A principled framework for comparing Variable Importance

Angel REYERO LOBO

Explanability for high-dimensional statistics workshop
IMT & Inria Paris-Saclay
Joint work with:
Pierre NEUVIAL & Bertrand THIRION.

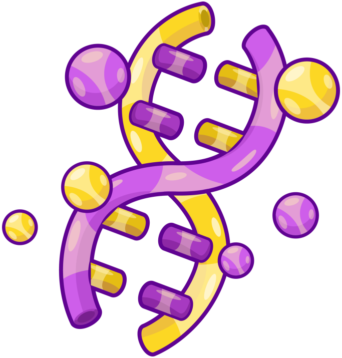
December 9, 2025

Contents

Index

Variable Importance Measures (VIM)

How can we define / learn the importance of each covariate X^j with respect to an outcome y?



X

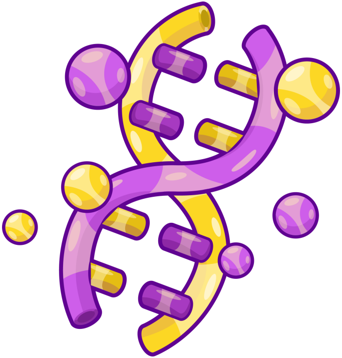
y

Try to study their relationship using a ML model:

$$\widehat{m} \in \operatorname*{argmin}_{f \in \mathscr{F}} \widehat{\mathbb{E}}\left[\mathscr{L}(f(X), y)\right]. \tag{1}$$

Variable Importance Measures (VIM)

How can we define / learn the importance of each covariate X^j with respect to an outcome y?



X

y

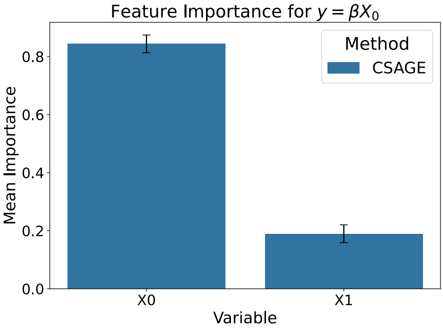
💡 Try to study their relationship using a ML model:

$$\widehat{m} \in \underset{f \in \mathscr{F}}{\operatorname{argmin}} \widehat{\mathbb{E}} \left[\mathscr{L}(f(X), y) \right]. \tag{1}$$

What does it mean to be important?

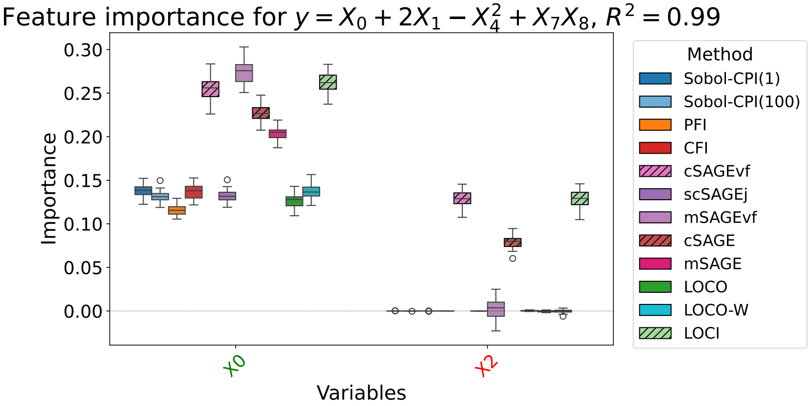
"Feature importance as how much predictive power it provides to the model. We can then define "important" features as those whose absence degrades m's performance."

Covert et al. (2020) NeurIPS



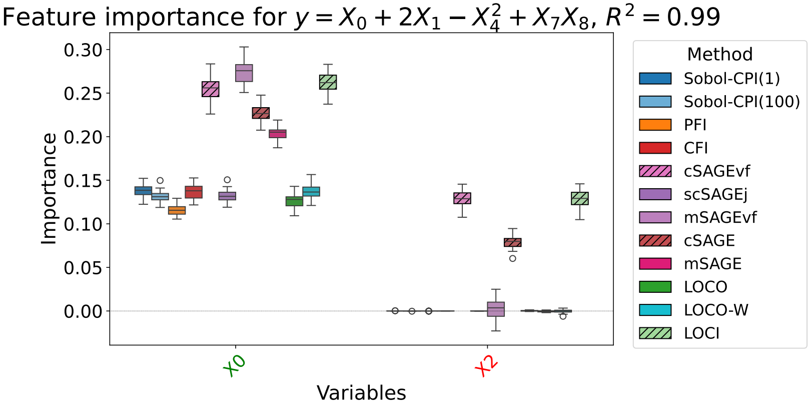
⚠ Gap between variable importance and variable selection.

Motivation



1. From X_0: How to compare VIMs?
2. From X_2: What's the minimum for a VIM?

Motivation

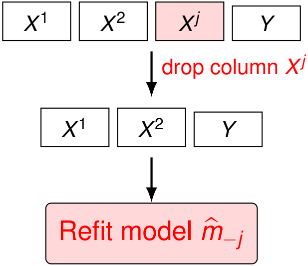


1. From X_0: How to compare VIMs?

2. From X_2: What's the minimum for a VIM?
   - **Minimal axiom:** $\psi(j, P) = 0$ if and only if $X^j \perp\!\!\!\perp y \mid X^{-j}$.

Index

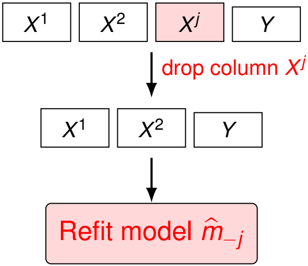
LOCO



$\Delta_{\mathrm{LOCO}} = \mathrm{perf}(\text{refit } \widehat{m}_{-j}) - \mathrm{perf}(\text{orig } \widehat{m})$

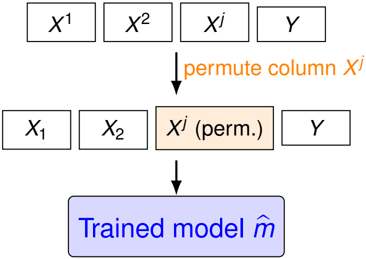
LOCO and PFI

LOCO



$$\Delta_{\text{LOCO}} = \text{perf}(\text{refit } \widehat{m}_{-j}) - \text{perf}(\text{orig } \widehat{m})$$

PFI



$$\Delta_{\text{PFI}} = \text{perf}(\widehat{m}(\text{perm } X^{(j)})) - \text{perf}(\widehat{m}(\text{orig X}))$$

How can we compare VIMs?

- **Leave One Covariate Out(LOCO)**:

$$\widehat{\psi}^{j}_{\mathrm{LOCO}} = \frac{1}{n_{\mathrm{test}}} \sum_{i=1}^{n_{\mathrm{test}}} \mathscr{L}(y_i, \widehat{m}_{-j}(x_i^{-j})) - \mathscr{L}(y_i, \widehat{m}(x_i)).$$

- **Permutation Feature Importance(PFI)**:

$$\widehat{\psi}^{j}_{\mathrm{PFI}} = \frac{1}{n_{\mathrm{test}}} \sum_{i=1}^{n_{\mathrm{test}}} \mathscr{L}(y_i, \widehat{m}(x_i^{(j)})) - \mathscr{L}(y_i, \widehat{m}(x_i)).$$

where the j-th covariate is permuted.

➔ LOCO uses *refitting* and PFI uses *perturbation*.

How can we compare VIMs?

*“**LOCO differs from the other methods [...] since most of the other methods don’t require retraining the model**. However, due to retraining the model, the interpretation shifts from only interpreting that one single model to interpreting the learner and how model training reacts to changes in the features.”*

Molnar (2025), Interpretable Machine Learning, 3rd Edition

How can we compare VIMs?

> *"**LOCO differs from the other methods [...] since most of the other methods don't require retraining the model**. However, due to retraining the model, the interpretation shifts from only interpreting that one single model to interpreting the learner and how model training reacts to changes in the features."*
>
> *Molnar (2025), Interpretable Machine Learning, 3rd Edition*

Basically, compare VIMs based on the **inference** procedure used.

Inconsistencies in comparison by inference

The Total Sobol Index can be estimated in many different ways!

$$\psi_{\mathrm{TSI}} := \mathbb{E}\left[\mathrm{Var}(y \mid X^{-j})\right]$$

Inconsistencies in comparison by inference

The Total Sobol Index can be estimated in many different ways!

$$
\begin{aligned}
\psi_{\mathrm{TSI}} &:= \mathbb{E}\left[\mathrm{Var}(y \mid X^{-j})\right] \\
&= \mathbb{E}\left[\left(m_{-j}(X^{-j}) - y\right)^2\right] - \mathbb{E}\left[(m(X) - y)^2\right] && \textit{refitting} \\
&= \frac{1}{2}\left(\mathbb{E}\left[\left(m(\widetilde{X}^{(j)}) - y\right)^2\right] - \mathbb{E}\left[(m(X) - y)^2\right]\right) && \textit{perturbation}
\end{aligned}
$$

Inconsistencies in comparison by inference

The Total Sobol Index can be estimated in many different ways!

$$
\begin{aligned}
\psi_{\mathrm{TSI}} :&= \mathbb{E}\left[\mathrm{Var}(y \mid X^{-j})\right] \\
&= \mathbb{E}\left[\left(m_{-j}(X^{-j}) - y\right)^2\right] - \mathbb{E}\left[(m(X) - y)^2\right] && \textit{refitting} \\
&= \frac{1}{2}\left(\mathbb{E}\left[\left(m(\widetilde{X}^{(j)}) - y\right)^2\right] - \mathbb{E}\left[(m(X) - y)^2\right]\right) && \textit{perturbation} \\
&= \mathbb{E}\left[\left(\mathbb{E}\left[m(X) \mid X^{-j}\right] - y\right)^2\right] - \mathbb{E}\left[(m(X) - y)^2\right] && \textit{marginalization} \\
&= \mathbb{E}\left[(m_{-j}(X^{-j}) - m(X))^2\right] && \textit{variance} \\
&= \sigma^2(R^2_{-j} - R^2).
\end{aligned}
$$

General pipeline

Theoretical index → Estimation → Type-I error

1. **Theoretical Index:**
   - Define goals and choose a matching theoretical quantity.
   - Verify if it satisfies the minimal axiom.

General pipeline

Theoretical index → Estimation → Type-I error

1. **Theoretical Index:**
   - Define goals and choose a matching theoretical quantity.
   - Verify if it satisfies the minimal axiom.
2. **Estimation:**
   - Select a procedure aligned with your desired inference properties:
     - E.g., double robustness, computational feasibility, extrapolation issues, benefit from unlabeled data, or simpler relationships between inputs than between inputs and outputs, . . .

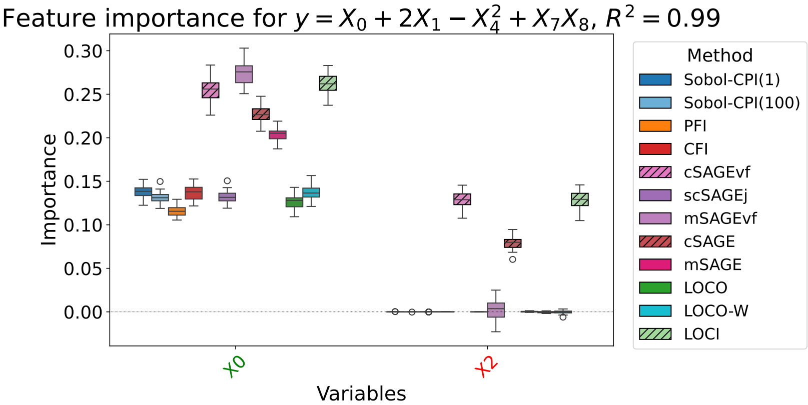
General pipeline

Theoretical index → **Estimation** → **Type-I error**

1. **Theoretical Index:**
   - Define goals and choose a matching theoretical quantity.
   - Verify if it satisfies the minimal axiom.
2. **Estimation:**
   - Select a procedure aligned with your desired inference properties:
     - E.g., double robustness, computational feasibility, extrapolation issues, benefit from unlabeled data, or simpler relationships between inputs than between inputs and outputs, ...
3. **Type-I error:**
   - Provide statistical guarantees for the important covariates.

Index

Simulated data



- From X_0:
  - Sobol-CPI(1)$\simeq$ Sobol-CPI(100)$\simeq$ CFI $\simeq$ scSAGEj $\simeq$ LOCO(W).
  - cSAGEvf $\simeq$ LOCI.
- From X_2:
  - cSAGEvf, cSAGE and LOCI do not satisfy the minimal axiom!

Real data

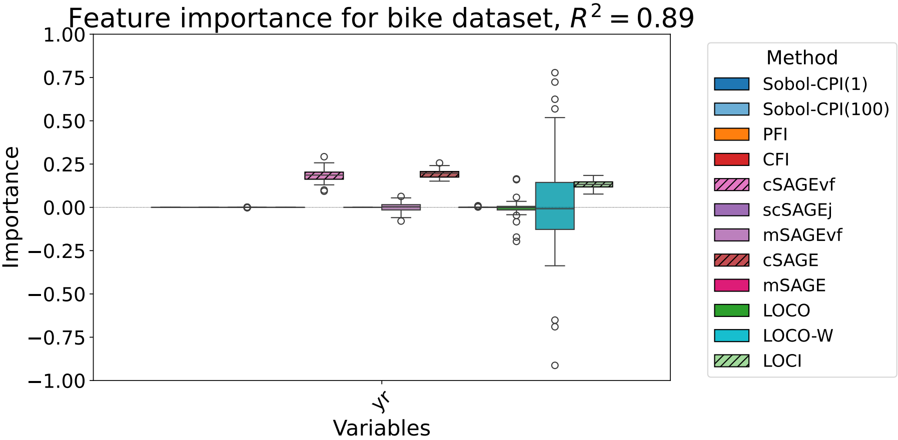


Figure 1: **Boxplots of the VIMs for the feature *year*:** Methods satisfying the minimal axiom assign no importance to this variable.

Real data

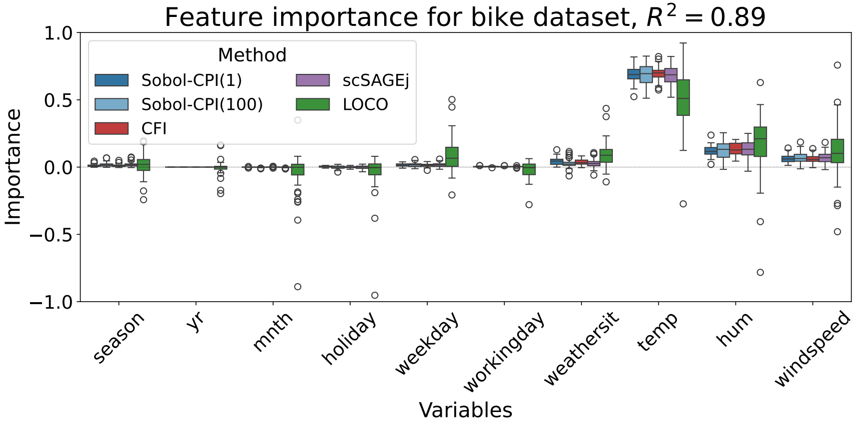


Figure 2: **Boxplots of the VIMs estimating ψ_{TSI} for all features:** Refitting approaches exhibit poorer inference properties.

Index

Conclusion

1. **How to compare VIMs?**

   Theoretical index → Estimation → Type-I error

   1. Conceptual comparison in the **theoretical index**.
   2. Inference comparison in the **estimation**.
   3. **Statistical guarantees** for the selected features.

   ✓ We provide a guide to help practitioners select a meaningful VIM.

2. **What's the minimum for a VIM?**

   - **Minimal axiom:** $\psi(j, P) = 0$ if and only if $X^j \perp\!\!\!\perp y \mid X^{-j}$.

   ✓ Intuitive: Important if its absence degrades the model.

   ✓ Link between **variable selection** and **variable importance**!

Index

References

Ian Covert, Scott M Lundberg, and Su-In Lee. Understanding global feature contributions with additive importance measures. *Advances in Neural Information Processing Systems*, 33:17212–17223, 2020.

Fiona Katharina Ewald, Ludwig Bothmann, Marvin N. Wright, Bernd Bischl, Giuseppe Casalicchio, and Gunnar König. *A Guide to Feature Importance Methods for Scientific Inference*, page 440–464. Springer Nature Switzerland, 2024. ISBN 9783031637971. doi: 10.1007/978-3-031-63797-1_22. URL http://dx.doi.org/10.1007/978-3-031-63797-1_22.

Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J. Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523): 1094–1111, 2018. doi: 10.1080/01621459.2017.1307116. URL https://doi.org/10.1080/01621459.2017.1307116.

References

Christoph Molnar. *Interpretable Machine Learning*. 3 edition, 2025. ISBN 978-3-911578-03-5. URL `https://christophm.github.io/interpretable-ml-book`.

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

Thank You, Questions?

Theoretical indices

Index	Definition	MA
ψ_{TSI}	$\mathbb{E}\left[\ell(m_{-j}(X^{-j}),y)\right]-\mathbb{E}[\ell(m(X),y)]$	Yes
ψ_{SAGE}	$\sum_{S\subset-\{j\}} w_S\left(\mathbb{E}\left[\ell(y,\mathbb{E}\left[m(X)\mid X^S\right])\right]-\mathbb{E}\left[\ell(y,\mathbb{E}\left[m(X)\mid X^{S\cup\{j\}}\right])\right]\right)$	No
ψ_{LOCI}	$\mathbb{E}[\ell(y,\mathbb{E}[y])]-\mathbb{E}\left[\ell(y,m_j(X^j))\right]$	No
ψ_{mSAGEvf}	$\mathbb{E}[\ell(y,\mathbb{E}[y])]-\mathbb{E}\left[\ell(y,\mathbb{E}\left[m(X^{(-j)})\right])\right]$	Yes
ψ_{mSAGE}	$\sum_{S\subset-\{j\}} w_S\left(\mathbb{E}\left[\ell(y,\mathbb{E}\left[m(X^{(-S)})\mid X^S\right])\right]-\mathbb{E}\left[\ell(y,\mathbb{E}\left[m(X^{(-S)})\mid X^{S\cup\{j\}}\right])\right]\right)$	Yes
ψ_{PFI}	$\mathbb{E}\left[\ell(m(X^{(j)}),y)\right]-\mathbb{E}[\ell(m(X),y)]$	Yes

Inference

Method	Theoretical quantity	Index	Estim
cSAGE	$\sum_{S \subset -\{j\}} w_S \left(v(S \cup \{j\}) - v(S)\right)$	ψ_{SAGE}	M
cSAGEvf	$v(\{j\})$	ψ_{LOCI}	M
mSAGEvf	$v^m(\{j\})$	ψ_{mSAGEvf}	M
mSAGE	$\sum_{S \subset -\{j\}} w_S \left(v^m(S \cup \{j\}) - v^m(S)\right)$	ψ_{mSAGE}	M
scSAGEvf	$v(-\{j\} \cup \{j\}) - v(-\{j\})$	ψ_{TSI}	M
LOCO	$\mathbb{E}\left[\ell(m_{-j}(X^{-j}), y)\right] - \mathbb{E}\left[\ell(m(X), y)\right]$	ψ_{TSI}	R
LOCO-W	$\mathbb{E}\left[\ell(m_{-j}(X^{-j}), y)\right] - \mathbb{E}\left[\ell(m(X), y)\right]$	ψ_{TSI}	R
LOCI	$\mathbb{E}\left[\ell(m_j(X^j), y)\right] - \mathbb{E}\left[\ell(m(X), y)\right]$	ψ_{LOCI}	R
PFI	$\mathbb{E}\left[\ell(m(X^{(j)}), y)\right] - \mathbb{E}\left[\ell(m(X), y)\right]$	ψ_{PFI}	P
CFI	$\mathbb{E}\left[\ell(m(\widetilde{X}^{(j)}), y)\right] - \mathbb{E}\left[\ell(m(X), y)\right]$	ψ_{TSI}	P
Sobol-CPI(n-cal)	$\frac{n_{\text{cal}}}{n_{\text{cal}}+1}\left(\mathbb{E}\left[\ell\left(\frac{1}{n_{\text{cal}}}\sum_{k=1}^{n_{\text{cal}}} m(\widetilde{X}_k^{(j)}), y\right)\right] - \mathbb{E}\left[\ell(m(X), y)\right]\right)$	ψ_{TSI}	P/M