Beyond Coefficients: Understanding Variable Importance in Modern ML

CoMeEcon

Angel Reyero Lobo

Institut de Mathémathiques de Toulouse & MIND, Inria Paris-Saclay
Joint work with:
Joseph Paillard

9th of June

Contents

Index

About me

- I am a PhD student at the Institut de Mathématiques de Toulouse and in the MIND team at Inria Paris-Saclay
- My advisors are Pierre Neuvial (IMT) and Bertrand Thirion (Inria Paris-Saclay)
- I am an ELLIS PhD student, so I did a 6-month research visit: I came to Amsterdam to CWI to visit Peter Grünwald
- My research interests are: Interpretable Machine Learning, Conditional Independence Testing, e-values, and Missing Data




Index

Setting: Supervised learning in tabular data

We have a dataset $\{(x_i, y_i)\}_{i=1}^n$,where $x_i \in \mathbb{R}^p$ and $y_i \in \mathbb{R}$.
The goal is to learn a function

$$\widehat{f} : \mathscr{X} \to \mathscr{Y}$$

such that

$$\widehat{f}(x) \approx y$$

for unseen data points (x, y).

To evaluate the predictive performance of a model $f : \mathscr{X} \to \mathscr{Y}'$, we use a loss function

$$\ell : \mathscr{Y} \times \mathscr{Y}' \to \mathbb{R}_+,$$

for example:

- Quadratic loss: $\ell(y, \widehat{y}) = (y - \widehat{y})^2$
- Classification 0–1 loss: $\ell(y, \widehat{y}) = \mathbf{1}_{\{y \neq \widehat{y}\}}$

Explainable AI and Scientific Discovery

Two main goals of explainability:

- **Local Variable Importance**
  - Explains a single prediction for an individual instance
    - e.g., Why did a credit scoring model reject Kayané's loan application?

Explainable AI and Scientific Discovery

Two main goals of explainability:

- **Local Variable Importance**
  - Explains a single prediction for an individual instance
    - e.g., Why did a credit scoring model reject Kayané's loan application?
  - Often leads to *counterfactual explanations*:
    - What minimal changes would flip the decision?
    - e.g., Research does not pay enough, she should change the job.

Explainable AI and Scientific Discovery

Two main goals of explainability:

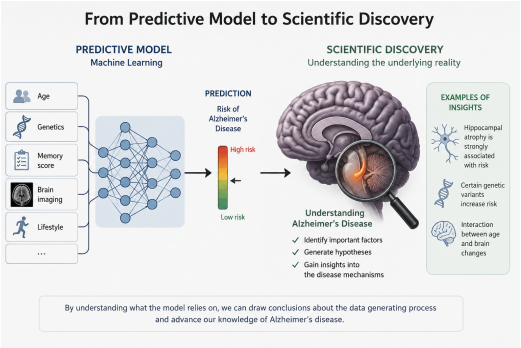
- **Local Variable Importance**
  - Explains a single prediction for an individual instance
    - e.g., Why did a credit scoring model reject Kayané's loan application?
  - Often leads to *counterfactual explanations*:
    - What minimal changes would flip the decision?
    - e.g., Research does not pay enough, she should change the job.
- **Global Variable Importance**
  - Explains how features influence the model overall
  - e.g., Is income generally important for credit approval? what about eye color?

Focus of this talk: Global Variable Importance
Günther et al. (2025) pessimistic about Local Variable Importance for complex models.

Explainable AI and Scientific Discovery

Two main motivations for studying **Global Variable Importance**:

- **Explainable AI (xAI)**:
  - understand the model itself, i.e., identify the features on which the model relies.
- **Scientific Discovery**:
  - use a predictive model to understand reality. Understanding which variables drive predictions can provide insights into the underlying data-generating process.



Linear Models

Linear Model assumption: $Y = \sum_{j=1}^{p} \beta_j X^j + \varepsilon$,where ε is centered independent noise.

We estimate the relationship using

$$\widehat{f}(X) = \sum_{j=1}^{p} \widehat{\beta}_j X^j.$$

Linear Models

Linear Model assumption: $Y = \sum_{j=1}^{p} \beta_j X^j + \varepsilon$,where ε is centered independent noise.

We estimate the relationship using

$$\widehat{f}(X) = \sum_{j=1}^{p} \widehat{\beta}_j X^j.$$

Interpretability comes naturally:

$$|\widehat{\beta}_j|$$

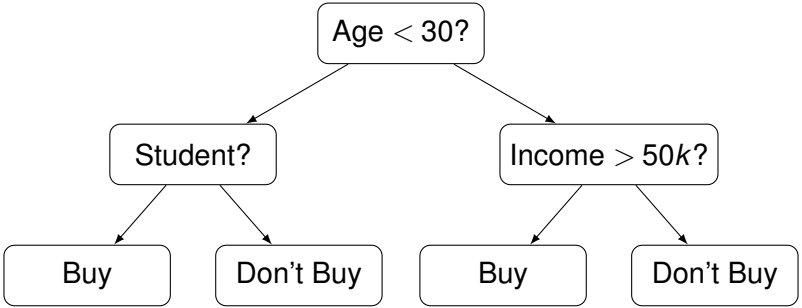
measures the influence of feature X^j on the prediction.

The same idea extends to:

- Generalized Linear Models (e.g. Logistic Regression),
- Regularized linear models (Ridge, Lasso).

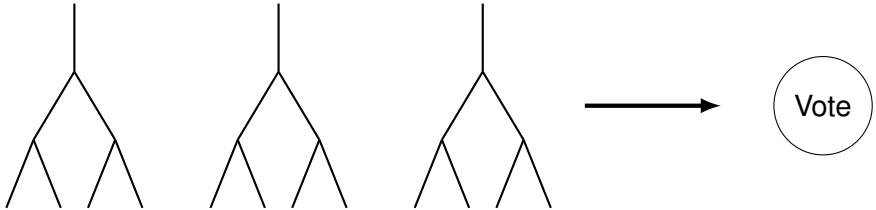
Limitation: linear models struggle to capture nonlinear interactions.

Decision Trees



- Easy to visualize and explain,
- Naturally capture interactions and nonlinearities,
- But often unstable and less predictive than ensemble methods.

Random Forests



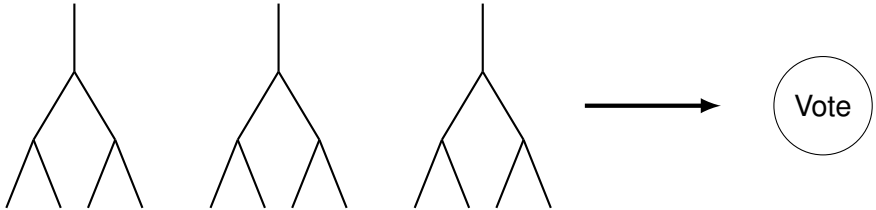
A random forest aggregates many decision trees trained on random subsets of:

- observations (bootstrap sampling),
- features (feature subsampling).

Advantages:

- Strong predictive performance,
- Captures complex interactions,
- More stable than a single tree.

Random Forests



A random forest aggregates many decision trees trained on random subsets of:

- observations (bootstrap sampling),
- features (feature subsampling).

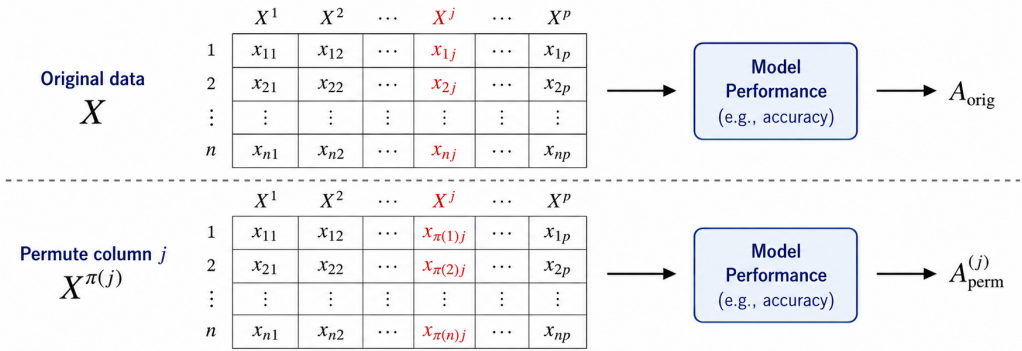
Advantages:

- Strong predictive performance,
- Captures complex interactions,
- More stable than a single tree.

? How can we assign an importance to X^j in the RF model?

Mean Decrease Accuracy (MDA)

Mean Decrease Accuracy (MDA)



$$\text{MDA}_j = A_{\text{orig}} - A^{(j)}_{\text{perm}}$$

If permuting feature j decreases the performance (i.e., $A^{(j)}_{\text{perm}} < A_{\text{orig}}$), then X^j is **important**.

Mean Decrease Accuracy (MDA)

Idea: a feature is important if shuffling its values decreases the model accuracy.

For a feature X^j:

1. Compute the prediction error on the test set,

$$\mathrm{Err}_{\mathrm{orig}} = \frac{1}{n_{\mathrm{test}}} \sum_{i=1}^{n_{\mathrm{test}}} \ell(\widehat{f}(x_i), y_i).$$

2. Randomly permute the values of the feature X^j among individuals: $X^{\pi(j)}$.
3. Recompute the prediction error,

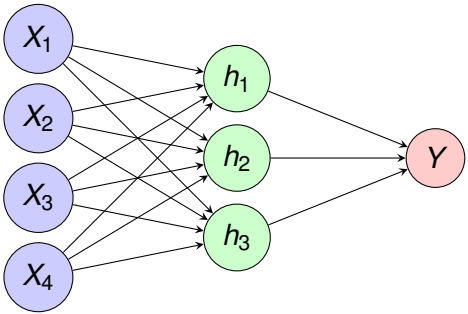
$$\mathrm{Err}^{(j)}_{\mathrm{perm}} = \frac{1}{n_{\mathrm{test}}} \sum_{i=1}^{n_{\mathrm{test}}} \ell(\widehat{f}(x_i^{\pi(j)}), y_i).$$

4. Define the importance score: $\mathrm{MDA}_j = \mathrm{Err}^{(j)}_{\mathrm{perm}} - \mathrm{Err}_{\mathrm{orig}}$.

Large increase in error $\Rightarrow$ important feature.

Model-agnostic Variable Importance

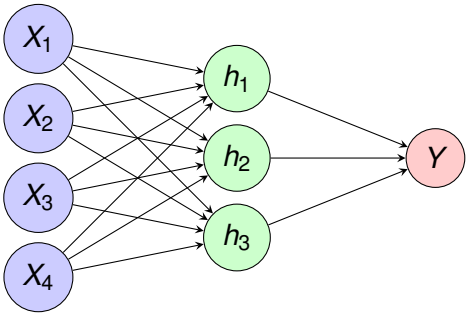
Modern machine learning models include:
Gradient Boosting, Neural Networks,
SuperLearners, foundation models, etc.



Question: how do we interpret such complex black-box models?

Model-agnostic Variable Importance

Modern machine learning models include:
Gradient Boosting, Neural Networks,
SuperLearners, foundation models, etc.



Question: how do we interpret such complex black-box models?

A key idea is to generalize Mean Decrease Accuracy (MDA) to *any* predictor $\widehat{f}$.

This leads to **Permutation Feature Importance (PFI)**.

Permutation Feature Importance (PFI)

PFI measures the increase in prediction loss when a feature is randomly permuted.

Let $\mathscr{D}_{\text{test}} = \{(x_i, y_i)\}_{i=1}^{n_{\text{test}}}$.

Define the importance of feature j as:

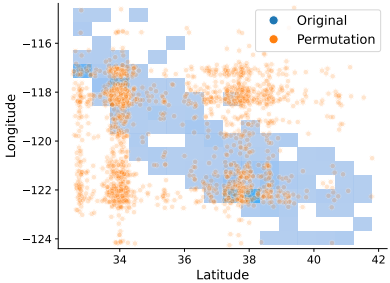
$$\text{PFI}_j = \frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} \ell\big(\widehat{f}(x_i^{\pi(j)}), y_i\big) - \frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} \ell\big(\widehat{f}(x_i), y_i\big).$$

where $x_i^{\pi(j)}$ is obtained by replacing the j-th feature value of x_i with the value of feature j from another randomly chosen observation.

Permutation Feature Importance (PFI): Limitations

Although PFI is model-agnostic and widely used, it has several issues:

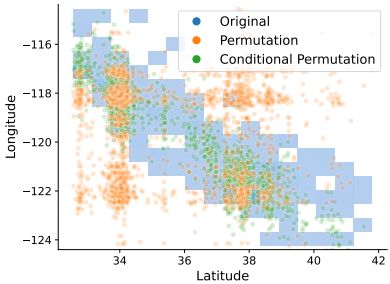
- **Theoretical limitations:** PFI does not correspond to an interpretable estimand (Bénard et al. (2022)).
- **Extrapolation issues:** permuting features may create unrealistic samples outside the data distribution (Strobl et al. (2008); Hooker et al. (2021)).



Permutation Feature Importance (PFI): Limitations

Although PFI is model-agnostic and widely used, it has several issues:

- **Theoretical limitations:** PFI does not correspond to an interpretable estimand (Bénard et al. (2022)).
- **Extrapolation issues:** permuting features may create unrealistic samples outside the data distribution (Strobl et al. (2008); Hooker et al. (2021)).



Key idea: permute while keeping realistic observations.

Conditional Feature Importance (CFI)

PFI

$$P_{(X^{\pi(j)},Y)} = P_{(Y,X^{-j})}\, P_{X^j}$$

- × Breaks X^j–Y
- × Breaks X^j–X^{-j}
- × Unrealistic samples

Conditional Feature Importance (CFI)

PFI

$$P_{(X^{\pi(j)},Y)} = P_{(Y,X^{-j})}\, P_{X^j}$$

- × Breaks X^j–Y
- × Breaks X^j–X^{-j}
- × Unrealistic samples

CFI

$$P_{(\tilde{X}^{(j)},Y)} = P_{(Y,X^{-j})}\, P_{X^j|X^{-j}}$$

- ✓ Preserves X^j–X^{-j}
- ✓ Realistic samples
- ✓ Breaks X^j–$Y \mid X^{-j}$

Key idea: condition on the remaining features X^{-j} before resampling X^j.

Conditional Feature Importance (CFI)

CFI measures the increase in prediction loss when a feature is *conditionally* permuted.

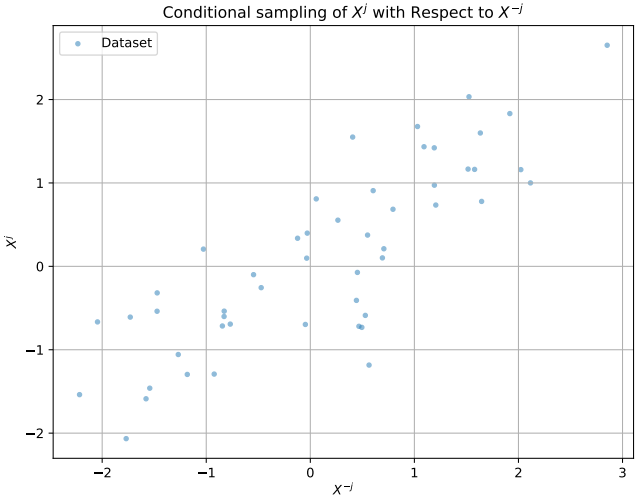
Let $\mathscr{D}_{\text{test}} = \{(x_i, y_i)\}_{i=1}^{n_{\text{test}}}$.

Define the importance of feature j as:

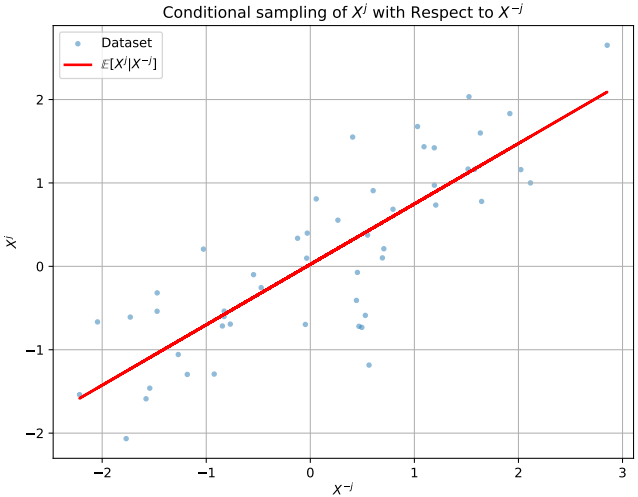
$$\psi_{\text{CFI}}(j) = \frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} \ell\big(\widehat{f}(\tilde{x}_i^{(j)}), y_i\big) - \frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} \ell\big(\widehat{f}(x_i), y_i\big).$$

where $\tilde{x}_i^{(j)}$ replaces the j-th feature value of x_i with a new sample from $\mathscr{L}(X^j \mid X^{-j})$.

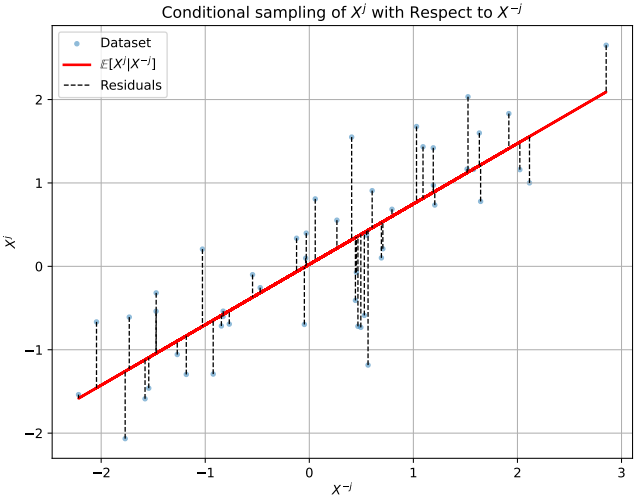
Conditional Permutation



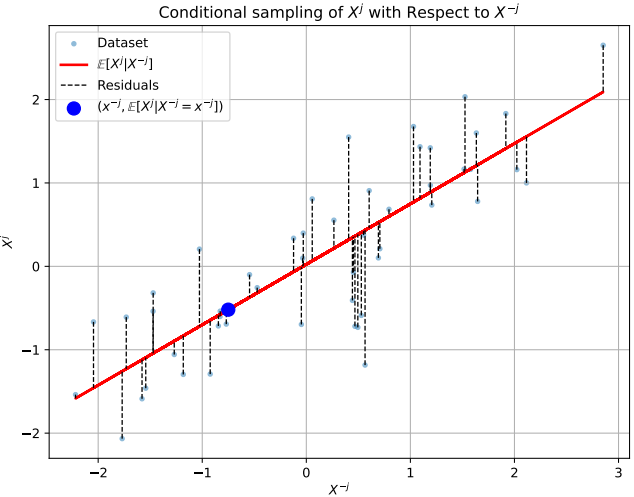
Conditional Permutation



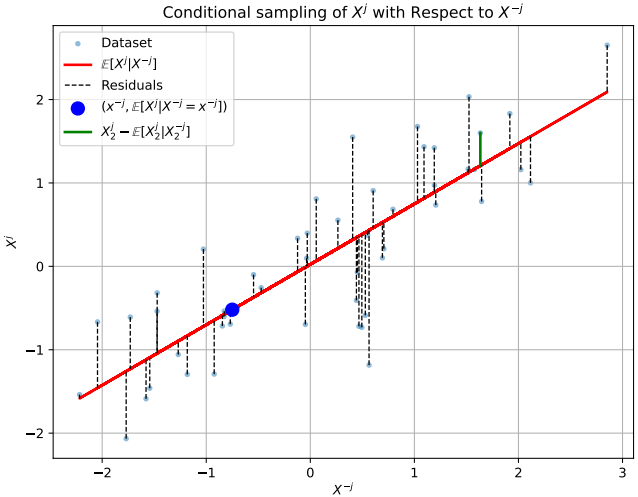
Conditional Permutation



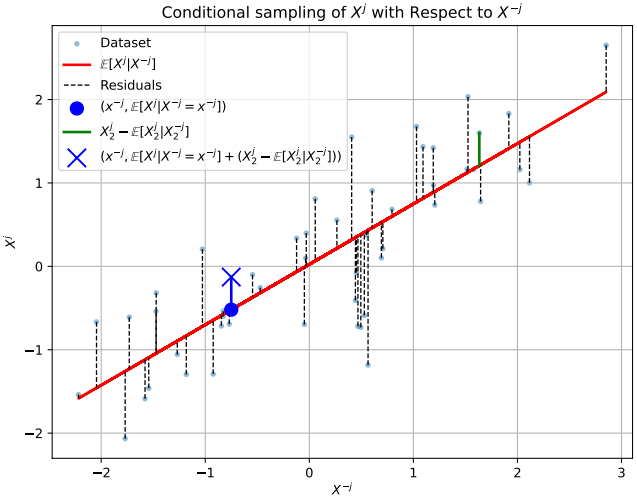
Conditional Permutation



Conditional Permutation



Conditional Permutation



Conditional Feature Importance using conditional permutation

Step 1: Learn the conditional mean

$$\widehat{v}_j(X^{-j}) \approx \mathbb{E}[X^j \mid X^{-j}]$$

$$\Downarrow$$

Step 2: Permute residuals

$$\widehat{\varepsilon}_j = X^j - \widehat{v}_j(X^{-j}), \qquad \widetilde{X}^{(j)} = \widehat{v}_j(X^{-j}) + \widehat{\varepsilon}_j^{\pi}$$

$$\Downarrow$$

Step 3: Measure loss increase

$$\psi_{\mathrm{CFI}}(j) = \frac{1}{n_{\mathrm{test}}} \sum_{i=1}^{n_{\mathrm{test}}} \ell\big(\widehat{f}(\widetilde{x}_i^{(j)}), y_i\big) - \frac{1}{n_{\mathrm{test}}} \sum_{i=1}^{n_{\mathrm{test}}} \ell\big(\widehat{f}(x_i), y_i\big).$$

⚠ Permute residuals, not features.

Leave One Covariate Out (LOCO)

LOCO measures the importance of a feature by evaluating a model that has been refit **without** that feature.

1 **Test data** $\{(x_i, y_i)\}_{i=1_{test}}^{n_{test}}$

	x^1	…	x^j	…	x^p	y
1	.	.	.	.	.	.
2	.	.	.	.	:	.
⋮	.	.	⋮	.	:	⋮
n_{test}						

All covariates (including j)

2 **Remove the j-th covariate**

	x^1	…	x^{j-1}		x^{j+1}	…	x^p	y
1	.	.			.	.	.	.
2	:	:			:	.	:	.
⋮	.	.	⋮		⋮	.	:	⋮
n_{test}								

All covariates except j

3 **Refit the model without X^j**

Fit a model using only X^{-j} (and y):

$\hat{f}_{-j} : \mathbb{R}^{p-1} \to \mathbb{R}$

4 **Predict and evaluate on test data**

Predictions $\hat{f}_{-j}(x_i^{-j})$

Compare with y_i using loss function $\ell(\cdot, \cdot)$

LOCO importance of covariate j

$$\psi_{\text{LOCO}}(j) = \frac{1}{n_{test}} \sum_{i=1}^{n_{test}} \left[\ell\left(\hat{f}_{-j}(x_i^{-j}), y_i\right) - \ell\left(\hat{f}(x_i), y_i\right) \right]$$

- x_i^{-j} : all covariates except j for observation i
- $\hat{f}_{-j}$: model refit without X^j
- $\hat{f}$: full model using all covariates
- $\ell(\cdot, \cdot)$: loss function

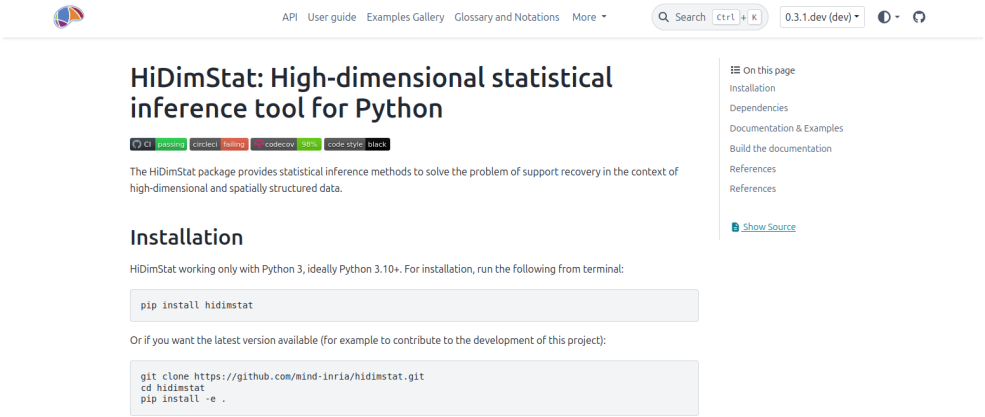
Leave-One-Covariate-Out (LOCO)

We measure feature importance by comparing model performance with and without X^j:

$$\psi_{\text{LOCO}}(j) = \frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} \left[\ell\big(\hat{f}_{-j}(x_i^{-j}), y_i\big) - \ell\big(\hat{f}(x_i), y_i\big) \right]$$

- It is computationally expensive, since we must retrain a model for each feature.
- It converges to the total Sobol' index, a standard estimand in sensitivity analysis.
- Due to optimization error, it is more unstable than CFI (Reyero-Lobo et al. (2025b)).

HiDimStat: https://hidimstat.github.io



Index

How Do We Compare Variable Importance Measures?

We have encountered several notions of importance:

- Linear model coefficients,
- Permutation-based importance (PFI/CFI),
- Leave-One-Covariate-Out importance (LOCO).

A natural question arises:

How can we compare these different notions of importance?

How Do We Compare Variable Importance Measures?

We have encountered several notions of importance:

- Linear model coefficients,
- Permutation-based importance (PFI/CFI),
- Leave-One-Covariate-Out importance (LOCO).

A natural question arises:

How can we compare these different notions of importance?

Some authors argue that LOCO and PFI/CFI are fundamentally different because

- LOCO compares *two predictive models*,
- PFI/CFI evaluates a *single model* under perturbed inputs.

How Do We Compare Variable Importance Measures?

We have encountered several notions of importance:

- Linear model coefficients,
- Permutation-based importance (PFI/CFI),
- Leave-One-Covariate-Out importance (LOCO).

A natural question arises:

How can we compare these different notions of importance?

Some authors argue that LOCO and PFI/CFI are fundamentally different because

- LOCO compares *two predictive models*,
- PFI/CFI evaluates a *single model* under perturbed inputs.

Exercise. Assume

$$Y = f_\star(X) + \varepsilon, \qquad \varepsilon \perp\!\!\!\perp X, \qquad \mathbb{E}[\varepsilon] = 0, \qquad \mathbb{E}[\varepsilon^2] < \infty.$$

Show that under squared loss,

$$\boxed{\psi_{\mathrm{CFI}}(j) = 2\,\psi_{\mathrm{LOCO}}(j)}.$$

CFI and 2×LOCO are Equivalent

Under squared loss,

$$\psi_{\mathrm{CFI}}(j) = \mathbb{E}\Big[\big(f_\star(\tilde{X}^{(j)}) - Y\big)^2 - \big(f_\star(X) - Y\big)^2\Big].$$

Using

$$Y = f_\star(X) + \varepsilon, \qquad \mathbb{E}[\varepsilon] = 0,$$

we obtain

$$\psi_{\mathrm{CFI}}(j) = \mathbb{E}\Big[\big(f_\star(\tilde{X}^{(j)}) - f_\star(X)\big)^2\Big].$$

Since $\tilde{X}^j \overset{d}{=} X^j, \tilde{X}^j \perp\!\!\!\perp X^j \mid X^{-j}$, the variables $f_\star(\tilde{X}^{(j)}), f_\star(X)$ are i.i.d. given X^{-j}. Therefore, using Tower's property, we obtain that

$$\boxed{\psi_{\mathrm{CFI}}(j) = 2\,\mathbb{E}\Big[\mathrm{Var}\big(f_\star(X) \mid X^{-j}\big)\Big]}$$

Computing LOCO

Recall

$$\psi_{\text{LOCO}}(j) = \mathbb{E}\Big[\big(f_{-j,\star}(X^{-j}) - Y\big)^2 - \big(f_\star(X) - Y\big)^2\Big].$$

Since

$$f_{-j,\star}(X^{-j}) = \mathbb{E}\big[f_\star(X) \mid X^{-j}\big],$$

the additive independent centered noise yields

$$\psi_{\text{LOCO}}(j) = \mathbb{E}\Big[\big(f_{-j,\star}(X^{-j}) - f_\star(X)\big)^2\Big] = \mathbb{E}\Big[\text{Var}\big(f_\star(X) \mid X^{-j}\big)\Big].$$

Hence,

$$\boxed{\psi_{\text{CFI}}(j) = 2\,\psi_{\text{LOCO}}(j)}.$$

How Should We Compare VIMs?

Exercise.
Assume the linear model $Y = \sum_{k=1}^{p} \beta_k X^k + \varepsilon$. Show that

$$\psi_{\text{LOCO}}(j) = \beta_j^2 \, \mathbb{E}\Big[\text{Var}(X^j \mid X^{-j})\Big].$$

How Should We Compare VIMs?

Exercise.
Assume the linear model $Y = \sum_{k=1}^{p} \beta_k X^k + \varepsilon$. Show that

$$\psi_{\text{LOCO}}(j) = \beta_j^2 \, \mathbb{E}\Big[\text{Var}(X^j \mid X^{-j})\Big].$$

Indeed,

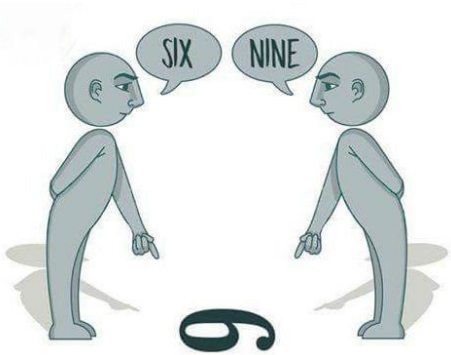
$$\psi_{\text{LOCO}}(j) = \mathbb{E}\Big[\text{Var}\big(f_\star(X) \mid X^{-j}\big)\Big] = \mathbb{E}\left[\text{Var}\Big(\sum_{k=1}^{p} \beta_k X^k \mid X^{-j}\Big)\right] = \beta_j^2 \mathbb{E}\Big[\text{Var}(X^j \mid X^{-j})\Big].$$

- Stronger correlations imply smaller LOCO importance.
- Regression coefficients do *not* account for redundancy among predictors.
- Different VIMs quantify fundamentally different notions of importance.

For a general framework to compare VIMs, see Reyero-Lobo et al. (2025a).

The Rashomon Effect

- Many different models can explain the same dataset **equally well**.
- However, these models may rely on **different patterns or features** in the data.
- This phenomenon is known as the **Rashomon effect**.



Aggregate Models or Aggregate Importances?

For each trained model $\{\widehat{f}_b\}_{b=1}^B$, we obtain an importance estimate

$$\widehat{\psi}_b := \psi(\widehat{f}_b).$$

This raises an important question:

- **Aggregate the importances**

$$\widehat{\psi} = \frac{1}{B}\sum_{b=1}^{B} \widehat{\psi}_b$$

- **Aggregate the models first, then compute the importance**

$$\widehat{\psi} = \psi\left(\frac{1}{B}\sum_{b=1}^{B} \widehat{f}_b\right)$$

Should we aggregate the models or the importance measures?

Aggregate Models, Not Explanations

We proved that the excess risk satisfies

$$\frac{1}{n}\sum \ell(\widehat{f}(x_i), y_i) - \mathbb{E}[\ell(f_\star(X), Y)] = \varepsilon + O_P(n^{-1/2}) = \mathbb{E}\left[\|\widehat{f} - f_\star\|^2\right] + O_P(n^{-1/2}).$$

- The term $O_P(n^{-1/2})$ corresponds to the **test/estimation error**.
- The dominant term ε comes from the **training procedure**.
- The importance bias comes mainly from model bias!
  We need model-agnostic measures; otherwise, we can't explain the distribution.

Key message: Most of the error in ML models comes from **training the model**, not from test estimation.

Aggregating Before vs After

We compare two strategies:

- **Aggregate models first (Ensemble)**

$$\psi_n^{\text{ens}} = \psi(\widehat{f}_{\text{ens}})$$

- **Aggregate explanations after training (Sub-models)**

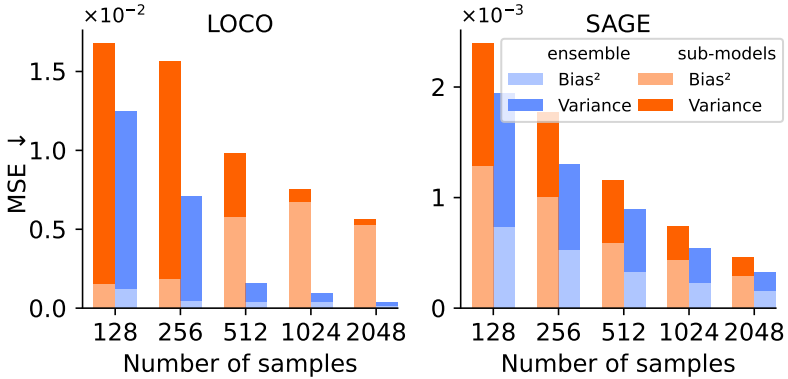
$$\psi_n^{\text{sub}} = \frac{1}{B}\sum_{b=1}^{B} \psi(\widehat{f}_b)$$

We obtain the relation

$$\psi_n^{\text{ens}} - \psi_\star = (\psi_n^{\text{sub}} - \psi_\star)\left(\rho + \frac{1-\rho}{B}\right) + O_P(n^{-1/2})$$

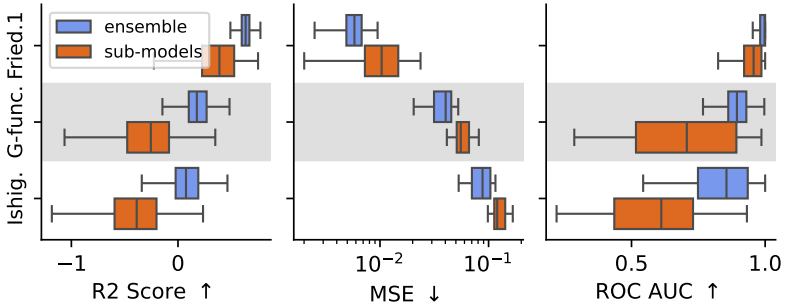
Conclusion: Aggregating models first (ensemble) leads to a smaller error.

Simulations



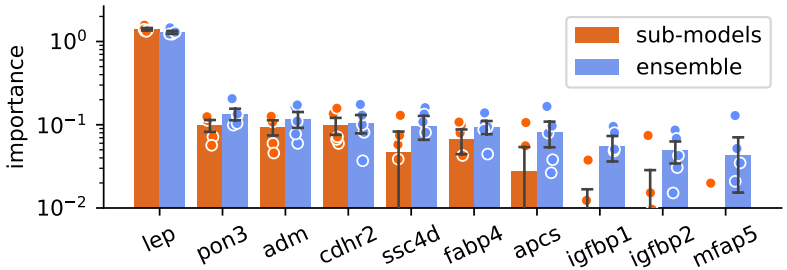
Aggregation of models (ensemble) vs aggregation of explanations (sub-models).

Simulations



Simulation results comparing aggregation of models (ensemble) with aggregation of explanations(sub-models).

Real data



UKBiobank comparing aggregation of models (ensemble) with aggregation of explanations (sub-models).

Statistical Guarantees for Scientific Discovery

We often use machine learning models for **scientific discovery**:

Which variables truly influence the outcome?

Marginal independence, $X^j \perp\!\!\!\perp Y$, is not enough:

- Variables spuriously correlated with important ones are considered important.

Statistical Guarantees for Scientific Discovery

We often use machine learning models for **scientific discovery**:

Which variables truly influence the outcome?

Marginal independence, $X^j \perp\!\!\!\perp Y$, is not enough:

- Variables spuriously correlated with important ones are considered important.

Instead, we test **Conditional Independence Testing (CIT)**

$$H_0 : \quad X^j \perp\!\!\!\perp Y \mid X^{-j}.$$

- X^j is important if it provides information about Y that cannot be recovered from X^{-j}.
- If features are too correlated, consider **grouping** them.

To make scientific claims, we need **statistical guarantees** (e.g. type-I or FDR control).
Many examples to do so in HiDimStat:
`https://hidimstat.github.io/stable/generated/gallery/examples`

The Model-X Principle

Most modern CIT procedures rely on the **Model-X assumption**:

$$\widetilde{X}^{(j)} \sim \mathscr{L}(X^j \mid X^{-j}).$$

We can sample copies of X^j while preserving its dependence with the rest.

The Model-X Principle

Most modern CIT procedures rely on the **Model-X assumption**:

$$\widetilde{X}^{(j)} \sim \mathscr{L}(X^j \mid X^{-j}).$$

We can sample copies of X^j while preserving its dependence with the rest.
Given an importance statistic $T(X^j, Y, X^{-j})$, we compare it with $T(\widetilde{X}^{(j)}, Y, X^{-j})$.
Under the null hypothesis ($H_0 : \quad X^j \perp\!\!\!\perp Y \mid X^{-j}$,) the two quantities are exchangeable:

$$T(X^j, Y, X^{-j}) \overset{\mathscr{L}}{=} T(\widetilde{X}^{(j)}, Y, X^{-j}).$$

This simple observation is the foundation of

- Conditional Randomization Tests (CRT),
- Model-X Knockoffs.

What should the test statistic $T(X^j, Y, X^{-j})$ be?

The power of the procedure depends on the choice of $T(X^j, Y, X^{-j})$.
Traditionally, one often uses the absolute value of a regression coefficient:

$$T = |\hat{\beta}_j|.$$

What should the test statistic $T(X^j, Y, X^{-j})$ be?

The power of the procedure depends on the choice of $T(X^j, Y, X^{-j})$.
Traditionally, one often uses the absolute value of a regression coefficient:

$$T = |\hat{\beta}_j|.$$

But isn't this just an
Interpretability / Variable Importance measure?

For linear models, $T = |\hat{\beta}_j|$ is a natural importance score.
Question:
How can we construct powerful CIT using importance measures from modern ML?

Conditional Independence Testing with Interpretable ML

Many questions arise when using interpretable ML for conditional independence testing:

- How can we test whether $\psi_{\text{CFI}}(j) = 0$?
- How can we make the Conditional Randomization Test model-agnostic?
- How can we make knockoffs model-agnostic?
- How can we obtain sequentially valid procedures?

Conditional Independence Testing with Interpretable ML

Many questions arise when using interpretable ML for conditional independence testing:

- How can we test whether $\psi_{\text{CFI}}(j) = 0$?
  - The CLT is not valid due to the vanishing variance under the null.
  - We can use variance corrections or nonparametric tests (Reyero-Lobo et al. (2025b)).
- How can we make the Conditional Randomization Test model-agnostic?
- How can we make knockoffs model-agnostic?
- How can we obtain sequentially valid procedures?

Conditional Independence Testing with Interpretable ML

Many questions arise when using interpretable ML for conditional independence testing:

- How can we test whether $\psi_{\text{CFI}}(j) = 0$?
  - The CLT is not valid due to the vanishing variance under the null.
  - We can use variance corrections or nonparametric tests (Reyero-Lobo et al. (2025b)).
- How can we make the Conditional Randomization Test model-agnostic?
  - The Holdout Randomization Test (HRT) is model-agnostic but requires a train–test split.
  - Semi-knockoffs (Reyero-Lobo et al. (2026)) avoid the need for a train–test split.
- How can we make knockoffs model-agnostic?
- How can we obtain sequentially valid procedures?

Conditional Independence Testing with Interpretable ML

Many questions arise when using interpretable ML for conditional independence testing:

- How can we test whether $\psi_{\text{CFI}}(j) = 0$?
  - The CLT is not valid due to the vanishing variance under the null.
  - We can use variance corrections or nonparametric tests (Reyero-Lobo et al. (2025b)).
- How can we make the Conditional Randomization Test model-agnostic?
  - The Holdout Randomization Test (HRT) is model-agnostic but requires a train–test split.
  - Semi-knockoffs (Reyero-Lobo et al. (2026)) avoid the need for a train–test split.
- How can we make knockoffs model-agnostic?
  - **Semi-knockoffs** can control the FDR through the knockoff threshold.
  - They do not need conditional densities but conditional expectations.
- How can we obtain sequentially valid procedures?

Conditional Independence Testing with Interpretable ML

Many questions arise when using interpretable ML for conditional independence testing:

- How can we test whether $\psi_{\text{CFI}}(j) = 0$?
  - The CLT is not valid due to the vanishing variance under the null.
  - We can use variance corrections or nonparametric tests (Reyero-Lobo et al. (2025b)).
- How can we make the Conditional Randomization Test model-agnostic?
  - The Holdout Randomization Test (HRT) is model-agnostic but requires a train–test split.
  - Semi-knockoffs (Reyero-Lobo et al. (2026)) avoid the need for a train–test split.
- How can we make knockoffs model-agnostic?
  - **Semi-knockoffs** can control the FDR through the knockoff threshold.
  - They do not need conditional densities but conditional expectations.
- How can we obtain sequentially valid procedures?
  - We can use **e-values**.
  - Joint work with: Sebastian Arias, Michele Meziu, and Peter Grünwald.

A Quick Introduction to the E-World

E-value. A random variable S satisfying $S \geq 0, \quad \mathbb{E}_P[S] \leq 1 \quad \forall P \in \mathscr{H}_0$.
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 H_0 whenever $S \geq 1/\alpha$.

A Quick Introduction to the E-World

E-value. A random variable S satisfying $S \geq 0, \quad \mathbb{E}_P[S] \leq 1 \quad \forall P \in \mathscr{H}_0$.
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 H_0 whenever $S \geq 1/\alpha$.

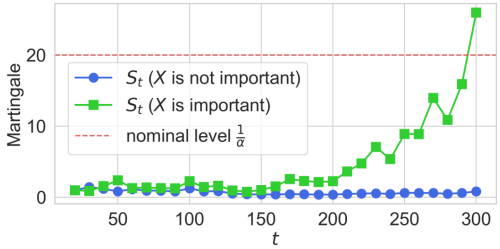
E-process. A sequence $(S_t)_{t \geq 1}$ such that

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

for every stopping time τ and every $P \in \mathscr{H}_0$.
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 $S_t \geq 1/\alpha$.



Sequential Conditional Independence Testing with ML

Goal: Test $Y \perp\!\!\!\perp X \mid Z$ assuming Model-X ($P_{X|Z}$ is known and $\tilde{X} \sim P_{X|Z}$).

Existing approaches:

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

$$\frac{f_{Y|X,Z}}{f_{Y|Z}}.$$

Sequential Conditional Independence Testing with ML

Goal: Test $Y \perp\!\!\!\perp X \mid Z$ assuming Model-X ($P_{X|Z}$ is known and $\tilde{X} \sim P_{X|Z}$).

Existing approaches:

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

$$\frac{f_{Y|X,Z}}{f_{Y|Z}}.$$

- **e-CRT** Shaer et al. (2023): based on importance statistics

$$q = T(X, Y, Z), \quad \tilde{q} = T(\tilde{X}, Y, Z),$$

and an antisymmetric score $g(\tilde{q}, q) = -g(q, \tilde{q})$ giving an e-value $1 + g(q, \tilde{q})$.

Sequential Conditional Independence Testing with ML

Goal: Test $Y \perp\!\!\!\perp X \mid Z$ assuming Model-X ($P_{X|Z}$ is known and $\tilde{X} \sim P_{X|Z}$).

Existing approaches:

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

$$\frac{f_{Y|X,Z}}{f_{Y|Z}}.$$

- **e-CRT** Shaer et al. (2023): based on importance statistics

$$q = T(X, Y, Z), \quad \tilde{q} = T(\tilde{X}, Y, Z),$$

and an antisymmetric score $g(\tilde{q}, q) = -g(q, \tilde{q})$ giving an e-value $1 + g(q, \tilde{q})$.

Limitations: Model-X is growth-rate optimal under correct specification, but can be suboptimal under misspecification. e-CRT is flexible but leaves both T and g unspecified.

e-CRT vs Model-X

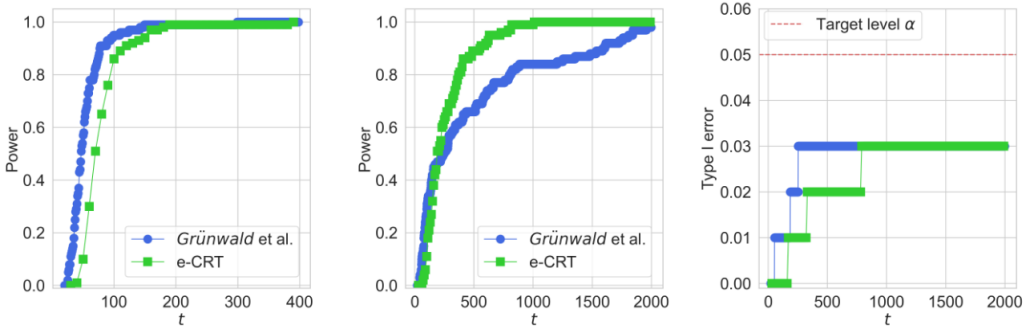


Figure 1: **From Shaer et al. (2023):** In the well-specified setting (left), Model-X is more powerful. However, in the misspecified setting (center), where the conditional density cannot be learned exactly, e-CRT can achieve higher power.

Sequential Conditional Independence Testing with ML

Our proposal:

$$\frac{2\,\hat{q}(t,\tilde{t})}{\hat{q}(t,\tilde{t})+\hat{q}(\tilde{t},t)},$$

where

$$t=\ell(\hat{m}(X,Z),Y),\quad \tilde{t}=\ell(\hat{m}(\tilde{X},Z),Y),$$

and $\hat{q}$ is an estimate of the joint density of $(t,\tilde{t})$.

- We estimate a 2D density, not a conditional density $f_{Y|X,Z}$.
- We can derandomize it:

$$\mathbb{E}_{\tilde{X}|X,Y,Z}\left[\frac{2\,\hat{q}(t,\tilde{t})}{\hat{q}(t,\tilde{t})+\hat{q}(\tilde{t},t)}\right].$$

Joint work with: Sebastian Arias, Michele Meziu, Peter Grünwald.

Index

Conclusion

- Machine learning has become a powerful tool for **scientific discovery**:
  - It can capture complex patterns that may not have been identified previously.
- Modern models are highly complex, motivating **model-agnostic** interpretability.
- The general pipeline of IML should be:
  1. define what we mean by *importance*, which determines the **target quantity** of interest
  2. compare different **estimators** (e.g., LOCO and CFI) and their properties, e.g. stability, computational cost, robustness to misspecification.
  3. derive rigorous **statistical guarantees**.

Conditional independence testing uses interpretability to construct powerful tests.

Everything is already implemented in `https://hidimstat.github.io`!

Index

References

Clément Bénard, Sébastien Da Veiga, and Erwan Scornet. Mean decrease accuracy for random forests: inconsistency, and a practical solution via the sobol-mda. *Biometrika*, 109(4):881–900, 02 2022.

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.

Eric Günther, Balázs Szabados, Robi Bhattacharjee, Sebastian Bordt, and Ulrike von Luxburg. Informative post-hoc explanations only exist for simple functions, 2025. URL https://arxiv.org/abs/2508.11441.

References

Giles Hooker, Lucas Mentch, and Siyu Zhou. Unrestricted permutation forces extrapolation: variable importance requires at least one more model, or there is no free variable importance. *Statistics and Computing*, 31(82):1–16, 2021. doi: 10.1007/s11222-021-10057-z. URL https://doi.org/10.1007/s11222-021-10057-z.

Joseph Paillard, Angel Reyero Lobo, Denis A. Engemann, and Bertrand Thirion. Aggregate models, not explanations: Improving feature importance estimation. In *International Conference on Machine Learning (ICML)*, 2026.

Angel Reyero-Lobo, Pierre Neuvial, and Bertrand Thirion. A principled approach for comparing variable importance, 2025a. URL https://arxiv.org/abs/2507.17306.

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

References

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

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.

Carolin Strobl, Anne-Laure Boulesteix, Thomas Kneib, Thomas Augustin, and Achim Zeileis. Conditional variable importance for random forests. *BMC Bioinformatics*, 9(1): 307, 2008. ISSN 1471-2105. doi: 10.1186/1471-2105-9-307. URL https://doi.org/10.1186/1471-2105-9-307.

Thank you!

Questions?

angel.reyero-lobo@inria.fr
joseph.paillard@roche.com