
💡 Study their relationship using an **ML model**: $\hat{m} \in \operatorname{argmin}_{f \in \mathcal{F}} \hat{\mathbb{E}}[\mathcal{L}(f(X), y)]$

Conditional Independent Testing

$$\mathcal{H}_1 := \{j : X^j \not\perp y \mid \mathbf{X}^{-j}\}.$$

⚠ **Highly correlated** genomic data

⇒ **marginally**, many genes are important. Isn't it only because of correlation?

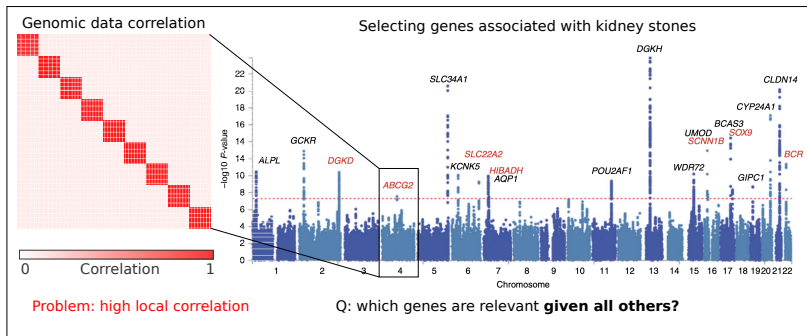


Conditional Independent Testing

$$\mathcal{H}_1 := \{j : X^j \not\perp y \mid \mathbf{X}^{-j}\}.$$

⚠ **Highly correlated** genomic data

⇒ **marginally**, many genes are important. Isn't it only because of correlation?



We need to select relevant variables **conditionally** on all the others!

1 Introduction

2 Semi-knockoffs

- Model-X assumption and Exchangeability under the Null
- Oracle Semi-knockoffs
- Practical Semi-knockoffs
- False Discovery Rate (FDR) control

3 Experiments

4 References

Related work: Model-X and Exchangeability under the Null

Model-X assumption: we can sample a r.v. $\tilde{X}^j \sim \mathcal{L}(X^j | X^{-j})$.

Under the null, for any importance statistic $T : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}$, the statistics

$$T(\textcolor{violet}{X}^j, X^{-j}, y) \quad \text{and} \quad T(\tilde{X}^j, X^{-j}, y)$$

are **exchangeable**.

We need **explainable ML** to quantify the contribution of a feature in T

Related work: Model-X and Exchangeability under the Null

Model-X assumption: we can sample a r.v. $\tilde{X}^j \sim \mathcal{L}(X^j | X^{-j})$.

Under the null, for any importance statistic $T : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}$, the statistics

$$T(\textcolor{violet}{X}^j, X^{-j}, y) \quad \text{and} \quad T(\tilde{X}^j, X^{-j}, y)$$

are **exchangeable**.

We need **explainable ML** to quantify the contribution of a feature in T

- **Coefficient**-based importance: $T(X^j, X^{-j}, y) = |\hat{\beta}^j|$

Related work: Model-X and Exchangeability under the Null

Model-X assumption: we can sample a r.v. $\tilde{X}^j \sim \mathcal{L}(X^j | X^{-j})$.

Under the null, for any importance statistic $T : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}$, the statistics

$$T(\mathbf{x}^j, X^{-j}, y) \quad \text{and} \quad T(\tilde{X}^j, X^{-j}, y)$$

are **exchangeable**.

We need **explainable ML** to quantify the contribution of a feature in T

- **Coefficient**-based importance: $T(X^j, X^{-j}, y) = |\hat{\beta}^j|$
- **Performance**-based importance: $T(X^j, X^{-j}, y) = \ell(\hat{m}(X^j, X^{-j}), y)$

Compare

$$\left\{ \ell\left(\hat{m}(\mathbf{x}^j, X^{-j}), y\right) \right\}_{i=1}^{n_{\text{test}}} \quad \text{with} \quad \left\{ \ell\left(\hat{m}(\tilde{X}^j, X^{-j}), y\right) \right\}_{i=1}^{n_{\text{test}}}, \quad \tilde{X}^j \sim \mathcal{L}(X^j | X^{-j}).$$

Related work: Model-X and Exchangeability under the Null

Model-X assumption: we can sample a r.v. $\tilde{X}^j \sim \mathcal{L}(X^j | X^{-j})$.

Under the null, for any importance statistic $T : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}$, the statistics

$$T(\mathbf{x}^j, X^{-j}, y) \quad \text{and} \quad T(\tilde{X}^j, X^{-j}, y)$$

are **exchangeable**.

We need **explainable ML** to quantify the contribution of a feature in T

- **Coefficient**-based importance: $T(X^j, X^{-j}, y) = |\hat{\beta}^j|$
- **Performance**-based importance: $T(X^j, X^{-j}, y) = \ell(\hat{m}(X^j, X^{-j}), y)$

Compare

$$\left\{ \ell(\hat{m}(\mathbf{x}^j, X^{-j}), y) \right\}_{i=1}^{n_{\text{test}}} \quad \text{with} \quad \left\{ \ell(\hat{m}(\tilde{X}^j, X^{-j}), y) \right\}_{i=1}^{n_{\text{test}}}, \quad \tilde{X}^j \sim \mathcal{L}(X^j | X^{-j}).$$

A train–test split is required to preserve exchangeability.

Exchangeability under the Null

- Current methods (e.g., HRT) rely on exchangeability between

$$\left\{ \ell \left(\hat{m}(\tilde{X}^j, X^{-j}), y \right) \right\}_{i=1}^{n_{\text{test}}} \quad \text{with} \quad \left\{ \ell \left(\hat{m}(X^j, X^{-j}), y \right) \right\}_{i=1}^{n_{\text{test}}}, \quad \tilde{X}^j \sim \mathcal{L}(X^j \mid X^{-j}).$$

- ☹ This requires a train–test split, which significantly reduces the method’s performance.

Exchangeability under the Null

- Current methods (e.g., HRT) rely on exchangeability between

$$\left\{ \ell\left(\hat{m}(\tilde{X}^j, X^{-j}), y\right) \right\}_{i=1}^{n_{\text{test}}} \quad \text{with} \quad \left\{ \ell\left(\hat{m}(X^j, X^{-j}), y\right) \right\}_{i=1}^{n_{\text{test}}}, \quad \tilde{X}^j \sim \mathcal{L}(X^j \mid X^{-j}).$$

☹ This requires a train–test split, which significantly reduces the method’s performance.

- 💡 **Semi-knockoffs:** We instead propose comparing

$$\left\{ \ell\left(\hat{m}(\tilde{X}_1^j, X^{-j}), y\right) \right\}_1^n \quad \text{vs} \quad \left\{ \ell\left(\hat{m}(\tilde{X}_2^j, X^{-j}), y\right) \right\}_1^n \quad \text{with}$$
$$\tilde{X}_1^j \sim \mathcal{L}(X^j \mid X^{-j}), \quad \tilde{X}_2^j \sim \mathcal{L}(X^j \mid X^{-j}, Y).$$

Key insight: If $X^j \perp\!\!\!\perp y \mid X^{-j}$, $\mathcal{L}(X^j \mid X^{-j}) = \mathcal{L}(X^j \mid X^{-j}, y)$, so they are **exchangeable**.

- We compare **perturbed** vs. **perturbed** instead of **perturbed** vs. **original** data.
- No train–test split is required.
- The method works with any predictive model $\hat{m}$.

No Need to Sample from $\mathcal{L}(X^j | X^{-j}, y)$

Problem: Directly sampling $\tilde{X}_2^{(j)} \sim \mathcal{L}(X^j | X^{-j}, y)$ is complex.

Exchangeability under H_0^j :

$$\rho_j(X^{-j}, y) := \mathbb{E}[X^j | X^{-j}, y] = \mathbb{E}[X^j | X^{-j}] =: \mathbf{v}_j(X^{-j})$$

Idea: Construct new coordinates using *residuals*.

$$\tilde{X}_1^{(j)} = \mathbf{v}_j(X^{-j}) + \underbrace{x_l^j - \mathbf{v}_j(x_l^{-j})}_{\varepsilon_{j,1}^l}, \quad \tilde{X}_2^{(j)} = \rho_j(X^{-j}, y) + \underbrace{x_k^j - \rho_j(x_k^{-j}, y_k)}_{\varepsilon_{j,2}^k}$$

- Residuals $\varepsilon_{j,1}, \varepsilon_{j,2}$ are then exchangeable.
- Avoids conditional sampling.

Algorithm (Type-I error control)

Algorithm 1 Semi-knockoffs Wilcoxon (SKO-W_{CX})

Input: Model $\hat{m}$, data $\{(X_i, y_i)\}_i^n$, significance level α , a feature j

(Regression 1) Fit $\hat{v}_j \approx \mathbb{E}[X^j \mid X^{-j}]$

(Regression 2) Fit $\hat{\rho}_j \approx \mathbb{E}[X^j \mid X^{-j}, Y]$

Compute residuals $\hat{\epsilon}_{j,1} = X^j - \hat{v}_j(X^{-j})$

Compute residuals $\hat{\epsilon}_{j,2} = X^j - \hat{\rho}_j(X^{-j}, Y)$

Sample $\pi_{j,1}$ and $\pi_{j,2}$ permutations of $\{1, \dots, n\}$

Construct

$$\tilde{X}'_{1,i} = \hat{v}_j(X_i^{-j}) + \hat{\epsilon}_{j,1,\pi_{j,1}(i)}$$

$$\tilde{X}'_{2,i} = \hat{\rho}_j(X_i^{-j}, y_i) + \hat{\epsilon}_{j,2,\pi_{j,2}(i)}$$

Return Wilcoxon test between $\left\{ \ell(\hat{m}(\tilde{X}'_{1,i}), y_i) \right\}_i$ and $\left\{ \ell(\hat{m}(\tilde{X}'_{2,i}), y_i) \right\}_i$

Algorithm (Derandomizing the permutations)

Algorithm 2 Derandomized Semi-knockoffs Wilcoxon ($\text{SKO-WCX-P}(B)$)

Input: Model $\hat{m}$, data $\{(X_i, y_i)\}_i^n$, significance level α , a feature j

(Regression 1) Fit $\hat{v}_j \approx \mathbb{E}[X^j \mid X^{-j}]$

(Regression 2) Fit $\hat{\rho}_j \approx \mathbb{E}[X^j \mid X^{-j}, Y]$

Compute residuals $\hat{\epsilon}_{j,1} = X^j - \hat{v}_j(X^{-j})$

Compute residuals $\hat{\epsilon}_{j,2} = X^j - \hat{\rho}_j(X^{-j}, Y)$

Sample $\{\pi_{j,1}^b\}_b^B$ and $\{\pi_{j,2}^b\}_b^B$ permutations of $\{1, \dots, n\}$

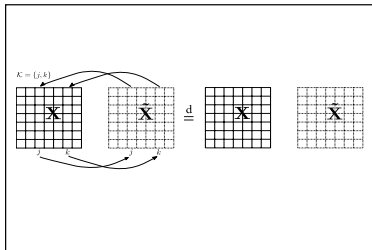
Construct

$$\tilde{X}_{1,i,b}^{(j)} = \hat{v}_j(X_i^{-j}) + \hat{\epsilon}_{j,1,\pi_{j,1}^b(i)}$$

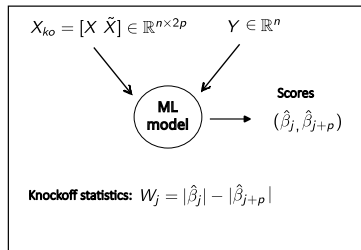
$$\tilde{X}_{2,i,b}^{(j)} = \hat{\rho}_j(X_i^{-j}, y_i) + \hat{\epsilon}_{j,2,\pi_{j,2}^b(i)}$$

Return Wilcoxon test between $\left\{ \frac{1}{B} \sum_b \ell(\hat{m}(\tilde{X}_{1,i,b}^{(j)}), y_i) \right\}_i$ and $\left\{ \frac{1}{B} \sum_b \ell(\hat{m}(\tilde{X}_{2,i,b}^{(j)}), y_i) \right\}_i$

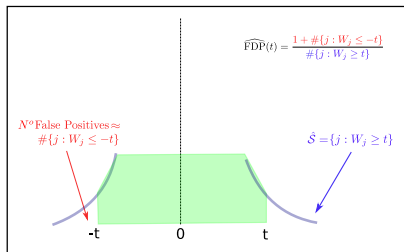
Knockoffs framework (Candès et al. 2018)



Knockoff variables



Building Knockoff statistics



Inferring on Knockoff statistics

False Discovery Rate (FDR) Control

For multiple testing, we aim to control the $\text{FDR}(\hat{S}) = \mathbb{E} \left[\frac{|\hat{S} \cap \mathcal{H}_0|}{|\hat{S}| \vee 1} \right]$.

Knockoffs ingredients:

(1) Knockoffs variables ($\tilde{X}$) (2) Knockoffs statistics (W_j) (3) Knockoffs threshold (T_q^*).

We then select

$$\hat{S} = \{j : W_j \geq T_q^*\}.$$

Theorem 1 (Knockoffs FDR control)

If the statistics W_j satisfy the sign-flip property

$$W_j \stackrel{d}{=} \varepsilon_j W_j, \quad \varepsilon_j \stackrel{\text{i.i.d.}}{\sim} \text{Unif}\{\pm 1\}, \quad j \in \mathcal{H}_0,$$

then the knockoffs procedure controls the FDR in finite samples.

False Discovery Rate (FDR) Control

For multiple testing, we aim to control the $\text{FDR}(\hat{S}) = \mathbb{E} \left[\frac{|\hat{S} \cap \mathcal{H}_0|}{|\hat{S}| \vee 1} \right]$.

Knockoffs ingredients:

(1) Knockoffs variables ($\tilde{X}$) (2) Knockoffs statistics (W_j) (3) Knockoffs threshold (T_q^*).

We then select

$$\hat{S} = \{j : W_j \geq T_q^*\}.$$

Theorem 1 (Knockoffs FDR control)

If the statistics W_j satisfy the sign-flip property

$$W_j \stackrel{d}{=} \varepsilon_j W_j, \quad \varepsilon_j \stackrel{\text{i.i.d.}}{\sim} \text{Unif}\{\pm 1\}, \quad j \in \mathcal{H}_0,$$

then the knockoffs procedure controls the FDR in finite samples.

⚠ Knockoffs variables are difficult to construct (Blain et al., 2025).

⚠ Most existing statistics are designed for linear models.

Semi-knockoffs for False Discovery Rate (FDR) Control

Our **semi-knockoff** statistic is

$$W_j^{\text{SKO}} := \frac{1}{n} \sum_{i=1}^n \left[\ell\left(\hat{m}(\tilde{X}_{1,i}^{(j)}), y_i\right) - \ell\left(\hat{m}(\tilde{X}_{2,i}^{(j)}), y_i\right) \right],$$

and satisfies the required symmetry, allowing the knockoff threshold to be applied.

- $\tilde{X}_{1,i}^{(j)}$ and $\tilde{X}_{2,i}^{(j)}$ are *not* valid knockoff variables.
- W_j^{SKO} is *not* a valid knockoff statistic.
- ☺ Still, we have FDR control and can be combined with *any* ML model.

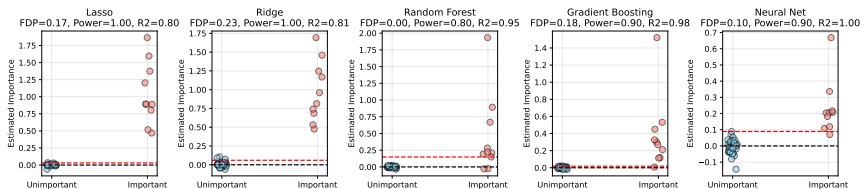


Figure 1: **Semi-knockoff statistics.** Null W_j^{SKO} (left) are symmetric around zero.

1 Introduction

2 Semi-knockoffs

- Model-X assumption and Exchangeability under the Null
- Oracle Semi-knockoffs
- Practical Semi-knockoffs
- False Discovery Rate (FDR) control

3 Experiments

4 References

Simulated data

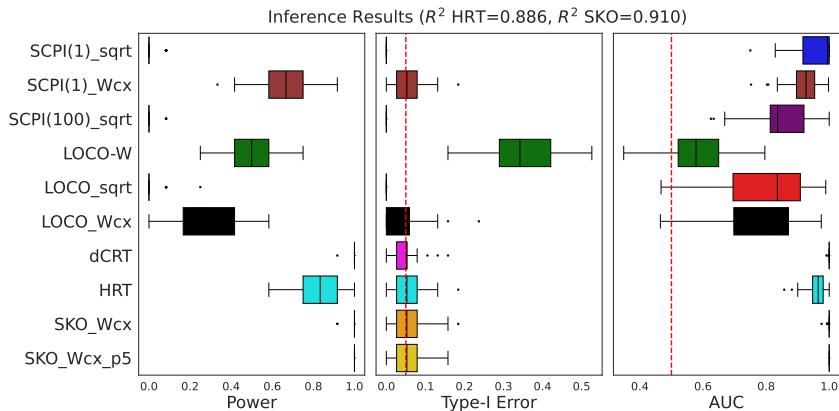


Figure 2: SKO controls type-I error with high power by avoiding a train–test split. Simulated linear model with correlated Gaussian covariates ($p = 50$, $n = 300$) and a gradient boosting.

Real data

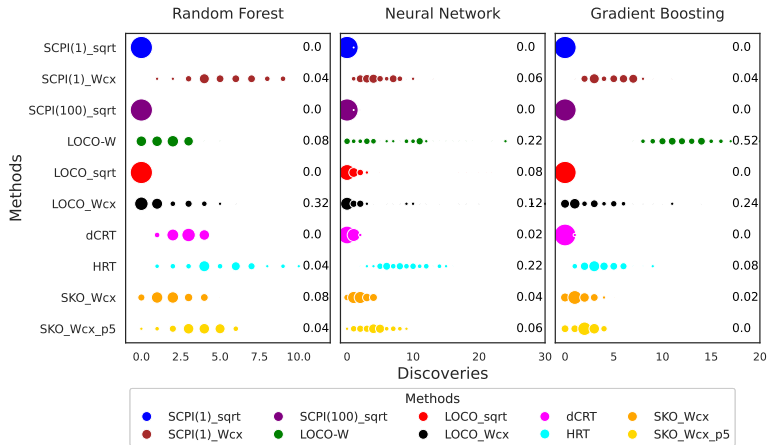


Figure 3: Discoveries on real data across ML models. The type-I error (shown to the right of each plot) is estimated using an artificial null feature correlated at 0.6 with the original features.

- **A flexible conditional independence test** compatible with *any* predictor $\hat{m}$.
- **Finite-sample guarantees**
 - *Type-I* error control via nonparametric paired tests (e.g., Wilcoxon).
 - *FDR* control via the knockoff threshold.
- **Conditional expectations** provide a practical alternative to densities for modern ML.
- **Theory for estimated imputers** ($\hat{v}_j, \hat{\rho}_j$)
 - *Stability* of regularized ML models when adding a null feature.
 - *Double robustness* yielding faster convergence rates.

1 Introduction

2 Semi-knockoffs

- Model-X assumption and Exchangeability under the Null
- Oracle Semi-knockoffs
- Practical Semi-knockoffs
- False Discovery Rate (FDR) control

3 Experiments

4 References

References

- Alexandre Blain, Angel Reyero Lobo, Julia Linhart, Bertrand Thirion, and Pierre Neuvial. When knockoffs fail: diagnosing and fixing non-exchangeability of knockoffs, 2025. URL <https://arxiv.org/abs/2407.06892>.
- Emmanuel Candès, Yingying Fan, Lucas Janson, and Jinchi Lv. Panning for gold: ‘model-X’ knockoffs for high dimensional controlled variable selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(3):551–577, 2018.
- Molei Liu, Eugene Katsevich, Lucas Janson, and Aaditya Ramdas. Fast and powerful conditional randomization testing via distillation. *Biometrika*, 109(2):277–293, June 2022. doi: 10.1093/biomet/asab039.
- Angel Reyero-Lobo, Bertrand Thirion, and Pierre Neuvial. Semi-knockoffs: a model-agnostic conditional independence testing method with finite-sample guarantees. In *Proceedings of the 43rd International Conference on Machine Learning (ICML)*, 2026.

Wesley Tansey, Victor Veitch, Haoran Zhang, Raul Rabadan, and David M. Blei and. The holdout randomization test for feature selection in black box models. *Journal of Computational and Graphical Statistics*, 31(1):151–162, 2022. doi: 10.1080/10618600.2021.1923520. URL <https://doi.org/10.1080/10618600.2021.1923520>.

Thank You, Questions?



- Angel Reyero-Lobo, Bertrand Thirion, and Pierre Neuvial. Semi-knockoffs: a model-agnostic conditional independence testing method with finite-sample guarantees. In Proceedings of the 43rd International Conference on Machine Learning(ICML), 2026.

Knockoffs framework

Knockoffs (from Candès et al. 2018):

① **Knockoff variables** ($\tilde{X}$): *imitations* of X independent of y :

- $[X, \tilde{X}]_{\text{swap}(s)} \stackrel{d}{=} [X, \tilde{X}]$ for any $s \subset [p]$.

$$\begin{bmatrix} X_1, X_2, X_3, \tilde{X}_1, \tilde{X}_2, \tilde{X}_3 \end{bmatrix}_{\text{swap}(1,3)} := \begin{bmatrix} \tilde{X}_1, X_2, \tilde{X}_3, X_1, \tilde{X}_2, X_3 \end{bmatrix}$$

- $\tilde{X} \perp\!\!\!\perp y \mid X$.

Example 2 (Gaussian knockoffs)

If $X \sim \mathcal{N}(0, \Sigma)$, then $[X, \tilde{X}] \sim \mathcal{N}(0, G)$ with $G = \begin{bmatrix} \Sigma & \Sigma - \text{diag}(\lambda) \\ \Sigma - \text{diag}(\lambda) & \Sigma \end{bmatrix}$ for $\lambda \in \mathbb{R}^p$
such that $G \succeq 0$:

- $[X, \tilde{X}]$ is **exchangeable**.
- $\tilde{X} \mid X$ is known!

Knockoffs framework

Knockoffs (from Candès et al. 2018):

- 1 **Knockoff variables** ($\tilde{X}$): *imitations* of X independent of y :
- 2 **Knockoffs statistics** ($W \in \mathbb{R}^p$): comparison of *predictability* between the **original** covariate and its **knockoff**.

Example 2 (Lasso Coefficients Difference)

Regress $y \in \mathbb{R}$ on $[X, \tilde{X}] \in \mathbb{R}^{2p}$ using LASSO and compute the estimated coefficient difference: $W_j = |\hat{\beta}_j| - |\hat{\beta}_{j+p}|$.

Knockoffs framework

Knockoffs (from Candès et al. 2018):

- 1 **Knockoff variables** ($\tilde{X}$): *imitations* of X independent of y :
- 2 **Knockoffs statistics** ($W \in \mathbb{R}^p$): comparison of *predictability* between the *original* covariate and its *knockoff*.
- 3 **Threshold**: it is data-dependent and given by

$$T_q^* = \min \left\{ t \in \mathcal{W} : \frac{1 + \#\{j : W_j \leq -t\}}{\#\{j : W_j \geq t\} \vee 1} \leq q \right\}.$$

and $+\infty$ if empty.

This procedure provides **FDR control**! ($\text{FDR}(\{j : W_j \geq T_q^*\}) \leq q$)

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

FDR control: HRT loses power due to the train–test split

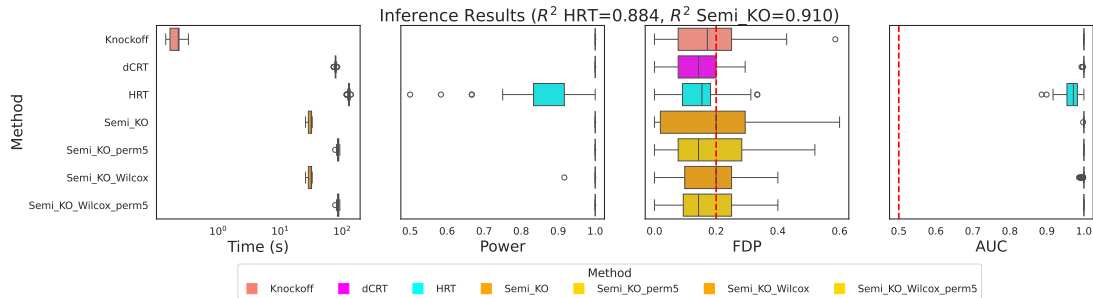


Figure 4: FDR control with adjacent support (Gradient Boosting). Let $X \sim \mathcal{N}(0, \Sigma)$ with $\Sigma_{ij} = 0.6^{|i-j|}$ and $y = \beta^\top X + \varepsilon$, where the first $0.25p$ entries of β are drawn uniformly from $[1, 2]$, the remaining entries are zero, and $\varepsilon \sim \mathcal{N}(0, 1)$. $\hat{m}$ is a gradient boosting, achieving $R^2 = 0.884$ with a train–test split and $R^2 = 0.910$ without splitting.

Knockoffs lose control when generating $\tilde{X}$ from highly correlated X

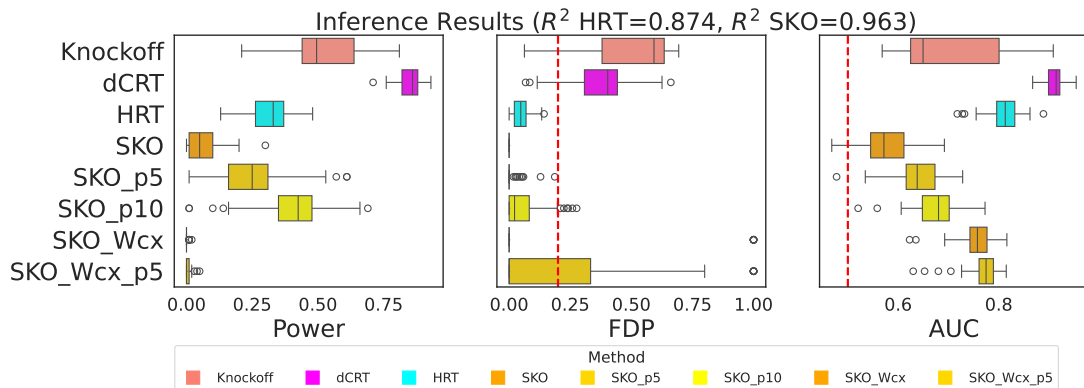


Figure 5: FDR control in a high-dimensional setting. Let $X \sim \mathcal{N}(0, \Sigma)$ with $\Sigma_{ij} = 0.6^{|i-j|}$ and $y = \beta^\top X + \varepsilon$, where $0.25p$ entries of β are selected uniformly at random and drawn from $\mathcal{N}(0, 1)$, the remaining are zero, and $\varepsilon \sim \mathcal{N}(0, 1)$. We use $n = 300$ samples and $p = 400$ features. $\hat{m}$ is a lasso, achieving $R^2 = 0.874$ with a train–test split and $R^2 = 0.963$ without splitting.