Some advances in Conditional Independence Testing with Machine Learning

Angel REYERO LOBO

PSI Treatment Effect Heterogeneity Study Group
IMT & Inria Paris-Saclay
Joint work with:
Sebastian Arias, Peter Grünwald, Michele Meziu, Pierre Neuvial & Bertrand Thirion

November 5, 2026



Contents

1. Introduction
2. Semi-knockoffs
   - Model-X assumption and Exchangeability under the Null
   - Oracle Semi-knockoffs
   - Practical Semi-knockoffs
   - Experiments
3. Sequential Conditional Independence Testing with Machine Learning
   - Sequential testing
   - Introduction to the e-values
   - Model-X e-value and e-CRT
   - Analysis under the alternative
   - Analysis under the null
   - Experiments
4. Conclusions
5. References

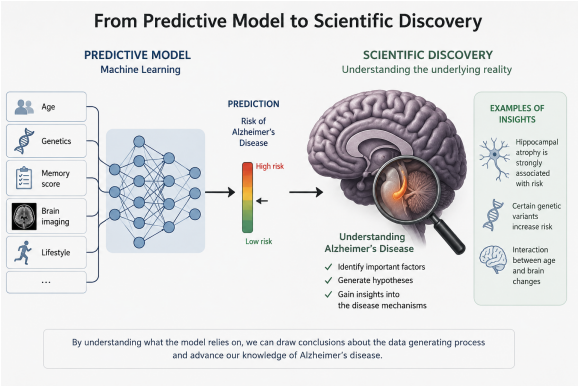
Index

1. Introduction
2. Semi-knockoffs
   - Model-X assumption and Exchangeability under the Null
   - Oracle Semi-knockoffs
   - Practical Semi-knockoffs
   - Experiments
3. Sequential Conditional Independence Testing with Machine Learning
   - Sequential testing
   - Introduction to the e-values
   - Model-X e-value and e-CRT
   - Analysis under the alternative
   - Analysis under the null
   - Experiments
4. Conclusions
5. References

Scientific Discovery with Machine Learning (ML)

Goal: identify which X^j (e.g. genes, brain regions) are *important* for Y (e.g. a disease).

- ML achieves **high predictive accuracy**.
- We want procedures that are:
  1. **Model-agnostic**: any model can be used.
  2. **Statistically valid**: type-I or FDR control.
  3. **Robust** to misspecifications.

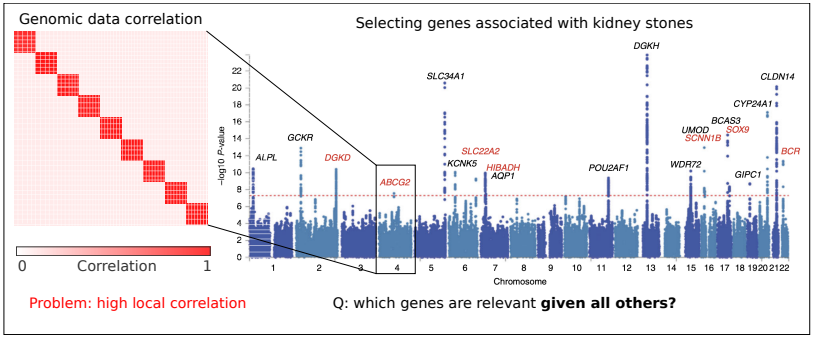


Conditional Independent Testing

$$\mathscr{P}^{\mathrm{CI}} := \left\{ P_{(X,Y)} : X^j \perp\!\!\!\perp Y \mid X^{-j} \right\} \text{ vs } \mathscr{Q}^{\mathrm{CI}} := \left\{ P_{(X,Y)} : X^j \not\perp\!\!\!\perp Y \mid X^{-j} \right\}.$$

⚠ **Highly correlated** genomic data

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

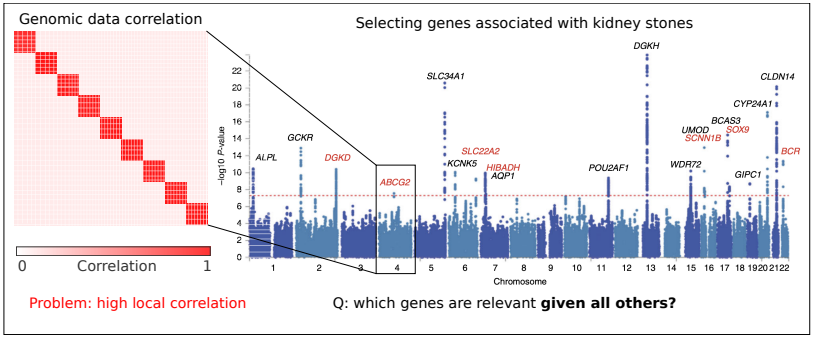


Conditional Independent Testing

$$\mathscr{P}^{\mathrm{CI}} := \left\{ P_{(X,Y)} : X^j \perp\!\!\!\perp Y \mid X^{-j} \right\} \text{ vs } \mathscr{Q}^{\mathrm{CI}} := \left\{ P_{(X,Y)} : X^j \not\perp\!\!\!\perp Y \mid X^{-j} \right\}.$$

Highly correlated genomic data

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



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

Index

1. Introduction
2. Semi-knockoffs
   - Model-X assumption and Exchangeability under the Null
   - Oracle Semi-knockoffs
   - Practical Semi-knockoffs
   - Experiments
3. Sequential Conditional Independence Testing with Machine Learning
   - Sequential testing
   - Introduction to the e-values
   - Model-X e-value and e-CRT
   - Analysis under the alternative
   - Analysis under the null
   - Experiments
4. Conclusions
5. References

Related work: the Model-X framework

Classical guarantees rely on *assumptions on* $Y \mid X$, e.g. a GLM: $g\big(\mathbb{E}[Y \mid X]\big) = \sum_j \beta^j X^j$.

Related work: the Model-X framework

Classical guarantees rely on *assumptions on* $Y \mid X$, e.g. a GLM: $g(\mathbb{E}[Y \mid X]) = \sum_j \beta^j X^j$.

Model-X (Candès et al., 2018)

Instead, assume *knowledge of the input distribution*:

- $P_{X^j|X^{-j}} = K(\cdot \mid X^{-j})$ is **known**, so we can **sample** from it.

Motivation: if the relationship *among inputs* is easier than the one between inputs and outcome, e.g. in genomics:

- abundant **unlabeled data** on the genes,
- inputs coming from the **same measurement device**.

Related work: the Model-X framework

Classical guarantees rely on *assumptions on* $Y \mid X$, e.g. a GLM: $g(\mathbb{E}[Y \mid X]) = \sum_j \beta^j X^j$.

Model-X (Candès et al., 2018)

Instead, assume *knowledge of the input distribution*:

- $P_{X^j|X^{-j}} = K(\cdot \mid X^{-j})$ is **known**, so we can **sample** from it.

Motivation: if the relationship *among inputs* is easier than the one between inputs and outcome, e.g. in genomics:

- abundant **unlabeled data** on the genes,
- inputs coming from the **same measurement device**.

Model-X conditional null hypothesis

$$\mathscr{P}_K^{\text{X-CI}} := \left\{ P_{(X,Y)} \,:\, X^j \perp\!\!\!\perp Y \mid X^{-j},\, P_{X^j|X^{-j}} = K(\cdot \mid X^{-j}) \right\}$$

Related work: Model-X and Exchangeability under the Null

Model-X assumption: we can sample a r.v. $\widetilde{X}^{(j)} \sim K(X^j \mid X^{-j})$.
Under the null, for *any* importance statistic $T : \mathscr{X} \times \mathscr{Y} \to \mathbb{R}$, the statistics

$$T(X^j, X^{-j}, Y) \quad \text{and} \quad T(\widetilde{X}^{(j)}, X^{-j}, Y)$$

are **exchangeable**:

$$P_{(X^j, X^{-j}, Y)} = P_{X^j|X^{-j},Y} \times P_{X^{-j},Y} = P_{X^j|X^{-j}} \times P_{X^{-j},Y} = K_{X^j|X^{-j}} \times P_{X^j,Y} = P_{(\widetilde{X}^{(j)}, X^{-j}, Y)}.$$

Related work: Model-X and Exchangeability under the Null

Model-X assumption: we can sample a r.v. $\widetilde{X}^{(j)} \sim K(X^j \mid X^{-j})$.
Under the null, for *any* importance statistic $T : \mathscr{X} \times \mathscr{Y} \to \mathbb{R}$, the statistics

$$T(X^j, X^{-j}, Y) \quad \text{and} \quad T(\widetilde{X}^{(j)}, X^{-j}, Y)$$

are **exchangeable**:

$$P_{(X^j, X^{-j}, Y)} = P_{X^j \mid X^{-j}, Y} \times P_{X^{-j}, Y} = P_{X^j \mid X^{-j}} \times P_{X^{-j}, Y} = K_{X^j \mid X^{-j}} \times P_{X^j, Y} = P_{(\widetilde{X}^{(j)}, X^{-j}, Y)}.$$

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

- **Coefficient**-based importance: $T(X^j, X^{-j}, Y) = |\widehat{\beta}^j|$
- **Performance**-based importance: $T(X^j, X^{-j}, Y) = \ell\big(\widehat{m}(X^j, X^{-j}), Y\big)$
  Compare

$$\left\{\ell\Big(\widehat{m}(X^j, X^{-j}), Y\Big)\right\}_{i=1}^{n_{\text{test}}} \qquad \text{with} \qquad \left\{\ell\Big(\widehat{m}(\widetilde{X}^{(j)}, X^{-j}), Y\Big)\right\}_{i=1}^{n_{\text{test}}}, \quad \widetilde{X}^{(j)} \sim K(X^j \mid X^{-j}).$$

Related work: Model-X and Exchangeability under the Null

Model-X assumption: we can sample a r.v. $\widetilde{X}^{(j)} \sim K(X^j \mid X^{-j})$.
Under the null, for *any* importance statistic $T : \mathscr{X} \times \mathscr{Y} \to \mathbb{R}$, the statistics

$$T(X^j, X^{-j}, Y) \quad \text{and} \quad T(\widetilde{X}^{(j)}, X^{-j}, Y)$$

are **exchangeable**:

$$P_{(X^j, X^{-j}, Y)} = P_{X^j \mid X^{-j}, Y} \times P_{X^{-j}, Y} = P_{X^j \mid X^{-j}} \times P_{X^{-j}, Y} = K_{X^j \mid X^{-j}} \times P_{X^j, Y} = P_{(\widetilde{X}^{(j)}, X^{-j}, Y)}.$$

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

- **Coefficient**-based importance: $T(X^j, X^{-j}, Y) = |\widehat{\beta}^j|$
- **Performance**-based importance: $T(X^j, X^{-j}, Y) = \ell\big(\widehat{m}(X^j, X^{-j}),\, Y\big)$
  Compare

$$\left\{\ell\Big(\widehat{m}(X^j, X^{-j}),\, Y\Big)\right\}_{i=1}^{n_{\text{test}}} \qquad \text{with} \qquad \left\{\ell\Big(\widehat{m}(\widetilde{X}^{(j)}, X^{-j}),\, Y\Big)\right\}_{i=1}^{n_{\text{test}}}, \quad \widetilde{X}^{(j)} \sim K(X^j \mid X^{-j}).$$

  A train–test split is required to preserve exchangeability.

Exchangeability under the Null

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

$$\left\{\ell\left(\widehat{m}(\widetilde{X}^{(j)}, X^{-j}),\, Y\right)\right\}_{i=1}^{n_{\text{test}}} \qquad \text{with} \qquad \left\{\ell\left(\widehat{m}(X^{j}, X^{-j}),\, Y\right)\right\}_{i=1}^{n_{\text{test}}}, \quad \widetilde{X}^{(j)} \sim K(X^{j} \mid X^{-j}).$$

☹ 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

$$\left\{\ell\left(\widehat{m}(\widetilde{X}^{(j)}, X^{-j}),\, Y\right)\right\}_{i=1}^{n_{\text{test}}} \quad \text{with} \quad \left\{\ell\left(\widehat{m}(X^j, X^{-j}),\, Y\right)\right\}_{i=1}^{n_{\text{test}}}, \quad \widetilde{X}^{(j)} \sim K(X^j \mid X^{-j}).$$

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

Semi-knockoffs: We instead propose comparing

$$\left\{\ell\left(\widehat{m}(\widetilde{X}_1^{(j)}, X^{-j}),\, Y\right)\right\}_1^n \quad \text{vs} \quad \left\{\ell\left(\widehat{m}(\widetilde{X}_2^{(j)}, X^{-j}),\, Y\right)\right\}_1^n \text{ with}$$
$$\widetilde{X}_1^{(j)} \sim K(X^j \mid X^{-j}), \qquad \widetilde{X}_2^{(j)} \sim K(X^j \mid X^{-j}, Y).$$

Key insight: If $X^j \perp\!\!\!\perp Y \mid X^{-j}$, $K(X^j \mid X^{-j}) = K(X^j \mid X^{-j}, Y)$, 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 $\widehat{m}$.

No Need to Sample from $K(X^j \mid X^{-j}, Y)$

Problem: Directly sampling $\widetilde{X}_2^{(j)} \sim K(X^j \mid X^{-j}, Y)$ is complex.

No Need to Sample from $K(X^j \mid X^{-j}, Y)$

Problem: Directly sampling $\widetilde{X}_2^{(j)} \sim K(X^j \mid X^{-j}, Y)$ is complex.

Exchangeability under $\mathscr{P}^{\mathrm{CI}}$:

$$\rho(X^{-j}, Y) := \mathbb{E}[X^j \mid X^{-j}, Y] = \mathbb{E}[X^j \mid X^{-j}] =: \nu(X^{-j})$$

No Need to Sample from $K(X^j \mid X^{-j}, Y)$

Problem: Directly sampling $\widetilde{X}_2^{(j)} \sim K(X^j \mid X^{-j}, Y)$ is complex.

Exchangeability under $\mathscr{P}^{\mathrm{CI}}$:

$$\rho(X^{-j}, Y) := \mathbb{E}[X^j \mid X^{-j}, Y] = \mathbb{E}[X^j \mid X^{-j}] =: \nu(X^{-j})$$

Idea: Construct new samples using *residuals*.

$$\widetilde{X}_1^{(j)} = \nu(X^{-j}) + \underbrace{x_l^j - \nu(x_l^{-j})}_{\varepsilon_{j,1}^l}, \quad \widetilde{X}_2^{(j)} = \rho(X^{-j}, Y) + \underbrace{x_k^j - \rho(x_k^{-j}, y_k)}_{\varepsilon_{j,2}^k}$$

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

Algorithm (Type-I error control)

Algorithm 1 Semi-knockoffs Wilcoxon (`SKO-Wcx`)

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

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

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

Compute residuals $\widehat{\varepsilon}_{j,1} = X^j - \widehat{\nu}_j(X^{-j})$

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

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

Construct

$$\widetilde{X}_{1,i}^{\prime(j)} = \widehat{\nu}_j(X_i^{-j}) + \widehat{\varepsilon}_{j,1,\pi_{j,1}(i)}$$
$$\widetilde{X}_{2,i}^{\prime(j)} = \widehat{\rho}_j(X_i^{-j}, Y_i) + \widehat{\varepsilon}_{j,2,\pi_{j,2}(i)}$$

Return Wilcoxon test $\left\{\ell\big(\widehat{m}(\widetilde{X}_{1,i}^{\prime(j)}, X_i^{-j}), Y_i\big)\right\}_i$ and $\left\{\ell\big(\widehat{m}(\widetilde{X}_{2,i}^{\prime(j)}, X_i^{-j}), Y_i\big)\right\}_i$

Algorithm (Derandomizing the permutations)

Algorithm 2 Derandomized Semi-knockoffs Wilcoxon (`SKO-Wcx-p`(B))

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

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

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

Compute residuals $\widehat{\varepsilon}_{j,1} = X^j - \widehat{\nu}_j(X^{-j})$

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

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

Construct

$$\widetilde{X}_{1,i,b}^{\prime(j)} = \widehat{\nu}_j(X_i^{-j}) + \widehat{\varepsilon}_{j,1,\pi_{j,1}^b(i)}$$

$$\widetilde{X}_{2,i,b}^{\prime(j)} = \widehat{\rho}_j(X_i^{-j}, Y_i) + \widehat{\varepsilon}_{j,2,\pi_{j,2}^b(i)}$$

Return Wilcoxon test $\left\{ \frac{1}{B}\sum_b \ell\big(\widehat{m}(\widetilde{X}_{1,i,b}^{\prime(j)}, X_i^{-j}), Y_i\big) \right\}_i$ and $\left\{ \frac{1}{B}\sum_b \ell\big(\widehat{m}(\widetilde{X}_{2,i,b}^{\prime(j)}, X_i^{-j}), Y_i\big) \right\}_i$

Nice Properties of Semi-Knockoffs

FDR control

Knockoff threshold (Candès et al., 2018) applies directly and yields FDR control.

- ✓ No p-value computation or correction
- ✓ No knockoff construction, hard with correlated features (Blain et al., 2025)

Nice Properties of Semi-Knockoffs

FDR control

Knockoff threshold (Candès et al., 2018) applies directly and yields FDR control.

- ✓ No p-value computation or correction
- ✓ No knockoff construction, hard with correlated features (Blain et al., 2025)

Robustness to Model-X misspecification

- Model-X methods require a *consistent* sampler (Berrett et al., 2019),

$$d(\hat{K}_{X^j|X^{-j}}, K_{X^j|X^{-j}}) \to 0,$$

  yet in practice rely on a regression sampler or a parametric model.
- Semi-Knockoffs relax this: it suffices that the samplers coincide *under the null*:

$$d\big(\hat{K}_{X^j|X^{-j},Y}, \hat{K}_{X^j|X^{-j}}\big) \to 0.$$

Simulated data

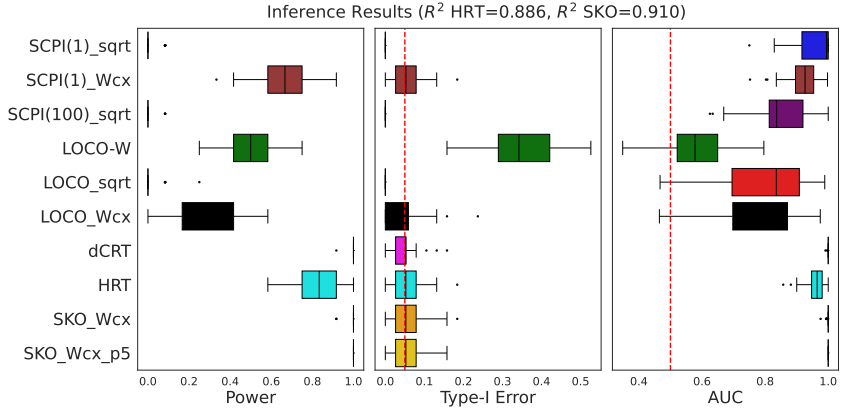


Figure 1: **SKO controls type-I error with high power by avoiding a train–test split.** Simulated linear model with correlated Gaussian covariates ($p = 50$, $n = 300$) and a gradient boosting.

Real data

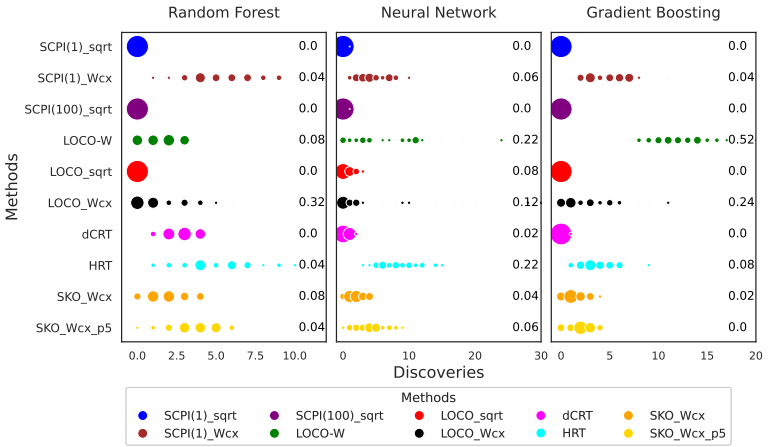


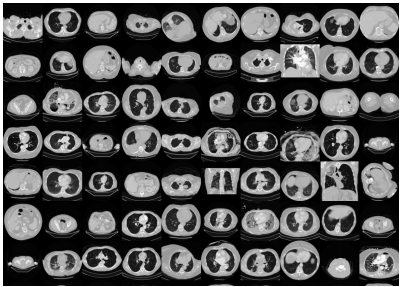
Figure 2: **Discoveries on real data across ML models.** The type-I error (shown to the right of each plot) is estimated using an artificial null feature correlated at 0.6 with the original features.

Index

1. Introduction
2. Semi-knockoffs
   - Model-X assumption and Exchangeability under the Null
   - Oracle Semi-knockoffs
   - Practical Semi-knockoffs
   - Experiments
3. Sequential Conditional Independence Testing with Machine Learning
   - Sequential testing
   - Introduction to the e-values
   - Model-X e-value and e-CRT
   - Analysis under the alternative
   - Analysis under the null
   - Experiments
4. Conclusions
5. References

Sequential testing and Safe Anytime-Valid Inference (SAVI)

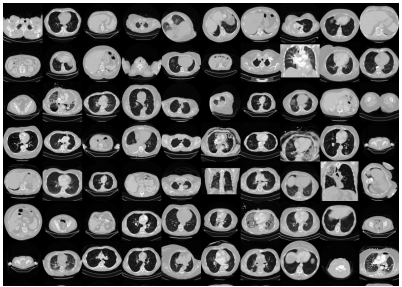
- We set $\alpha = 0.05$ and then collect some data.



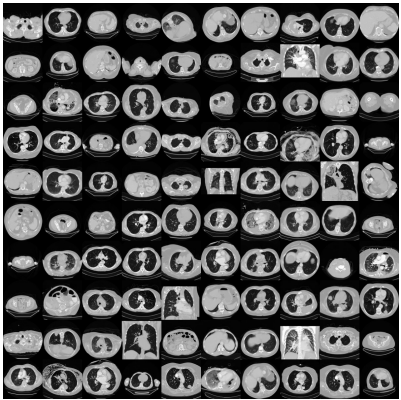
- The observed p-value is 0.07.
- Why don't we collect more data?

Sequential testing and Safe Anytime-Valid Inference (SAVI)

- We set $\alpha = 0.05$ and then collect some data.



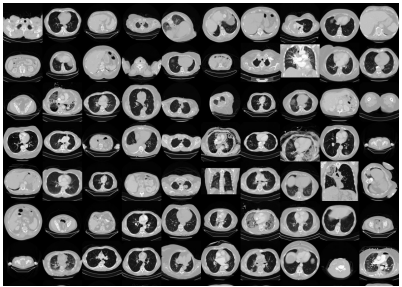
- The observed p-value is 0.07.
- Why don't we collect more data?



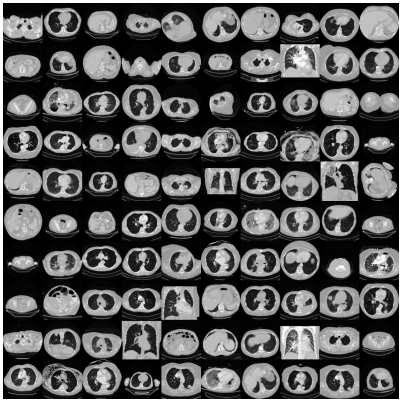
- The new p-value is 0.04.

Sequential testing and Safe Anytime-Valid Inference (SAVI)

- We set $\alpha = 0.05$ and then collect some data.



- The observed p-value is 0.07.
- Why don't we collect more data?



- The new p-value is 0.04.

✖ With probability tending to 1, we will reject even under $\mathscr{P}$.

A Quick Introduction to the e-World

E-value. A random variable S satisfying $S \geq 0, \quad \mathbb{E}_P[S] \leq 1 \quad \forall P \in \mathscr{P}$.
We can obtain type-I error control: by Markov's inequality,

$$\mathbb{P}_P(S \geq 1/\alpha) \leq \frac{\mathbb{E}_P[S]}{1/\alpha} \leq \alpha,$$

so we reject $\mathscr{P}$ whenever $S \geq 1/\alpha$.

A Quick Introduction to the e-World

E-value. A random variable S satisfying $S \geq 0, \quad \mathbb{E}_P[S] \leq 1 \quad \forall P \in \mathscr{P}$.
We can obtain type-I error control: by Markov's inequality,

$$\mathbb{P}_P(S \geq 1/\alpha) \leq \frac{\mathbb{E}_P[S]}{1/\alpha} \leq \alpha,$$

so we reject $\mathscr{P}$ whenever $S \geq 1/\alpha$.

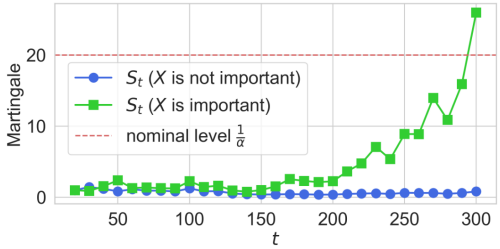
E-process. A sequence $(S_t)_{t\geq 1}$ such that

$$S_t \geq 0, \qquad \mathbb{E}_P[S_\tau] \leq 1$$

for every stopping time τ and every $P \in \mathscr{P}$.
By Ville's inequality,

$$\mathbb{P}_P\left(\sup_{t\geq 1} S_t \geq \frac{1}{\alpha}\right) \leq \alpha.$$

Reject as soon as $S_t \geq 1/\alpha$.



Related work: Sequential Conditional Independence Testing

Goal: Test $\mathscr{P}_K^{\text{X-CI}} := \left\{ P_{(X,Y)} \,:\, X^j \perp\!\!\!\perp Y \mid X^{-j},\, P_{X^j|X^{-j}} = K(\cdot \mid X^{-j}) \right\}$ vs a point alternative $\mathbb{Q}$.

Existing approaches:

- **Model-X e-value** (Grünwald et al. (2024)): based on likelihood ratio

$$\frac{f_{Y|X^j,X^{-j}}}{f_{Y|X^{-j}}}.$$

  - It is growth-rate optimal (GRO) against $\mathbb{Q}$ with density f, but how do we estimate it?

Related work: Sequential Conditional Independence Testing

Goal: Test $\mathscr{P}_K^{\text{X-CI}} := \left\{ P_{(X,Y)} \,:\, X^j \perp\!\!\!\perp Y \mid X^{-j},\, P_{X^j|X^{-j}} = K(\cdot \mid X^{-j}) \right\}$ vs a point alternative $\mathbb{Q}$.

Existing approaches:

- **Model-X e-value** (Grünwald et al. (2024)): based on likelihood ratio

$$\frac{f_{Y|X^j,X^{-j}}}{f_{Y|X^{-j}}}.$$

  - It is growth-rate optimal (GRO) against $\mathbb{Q}$ with density f, but how do we estimate it?
- **e-CRT** (Shaer et al. (2023)): based on importance statistics

$$q = T(X^j, X^{-j}, Y), \quad \widetilde{q} = T(\widetilde{X}^{(j)}, X^{-j}, Y),$$

  and an antisymmetric score $g(\widetilde{q}, q) = -g(q, \widetilde{q})$ giving an e-value $1 + g(q, \widetilde{q})$.
  - It is flexible but which T and g do we use?

Related work: Sequential Conditional Independence Testing

Goal: Test $\mathscr{P}_K^{\text{X-CI}} := \left\{ P_{(X,Y)} \,:\, X^j \perp\!\!\!\perp Y \mid X^{-j},\, P_{X^j|X^{-j}} = K(\cdot \mid X^{-j}) \right\}$ vs a point alternative $\mathbb{Q}$.

Existing approaches:

- **Model-X e-value** (Grünwald et al. (2024)): based on likelihood ratio

$$\frac{f_{Y|X^j,X^{-j}}}{f_{Y|X^{-j}}}.$$

  - It is growth-rate optimal (GRO) against $\mathbb{Q}$ with density f, but how do we estimate it?
- **e-CRT** (Shaer et al. (2023)): based on importance statistics

$$q = T(X^j, X^{-j}, Y), \quad \widetilde{q} = T(\widetilde{X}^{(j)}, X^{-j}, Y),$$

  and an antisymmetric score $g(\widetilde{q}, q) = -g(q, \widetilde{q})$ giving an e-value $1 + g(q, \widetilde{q})$.
  - It is flexible but which T and g do we use?
- Which one is better in practice? How do we incorporate ML?

Model-X e-variable in practice: Exponential e-variable

The **GRO e-variable** (Grünwald et al. (2024)) for testing $\mathscr{P}_K^{\mathrm{X-CI}}$ vs $\{\mathbb{Q}\}$ is:

$$\frac{f_{Y|X^j,X^{-j}}(Y|X^j,X^{-j})}{f_{Y|X^{-j}}(Y|X^{-j})} = \frac{f_{Y|X^j,X^{-j}}(Y|X^j,X^{-j})}{\mathbb{E}_{\widetilde{X}^{(j)}|X^{-j}}[f_{Y|X^j,X^{-j}}(Y|\widetilde{X}^{(j)},X^{-j})]},$$

where $\mathbb{Q}$ has density f and $\widetilde{X}^{(j)} \sim K_{X|Z}$.
We propose:

$$E^{\eta}_{\mathrm{exp}} := \frac{e^{-\eta \ell(\widehat{m}(X^j,X^{-j}),Y)}}{\mathbb{E}_{\widetilde{X}^{(j)}|X^{-j}}[e^{-\eta \ell(\widehat{m}(\widetilde{X}^{(j)},X^{-j}),Y)}]},$$

with a free parameter $\eta \geq 0$.

- Good **approximation of the GRO** if: $\ell(m(x^j,x^{-j}),y) \approx -\eta \log(f_{Y|X^j,X^{-j}}(y|x^j,x^{-j}))$.
- E.g. in regression with ℓ^2-loss and Gaussian noise, and classification with CE.

Testing Algorithm

Algorithm 3 Testing Algorithm

While $\prod_{i=1}^{n} E_i \leq \frac{1}{\alpha}$ **do**

1. Sample $(X_{n+1}^j, X_{n+1}^{-j}, Y_{n+1})$.
2. Update the model to $\widehat{m}_n$ using all past data $\{(X_k^j, X_k^{-j}, Y_k)\}_{k=1}^n$ or an online step.
3. Perform a prequential parameter update:

$$\widehat{\eta}_n = \arg\max_{\eta} \left(\prod_{i=1}^{n} E_{\exp}^{\eta}(X_i^j, X_i^{-j}, Y_i) \right)$$

4. Compute

$$E_{n+1} = E_{\exp}^{\widehat{\eta}_n}(X_{n+1}^j, X_{n+1}^{-j}, Y_{n+1}).$$

e-CRT / Antisymmetric e-variables

To leverage ML models, (Shaer et al. 2023) introduce **importance statistics**:

$$I := \ell(\hat{m}(X^j, X^{-j}), Y) \qquad \text{and} \qquad \widetilde{I} := \ell(\hat{m}(\widetilde{X}^{(j)}, X^{-j}), Y),$$

and e-variables of the form:

$$1 + g(I, \widetilde{I}),$$

with **antisymmetric** functions g ($g(a,b) = -g(b,a)$) such that $g(a,b) \geq -1$.

- Simplifies conditional independence question into **exchangeability**

$$Y \perp\!\!\!\perp X^j | X^{-j} \longrightarrow (I, \widetilde{I}) \overset{d}{=} (\widetilde{I}, I).$$

- Examples:

$$\underbrace{1 + \lambda \operatorname{sign}(\widetilde{I} - I)}_{\text{Sign}}, \qquad \underbrace{1 + \lambda \tanh(\widetilde{I} - I)}_{\text{Tanh}}, \qquad \underbrace{1 + \lambda(\widetilde{I} - I)}_{\textbf{Coin-betting} \text{ (bounded loss)}}.$$

Practical limitation: Function g is left unspecified, unclear how to choose it.

Analysis under the alternative

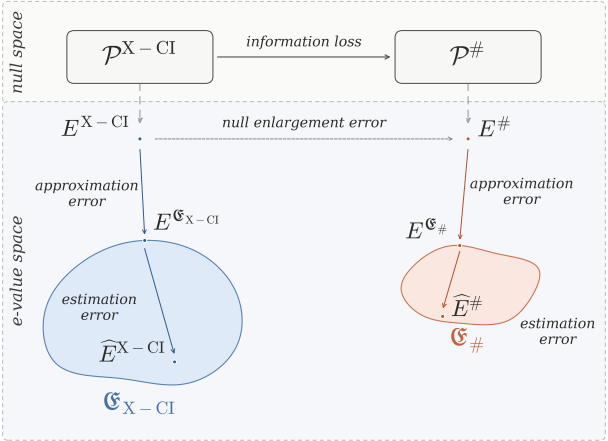


Figure 3: While distilling information induces a *null enlargement error*, it can improve power by reducing the *approximation* and *estimation errors*.

Analysis under the null

The Model-X kernel K is usually estimated $\widehat{K}_n$. Then, we do not have *exact* e-values.
What guarantees do we still have?

Analysis under the null

The Model-X kernel K is usually estimated $\widehat{K}_n$. Then, we do not have *exact* e-values.
What guarantees do we still have?

Triple robustness (informal theorem)

Consider the coin-betting e-process $E^{(n)} = \prod_{i=1}^{n} E_i$, with

$$E_i := 1 + \widehat{\lambda}_i \Big(\ell\big(\widehat{m}_i(\widetilde{X}_i^{(j)}, X_i^{-j}), Y_i\big) - \ell\big(\widehat{m}_i(X_i, X_i^{-j}), Y_i\big) \Big), \qquad \widetilde{X}_i^{(j)} \sim \widehat{K}_i.$$

Then

$$\underbrace{\mathbb{P}\Big(\sup_{1 \le n \le N} E^{(n)} \ge \tfrac{1}{\alpha} \Big) - \alpha}_{\text{Type-I error inflation}} \le \underbrace{O\Big(\alpha \sum_{i=1}^{N} \widehat{\lambda}_i \, a_i \, b_i \Big)}_{\text{triple robustness}},$$

where $a_i := d(\widehat{K}_i, K)$, $b_i := d(\widehat{m}_i, m)$, and m is the Bayes predictor.

Analysis under the null

The Model-X kernel K is usually estimated $\widehat{K}_n$. Then, we do not have *exact* e-values.
What guarantees do we still have?

Triple robustness (informal theorem)

Consider the coin-betting e-process $E^{(n)} = \prod_{i=1}^{n} E_i$, with

$$E_i := 1 + \widehat{\lambda}_i \Big(\ell\big(\widehat{m}_i(\widetilde{X}_i^{(j)}, X_i^{-j}), Y_i\big) - \ell\big(\widehat{m}_i(X_i, X_i^{-j}), Y_i\big) \Big), \qquad \widetilde{X}_i^{(j)} \sim \widehat{K}_i.$$

Then

$$\underbrace{\mathbb{P}\Big(\sup_{1 \le n \le N} E^{(n)} \ge \tfrac{1}{\alpha} \Big) - \alpha}_{\text{Type-I error inflation}} \le \underbrace{O\Big(\alpha \sum_{i=1}^{N} \widehat{\lambda}_i \, a_i \, b_i \Big)}_{\text{triple robustness}},$$

where $a_i := d(\widehat{K}_i, K)$, $b_i := d(\widehat{m}_i, m)$, and m is the Bayes predictor.

This parallels the double robustness of the CFI (Reyero-Lobo et al., 2025).

Experiments (HIV dataset)

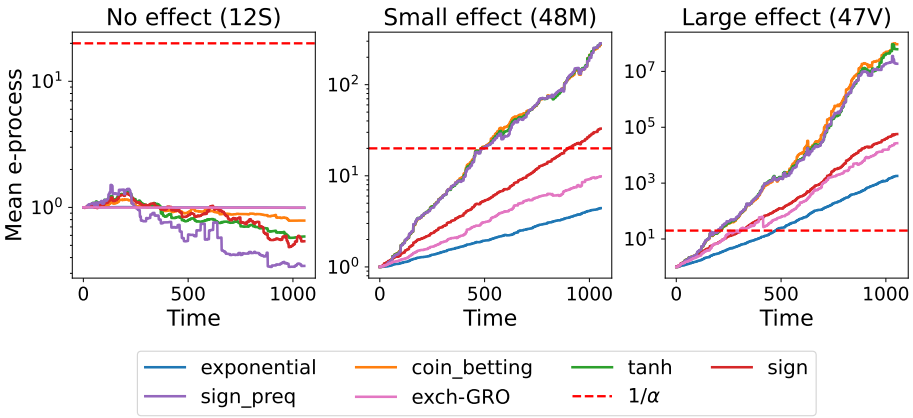


Figure 4: **Drug resistance for three HIV mutations**. On the left (null mutation) none of the e-processes reject. In the center (small signal) most antisymmetric reject the null and outperform the exponential, which does not. On the right (large signal) all the e-processes reject the null.

Index

1. Introduction
2. Semi-knockoffs
   - Model-X assumption and Exchangeability under the Null
   - Oracle Semi-knockoffs
   - Practical Semi-knockoffs
   - Experiments
3. Sequential Conditional Independence Testing with Machine Learning
   - Sequential testing
   - Introduction to the e-values
   - Model-X e-value and e-CRT
   - Analysis under the alternative
   - Analysis under the null
   - Experiments
4. Conclusions
5. References

Conclusion

What we need

- **Interpretable** ML, to gain insights about the underlying distribution.
- **Model-agnostic** methods, to leverage predictive black-box models.
- **Statistically valid** inference.

Conclusion

What we need

- **Interpretable** ML, to gain insights about the underlying distribution.
- **Model-agnostic** methods, to leverage predictive black-box models.
- **Statistically valid** inference.

Power in practice

- **Semi-Knockoffs** are data-efficient: no train-test split is needed.
- **Sequential CIT:** We analyse the e-power loss trade-offs in practice.

Conclusion

What we need

- **Interpretable** ML, to gain insights about the underlying distribution.
- **Model-agnostic** methods, to leverage predictive black-box models.
- **Statistically valid** inference.

Power in practice

- **Semi-Knockoffs** are data-efficient: no train-test split is needed.
- **Sequential CIT:** We analyse the e-power loss trade-offs in practice.

Robustness to Model-X misspecification

- **Semi-Knockoffs:** validity relies only on the samplers we choose.
- **Sequential CIT:** double/triple robustness guarantee comes from the ML model itself.

Index

1. Introduction
2. Semi-knockoffs
   - Model-X assumption and Exchangeability under the Null
   - Oracle Semi-knockoffs
   - Practical Semi-knockoffs
   - Experiments
3. Sequential Conditional Independence Testing with Machine Learning
   - Sequential testing
   - Introduction to the e-values
   - Model-X e-value and e-CRT
   - Analysis under the alternative
   - Analysis under the null
   - Experiments
4. Conclusions
5. References

References

Thomas B. Berrett, Yi Wang, Rina Foygel Barber, and Richard J. Samworth. The conditional permutation test for independence while controlling for confounders. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 82(1): 175–197, 10 2019. ISSN 1369-7412. doi: 10.1111/rssb.12340. URL https://doi.org/10.1111/rssb.12340.

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.

References

Peter Grünwald, Alexander Henzi, and Tyron Lardy. Anytime-valid tests of conditional independence under model-x. *Journal of the American Statistical Association*, 119 (546):1554–1565, 2024. doi: 10.1080/01621459.2023.2205607. URL https://doi.org/10.1080/01621459.2023.2205607.

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, Pierre Neuvial, and Bertrand Thirion. Conditional feature importance revisited: Double robustness, efficiency and inference, 2025. URL https://arxiv.org/abs/2501.17520.

Angel Reyero-Lobo, Michele Meziu, Sebastian Uriel Arias, and Peter Grünwald. Sequential conditional independence testing with machine learning models, 2026a. URL https://arxiv.org/abs/2610.11388.

References

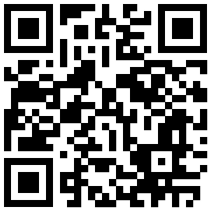
Angel Reyero-Lobo, Bertrand Thirion, and Pierre Neuvial. Semi-knockoffs: a model-agnostic conditional independence testing method with finite-sample guarantees. In *Proceedings of the 43rd International Conference on Machine Learning (ICML)*, 2026b.

Shalev Shaer, Gal Maman, and Yaniv Romano. Model-x sequential testing for conditional independence via testing by betting. In Francisco Ruiz, Jennifer Dy, and Jan-Willem van de Meent, editors, *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pages 2054–2086. PMLR, 25–27 Apr 2023. URL `https://proceedings.mlr.press/v206/shaer23a.html`.

References

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.
- Angel Reyero-Lobo, Michele Meziu, Sebastian Uriel Arias, and Peter Grünwald. Sequential conditional independence testing with machine learning models, 2026.

Index

6 Semi-Knockoffs
- Introduction to the knockoffs
- Semi-knockoffs FDR control

7 Sequential Conditional Independence Testing with ML
- E-power
- Exponential vs antisymmetric e-power
- Antisymmetric GRO e-variable
- Triple Robustness

Knockoffs framework

Knockoffs (from Candès et al. 2018):

1. **Knockoff variables ($\widetilde{X}$):** *imitations* of X independent of y:
   - $[X, \widetilde{X}]_{\text{swap}(s)} \overset{\text{d}}{=} [X, \widetilde{X}]$ for any $s \subset [p]$.

$$\left[X_1, X_2, X_3, \widetilde{X}_1, \widetilde{X}_2, \widetilde{X}_3\right]_{\text{swap}(1,3)} := \left[\widetilde{X}_1, X_2, \widetilde{X}_3, X_1, \widetilde{X}_2, X_3\right]$$

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

Example 1 (Gaussian knockoffs)

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

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

Knockoffs framework

Knockoffs (from Candès et al. 2018):

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

Example 1 (Lasso Coefficients Difference)

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

Knockoffs framework

Knockoffs (from Candès et al. 2018):

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

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

   and $+\infty$ if empty.

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

$$\mathrm{FDP}(\widehat{S}) := \frac{\left|\widehat{S} \cap \mathscr{H}_0\right|}{\left|\widehat{S}\right| \vee 1} \qquad \mathrm{FDR}(\widehat{S}) := \mathbb{E}\left[\mathrm{FDP}(\widehat{S})\right].$$

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

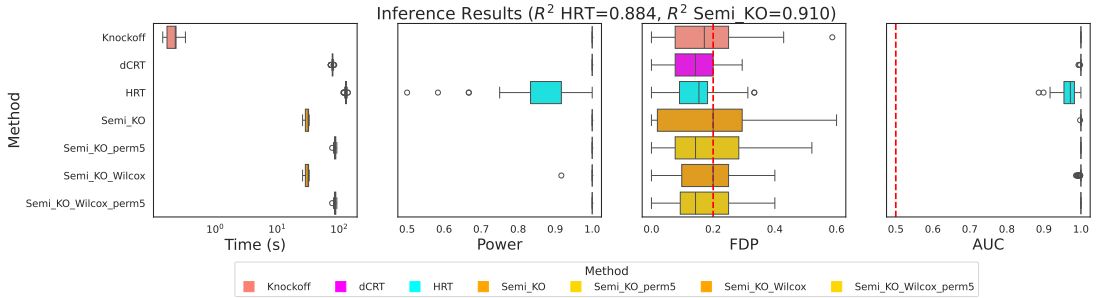


Figure 5: **FDR control with adjacent support (Gradient Boosting).** Let $X \sim \mathcal{N}(0, \Sigma)$ with $\Sigma_{ij} = 0.6^{|i-j|}$ and $y = \beta^\top X + \varepsilon$, where the first $0.25p$ entries of β are drawn uniformly from $[1,2]$, the remaining entries are zero, and $\varepsilon \sim \mathcal{N}(0,1)$. $\hat{m}$ 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 $\widetilde{X}$ from highly correlated X

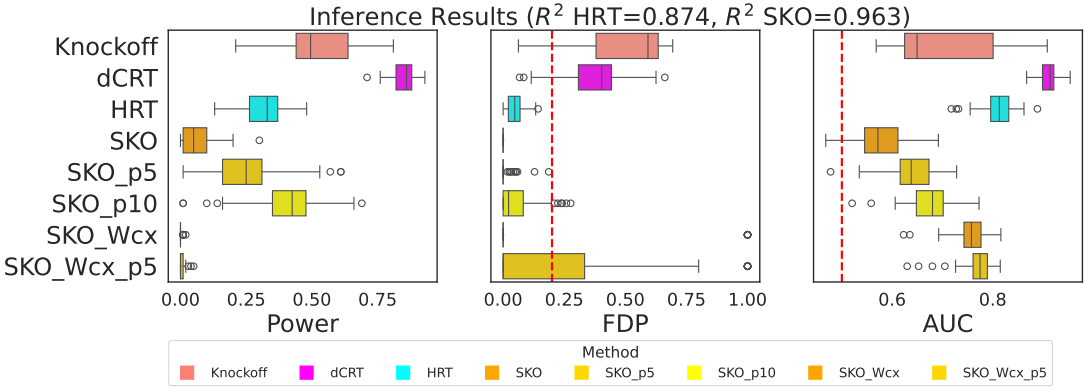


Figure 6: **FDR control in a high-dimensional setting.** Let $X \sim \mathcal{N}(0, \Sigma)$ with $\Sigma_{ij} = 0.6^{|i-j|}$ and $y = \beta^\top X + \varepsilon$, where $0.25p$ entries of β are selected uniformly at random and drawn from $\mathcal{N}(0,1)$, the remaining are zero, and $\varepsilon \sim \mathcal{N}(0,1)$. We use $n = 300$ samples and $p = 400$ features. $\widehat{m}$ is a lasso, achieving $R^2 = 0.874$ with a train–test split and $R^2 = 0.963$ without splitting.

Index

6 Semi-Knockoffs
- Introduction to the knockoffs
- Semi-knockoffs FDR control

7 Sequential Conditional Independence Testing with ML
- E-power
- Exponential vs antisymmetric e-power
- Antisymmetric GRO e-variable
- Triple Robustness

E-power

What about the power?

$$\left(\prod_{i=1} S_i\right)^{1/n} = \exp\left(\frac{1}{n}\sum_{i=1} \log S_i\right) \overset{n\to\infty}{\to} \exp\left\{\mathbb{E}^{\mathbb{Q}}[\log S^{(1)}]\right\} \tag{1}$$

One notion of power against a point alternative $\mathbb{Q}$ is the **log-utility**:

$$\mathbb{E}^{\mathbb{Q}}[\log S].$$

Exponential vs antisymmetric e-power

- If $\ell(m(x,z),y) = -\eta \log(f_{Y|X,Z}(y|x,z))$, $E^{\eta}_{\exp}$ is GRO (dominates all antisymmetric).
- Do antisymmetric e-variables ever dominate?

Thm: Exponential vs Coin-betting e-power

Define *predictiveness signal*:

$$\Delta := \mathbb{E}\left[\ell(m(\widetilde{X},Z),Y)\right] - \mathbb{E}[\ell(m(X,Z),Y)] \geq 0,$$

and loss variances: $\sigma_0^2 := \mathbb{V}(\ell(m(\widetilde{X},Z),Y)), \sigma_1^2 := \mathbb{V}(\ell(m(X,Z),Y))$. Optimal e-powers for $E^{\eta}_{\exp}$ and $E^{\lambda}_{\mathrm{cb}}$ are:

$$\mathbb{E}\left[\log E^{\eta^\star}_{\exp}\right] = \frac{\Delta^2}{2\sigma_0^2} + O(\Delta^3),\ \mathbb{E}\left[\log E^{\lambda^\star}_{\mathrm{cb}}\right] = \frac{1}{2}\frac{\Delta^2}{\sigma_1^2+\Delta^2} + O(\Delta^3).$$

$\rightarrow$ If $\frac{\sigma_0^2}{\sigma_1^2+\Delta^2} > 1$, **Coin-betting can be more e-powerful!**

e-CRT (Coin betting) vs Model-X (Exponential)

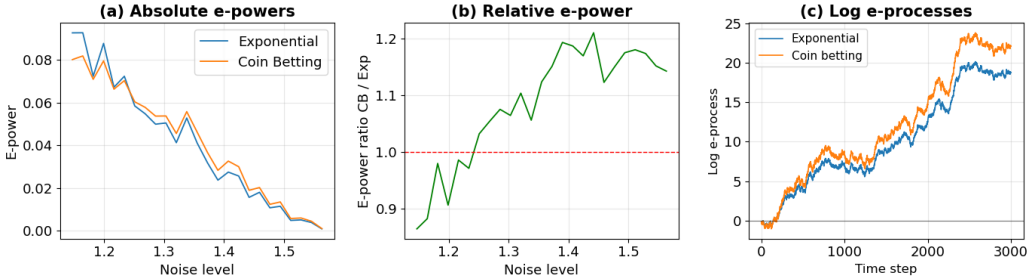


Figure 7: E-power comparison of the Exponential and Coin Betting across different noise levels.

- We prove that neither strictly dominates the other in practice.

e-CRT (Coin betting) vs Model-X (Exponential)

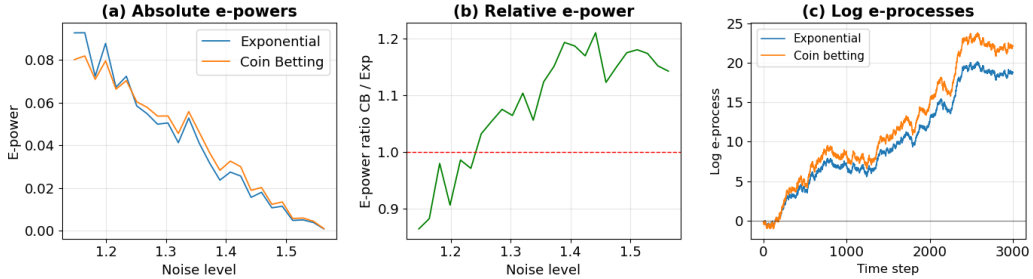


Figure 7: E-power comparison of the Exponential and Coin Betting across different noise levels.

- We prove that neither strictly dominates the other in practice.
- Which antisymmetric should we choose?

Exchangeability GRO e-variable

$$E^{\widehat{f}}_{\text{exch}} := 1 + \mathbb{E}_{\widetilde{X}|Z}\left[\frac{\widehat{f}(I,\widetilde{I}) - \widehat{f}(\widetilde{I},I)}{\widehat{f}(I,\widetilde{I}) + \widehat{f}(\widetilde{I},I)}\right] = \mathbb{E}_{\widetilde{X}|Z}\left[\frac{2\widehat{f}(I,\widetilde{I})}{\widehat{f}(I,\widetilde{I}) + \widehat{f}(\widetilde{I},I)}\right],$$

where $\widehat{f}$ approximates the density of $(I,\widetilde{I})$ via Kernel Density Estimation.

- The GRO e-variable for testing exchangeability null $(A,B) \overset{d}{=} (B,A)$ is:

$$\frac{2f(A,B)}{f(A,B)+f(B,A)}.$$

- For good approximations $\widetilde{f} \approx f$ and any antisymmetric g:

$$\mathbb{E}\left[\log\left(\frac{2\widehat{f}(I,\widetilde{I})}{\widehat{f}(I,\widetilde{I}) + \widehat{f}(\widetilde{I},I)}\right)\right] \geq \mathbb{E}\left[\log\left(1 + g(I,\widetilde{I})\right)\right].$$

- We simplify the null to $(I,\widetilde{I}) \overset{\text{d}}{=} (\widetilde{I},I)$ + Model-X assumption.
- $E^{\widehat{f}}_{\text{exch}}$ approximates the GRO **among antisymmetric** e-variables!

Analysis under the null

