A principled approach for comparing Variable Importance

Angel Reyero Lobo, Pierre Neuvial and Bertrand Thirion

Abstract. Variable importance measures (VIMs) aim to quantify the contribution of each input covariate to the predictability of a given output. With the growing interest in explainable AI, numerous VIMs have been proposed, many of which are heuristic in nature. This is often justified by the inherent subjectivity of the notion of importance. This raises important questions regarding usage: What makes a good VIM? How can we compare different VIMs?

In this paper, we address these questions by: (1) proposing an axiomatic framework that bridges the gap between variable importance and variable selection. This framework formalizes the intuitive principle that features providing no additional information should not be assigned importance. It helps avoid false positives due to spurious correlations, which can arise with popular methods such as Shapley values; and (2) introducing a general pipeline for constructing VIMs, which clarifies the objective of various VIMs and thus facilitates meaningful comparisons. This approach is natural in statistics, but the literature has diverged from it.

Finally, we provide an extensive set of examples to guide practitioners in selecting and estimating appropriate indices aligned with their specific goals and data.

Key words and phrases: Variable importance, Variable selection, Scientific discovery.

1. INTRODUCTION

Global variable importance aims to assign a measure of relevance to each feature with respect to a target. Since this relationship can be highly complex, machine learning (ML) models—proven effective at capturing such complexities in real-world scenarios—are often used as surrogates. These models enable the extraction of feature importance in intricate settings, such as genomic data, which are known for their high dimensionality and correlation structure. This is at the core of scientific discovery in data-driven approaches that are currently pervasive.

However, ML models are often opaque. While simple models like linear regression are fully interpretable through their coefficients, they rarely reflect the true data-generating process. Hence, there exists a trade-off between model complexity and interpretability (Molnar et al. (2022)). To mitigate this, there has been growing interest in *model-agnostic* variable importance measures (VIMs) (Williamson et al. (2023)).

Controlled variable selection has emerged as a method for filtering features with statistical guarantees (Candès et al. (2018)). Although variable selection and variable importance may appear to pursue the same goal, they have been treated separately in the literature. For instance, some variable importance measures aim to equitably distribute relevance across all features, while variable selection aims to identify a minimal subset of predictive variables (Bénard, Da Veiga and Scornet (2022)). One of the reasons is that some popular variable importance measures are based on complex combinatorial computations and game-theoretic axioms that make them hard to interpret and not well suited for this goal (Verdinelli and Wasserman (2024a)).

The most widely used importance measure is the Shapley value (Shapley (1953)), which, for global variable importance, is approximated by cSAGE (Covert, Lundberg and Lee (2020)). Figure 1 illustrates this approach in a simple linear setting with a single relevant covariate. No-

Angel Reyero Lobo is PhD candidate, Institut de Mathématiques de Toulouse and Inria Saclay, France (e-mail: angel.reyero-lobo@inria.fr). Pierre Neuvial is Research Director, Institut de Mathématiques de Toulouse, France (e-mail: pierre.neuvial@math.univ-toulouse.fr). Bertrand Thirion is Research Director, Inria Saclay, CEA, Université Paris Saclay, France (e-mail: bertrand.thirion@inria.fr).

tably, Shapley values assign nonzero importance to irrelevant yet correlated features. This potentially leads to false discoveries and questions its applicability to scientific discovery.

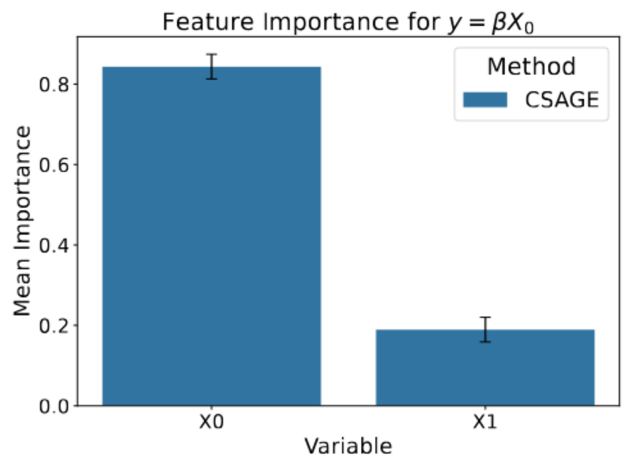

FIG 1. ***Bar plot of the estimated importance using SAGE:*** *The input consists of two Gaussian features correlated by* 0.6*, with a linear output* $y = \beta X_0$*. Left: importance of* X_0*; Right: importance of* X_1*. Note that Shapley values assigns nonzero importance to* X_1*, even though it does not appear in the true model. Importance was estimated using Gradient Boosting with* $n = 10{,}000$ *samples over 100 repetitions.*

Ideally, we want a definition of importance that considers the impact of missing a feature on predictive performance. This definition should align with the standard definitions of importance from Covert, Lundberg and Lee (2020); Molnar (2025). However, this view conflicts with current axiomatic frameworks (Shapley (1953); Verdinelli and Wasserman (2024a)). In response, we propose a **minimal axiom**: variables that do not provide any exclusive information should not be assigned importance. It bridges the gap between variable selection and importance by offering both statistical guarantees and rich information beyond binary decisions. It is minimal in the sense that it corresponds to a necessary condition for null importance, allowing for flexibility to accommodate subjective definitions of importance.

We conduct a comprehensive evaluation of state-of-the-art VIMs in the light of this axiom. Notably, Shapley values do not satisfy the proposed axiom, meaning that they assign importance to unimportant features, as seen in Figure 1. In contrast, perturbation-based indices do satisfy it. In particular, we argue that the commonly held belief—that Permutation Feature Importance (PFI, Mi et al. (2021)) is suited only for marginal importance (Ewald et al. (2024); Molnar (2025))—is misguided. Instead, we show that it aligns with conditional importance, and therefore the minimal axiom. This is done because **an implicit conditional filtering** emerges during the training of the model.

One substantial challenge in VIMs is the inconsistency across rankings produced by different methods, making it difficult to assess which features are truly important. Therefore, it is necessary to classify them to enable insightful comparisons. Existing categorizations mainly focus on the inference step—grouping methods by whether they involve refitting, marginalizing, or perturbing features (Ewald et al. (2024)). However, this overlooks the fact that the model is just a surrogate for the data-generating function, which is unknown in practice and must be estimated. In practice, it is not any model, but the minimizer of an empirical loss. Moreover, variable importance estimation often involves quantities like conditional distributions or expectations, which are often ignored in the theoretical analysis (Covert, Lundberg and Lee (2020); Hooker, Mentch and Zhou (2021); Ewald et al. (2024)). Furthermore, even methods that use the same estimation procedure can yield different rankings, as demonstrated in our experiments comparing PFI and CFI—both perturbation-based—as well as mSAGE and cSAGE—both marginalization-based.

To enable interpretable comparisons, we introduce a **principled approach for constructing VIMs**. This approach begins by **(1) explicitly defining a theoretical importance index** (which should satisfy the minimal axiom) that aligns with user objectives. Then, it focuses on **(2) estimating** this index, considering desirable inference properties such as robustness to model misspecification (Reyero Lobo, Neuvial and Thirion (2025)). Finally, to enable reliable scientific discovery, we emphasize **(3) recovering statistical guarantees**, such as type-I error control. We argue that the uncertainty of the model should therefore be accounted for only at the last two stages, not in the definition of the theoretical index itself.

Next, we study existing importance indices in the light of this three-step approach: we explain their interpretations, compare different estimation strategies, and outline the corresponding statistical guarantees to guide practitioners in real-world applications.

The main contributions of this paper are:

- to propose an interpretable, minimal axiom that unifies variable selection and importance.
- to introduce a principled approach for constructing VIMs and comparing them meaningfully.
- to apply the proposed methodology to an extensive set of VIMs, providing a new classification, distinct from recent presentation of the field.
- to illustrate using extensive numerical experiments with both real and synthetic datasets that the VIMs proven to satisfy the minimal axiom assign zero importance to the same null covariates, and the newly proposed classification is consistent with the obtained importance rankings.

1.1 Setting and Notation

Let $(X, y) \sim P \in \mathcal{M}$, where $X \in \mathcal{X} \subset \mathbb{R}^p$ is the input, $y \in \mathcal{Y} \subset \mathbb{R}$ is the output and the pair (X, y) is sampled from a distribution P belonging to a model class denoted $\mathcal{M}$. Given a set $S \subset [p]$, we denote by $-S = [p]\backslash S$ the subset $[p]$ restricted of S. If $S = \{j\}$, when it is clear from the context, j (resp. $-j$) denotes the subset $\{j\}$ (resp. $-\{j\}$). We denote by X^S the coordinates of X corresponding to the subset S. Similarly, we denote by X^j the j-th coordinate of X.

We define $\psi(j, P)$ as the importance index of the j-th covariate X^j for $j \in \{1, \ldots, p\}$. We denote by $\mathcal{F}$ an arbitrary space of functions from $\mathcal{X}$ to $\mathcal{Y}$. We denote by $m \in \mathcal{F}$ the conditional expectation $m(X) = \mathbb{E}[y \mid X]$.

We emphasize that no prior assumption is made on the function space $\mathcal{F}$. Thus, there is complete freedom in choosing the function $m \in \mathcal{F}$. However, in practice, inference challenges may arise when approximating m using a machine learning model $\widehat{m}$, motivating the use of a model-agnostic approach to accommodate complex data-generating processes. We denote by $\widehat{m}_n$ the estimator that makes explicit the dependence on the training sample size n. This estimation issue is further discussed in Section 4.

We also make the identifiability assumption that no input covariate is an exact function of the others. This ensures that the problem of assigning importance is well-defined—a standard assumption in the variable importance literature (see Candès et al. (2018); Verdinelli and Wasserman (2024a)). If necessary, this can be achieved by prior filtering of repeated features or a choice of a representative of the group (Meinshausen (2013)). Under this assumption, the set of important covariates

$$\mathcal{S} := \left\{ j \in \{1, \ldots, p\} \,\middle|\, X^j \not\perp\!\!\!\perp y \mid X^{-j} \right\}$$

is well defined (Candès et al. (2018)). This is known as the Markov blanket of y and it is the objective of variable selection.

In Section 2, we revisit the axiomatic approaches proposed in the literature. In Section 3, we introduce our new interpretable minimal axiom, designed to close the gap between variable importance and variable selection. Section 4 develops our principled framework for constructing and comparing Variable Importance Measures (VIMs). In Section 5, we apply this framework to state-of-the-art VIMs and discuss whether they satisfy the minimal axiom, with a summary provided in Section 6. Finally, Section 7 presents simulations and real data analyses illustrating our findings.

2. RELATED WORK

The notion of importance is inherently subjective, and formalizing what is desirable from a mathematical perspective can be challenging.

In this section, we begin by discussing the main axiomatic foundations that have been proposed in the literature. In Section 2.1, we review the Shapley value axioms, and in Section 2.2, we present the critique put forward by Verdinelli and Wasserman (2024a) along with their newly proposed axioms.

2.1 Shapley values axioms

Shapley (1953) introduced an axiomatic framework for the fair allocation of a total payout among players in a cooperative game, where each player's share is determined by their contributions across all possible subsets of players. Building on this idea, Covert, Lundberg and Lee (2020) applied the same framework to variable importance, interpreting the features as "players" and the "game" as the prediction task, where the value corresponds to the model's predictive performance.

Let $V : \mathcal{P}([p]) \to \mathbb{R}$ be a value function, where $\mathcal{P}$ denotes the power set operator and $S \in \mathcal{P}([p])$ represents a subset of indices. This value function assigns a measure of predictive power to each feature set S. It consists of the building blocks used to define the importance measure ψ, which aggregates importance across all subsets. For instance, Covert, Lundberg and Lee (2020) propose defining $V(S)$ as the difference in the model loss when using the covariates in S versus using no covariates (i.e., predicting by the mean).

The classical Shapley axioms are as follows:

AXIOM 1 (Efficiency). *The total importance is fully distributed among all features:* $\sum_j \psi(j, P) = V([p])$.

AXIOM 2 (Symmetry). *If two features j and k contribute equally to every subset, i.e.,* $V(S \cup \{j\}) = V(S \cup \{k\})$ *for all S, then* $\psi(j, P) = \psi(k, P)$.

AXIOM 3 (Dummy). *If a feature j does not contribute to any subset, i.e.,* $V(S \cup \{j\}) = V(S)$ *for all S, then* $\psi(j, P) = 0$.

AXIOM 4 (Linearity). *If two value functions V and V' yield values $\psi(j, P)$ and $\psi'(j, P)$ respectively, then the value of $V + V'$ is* $\psi(j, P) + \psi'(j, P)$.

This framework led to the now well-known Shapley values, which are the unique fulfilling this axiomatic framework. A common criticism of Shapley values is their computational burden due to the need to evaluate all possible feature subsets. However, in this work, we focus on conceptual critiques rather than computational concerns.

Verdinelli and Wasserman (2024a) criticized Shapley values from both inferential and conceptual perspectives. Inferentially, many combinations of features may lead to poor estimation, making it questionable to rely on all possible subsets. Conceptually, due to the complex weighted combinations, these values are difficult to interpret, as their numerical magnitudes cannot be directly associated with a tangible property, such as the percentage of explained variance or the effect of a feature within the model.

Moreover, the efficiency axiom (1) imposes an additivity constraint, which may be undesirable when variables are correlated or interact non-linearly (Kumar et al. (2020)). Such assumptions may distort the attribution of importance.

Due to the strict nature of the dummy axiom (3), when using the value function proposed by Covert, Lundberg and Lee (2020), a feature is assigned zero importance only in the very specific case where it is completely independent of both the target and all other features. Hence, Shapley values may assign nonzero importance to irrelevant covariates, as seen in Figure 1. This undermines their use in feature selection, and therefore scientific discovery.

2.2 Verdinelli & Wasserman axioms

Verdinelli and Wasserman (2024a) tackled the issue of *correlation distortion*—the phenomenon where the importance of a feature is underestimated due to its correlation with other inputs. They proposed a new axiomatic foundation for regression settings, explicitly aligning with the interpretation of importance as predictive power and aiming to mitigate the effects of correlation:

AXIOM 5 (Functional Dependence). *$\psi(j, P) = 0$ if and only if $m(X)$ is not a function of X^j.*

AXIOM 6 (Correlation-Free). *$\psi(j, P) = \psi(j, p(y \mid X)\, p_j(X^j)\, p_{-j}(X^{-j}))$, where p_j and p_{-j} denote the marginals of X^j and X^{-j}, respectively.*

AXIOM 7 (Linear Agreement). *If $m(X) = \sum_j \beta_j X^j$, then $\psi(j, P) = \beta_j^2$.*

The first axiom is intuitive: a feature is important if the prediction function $m(X) (:= \mathbb{E}[y \mid X])$ depends on it. The second axiom aims to eliminate correlation distortion by requiring that importance depends only on the marginals of X^j and X^{-j}, and not on their joint dependence structure. The third axiom ensures agreement with the coefficients in a linear model, grounding the measure in familiar settings.

In Section 3, we propose a generalization of the functional dependence axiom and argue that the correlation-free axiom may not be universally appropriate, as its necessity depends on one's subjective view on what variable importance should capture.

Additionally, Verdinelli and Wasserman (2024a) introduced several more loosely defined axioms motivated by computational and inferential considerations. However, these are not fully satisfied by their proposed VIM, a point we further discuss in Section 5.3.

3. AXIOMATIC FRAMEWORK FOR VARIABLE IMPORTANCE MEASURES

The notion of variable importance is inherently subjective. Consequently, an axiomatic framework—such as those previously proposed—can be overly restrictive or tailored to specific goals, making it difficult to generalize. For instance, in predicting house prices, features like the number of rooms, surface area, and neighborhood may exhibit complex interdependencies. While the first two may be correlated and suffer from correlation distortion, one could argue that either is sufficient on its own, or that the neighborhood offers specific information not captured by the others. This highlights that axioms like the correlation-free requirement (Axiom 6) may not be universally desirable. Note that this remark no longer applies in very high-dimensional settings, where extreme correlations can make the notion of individual feature importance ill-defined, and grouping features may become preferable (e.g. Chamma, Engemann and Thirion (2024)).

However, there is a consensus in the literature that feature importance should be at least related to a feature's predictive power Covert, Lundberg and Lee (2020); Verdinelli and Wasserman (2024a); Ewald et al. (2024). Notably, Covert, Lundberg and Lee (2020) explicitly define variable importance as follows:

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

In this section, we introduce a *minimal axiom*, which formalizes this intuitive and widely accepted notion. We propose it as a minimal requirement for any variable importance measure (VIM). This axiom captures the idea that a feature should be considered important only if it provides unique predictive information—i.e., if its absence degrades the model's performance.

AXIOM 8 (Minimal Axiom). *$\psi(j, P) = 0$ if and only if $X^j \perp\!\!\!\perp y \mid X^{-j}$.*

This axiom aligns with the goal of identifying features that are conditionally informative for prediction. Under standard conditions, it relates to the functional dependence axiom (5). For instance, when performing regression, it is standard to assume an additive, independent, centered noise:

ASSUMPTION 1a (Additive noise). *$y = m(X) + \epsilon$ with $\epsilon \perp\!\!\!\perp X$ and $\mathbb{E}[\epsilon] = 0$, for some $m \in \mathcal{F}$.*

Note that m is indeed the conditional expectation function, since $\mathbb{E}[y \mid X] = \mathbb{E}[m(X) + \epsilon \mid X] = m(X)$.

Similarly, in classification, we assume a model on the probability of belonging to the class:

ASSUMPTION 1b (Classification). $\mathbb{P}(y = 1 \mid X) = \sigma(m(X))$ *for some* $m \in \mathcal{F}$ *and* $\sigma : \mathbb{R} \to [0, 1]$ *bijective.*

For example, taking the logistic function as σ shows that it generalizes the logistic regression model.

Under either of these assumptions, the conditional dependence in the minimal axiom is related with the functional dependence:

PROPOSITION 3.1 (Conditional and Functional Independence). *Under Assumption 1a or 1b,* $X^j \perp\!\!\!\perp y \mid X^{-j}$ *if and only if* $m(X)$ *is not a function of* X^j.

These assumptions are necessary to obtain an equivalence. For instance, in regression, if Assumption 1a is violated, functional independence does not imply conditional independence. This is demonstrated by a counterexample from Ewald et al. (2024), where $y \sim \mathcal{N}(X^1, X^2)$ and $X^1 \perp\!\!\!\perp X^2$. In this case, $\mathbb{E}[y \mid X] = X^1$, implying that y is functionally independent of X^2. However, $y \not\perp\!\!\!\perp X^2 \mid X^1$, so conditional independence does not hold. Nevertheless, Assumptions 1a and 1b are both standard and sufficiently general, and any statistical selection procedure involving models implicitly relies on these assumptions. In what follows, we assume that the equivalence in Proposition 3.1 holds.

This minimal axiom is thus more general and helps bridge the gap between variable importance and variable selection—two concepts often treated separately in the literature. The former typically distributes predictive contribution heuristically, while the latter focuses on identifying a minimal subset of predictive features (Bénard, Da Veiga and Scornet (2022)). By unifying both under a shared statistical foundation, the minimal axiom allows these fields to benefit from one another. On the one hand, *controlled* variable selection offers rigorous statistical guarantees, such as control over Type-I error or the False Discovery Rate (Candès et al. (2018); Tansey et al. (2022)). On the other hand, variable importance provides more nuanced information than a simple yes/no decision. Both aspects are essential for advancing scientific understanding, as they enable interpretable, statistically sound conclusions about the roles of different variables.

Importantly, this approach is flexible enough to accommodate various interpretations of variable importance, depending on the inferential or practical objective. This will be illustrated in Section 5, where we examine several widely used VIMs and assess whether they satisfy the minimal axiom.

4. PRINCIPLED APPROACH TO DEFINE A VIM

4.1 Pitfalls of inference-based classification of VIM

Historically, many variable importance measures have emerged as heuristics for interpreting black-box models. As such, they often focus on quantifying the effects of perturbations or modifications to the information provided by a specific feature. However, since many of these measures are closely tied to particular estimation procedures, the literature has introduced several ad hoc categorizations that can obscure the core statistical objective. For example, it is common to distinguish between VIMs that rely on a single model via perturbations and those that require refitting new models. Some claim that such measures are inherently incomparable. For instance, Molnar (2025) stated that Leave-One-Covariate-Out (LOCO, Lei et al. (2018)) and PFI (Mi et al. (2021)) cannot be compared:

> *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.*

Similarly, they claimed that conditional PFI (CFI, Hooker, Mentch and Zhou (2021)) and LOCO are not directly comparable:

> *But since conditional PFI and LOCO work differently, they differ in their interpretations. Conditional PFI is an interpretation that only involves the model at hand. LOCO importance focuses more on the machine learning algorithm, since it involves retraining the model, and the interpretation now involves multiple models trained differently.*

Yet, this perspective overlooks a crucial point: the models used are surrogates for the true data-generating process, and most variable importance measures ultimately aim to estimate some underlying population-level index. For example, in (5) in Section 5.3, we show that conditional PFI coincides with the LOCO asymptotically. We propose reframing the development of VIMs by first focusing on the theoretical index they intend to estimate. That is, instead of starting from a model-specific heuristic, we advocate for a general structured approach for generating and interpreting VIMs.

4.2 Proposed approach

Our proposed method consists of three steps: first, defining the theoretical index; then, estimating it; and finally, providing statistical guarantees for the important covariates.

The starting point in this pipeline is the **theoretical index** itself, as it encodes the desired interpretation of *importance*. By choosing a meaningful and interpretable population quantity, one can ensure properties such as how correlation with other variables influences a feature's importance. It is at this stage that one should explicitly decide which axioms the VIM should satisfy, acknowledging the subjectivity inherent to the notion of importance. Indeed, we do not expect a single index to suit all use cases—this aligns with the "Assuming One-Fits-All Interpretability" pitfall identified by Molnar et al. (2022). This choice will yield an insightful ranking aligned with the specifically defined goals. Moreover, our minimal axiom should be verified at this stage, as it represents a minimal requirement for this theoretical quantity.

Given that the data-generating process is unknown in practice, the next step is to **estimate** the theoretical index. This step can involve a variety of estimation procedures, each with distinct statistical and computational properties. Some estimators aim for nonparametric efficiency (Williamson et al. (2023)), others focus on reducing variance (Paillard et al. (2025)), achieving robustness to model misspecification (Reyero Lobo, Neuvial and Thirion (2025)), or ensuring computational feasibility (Verdinelli and Wasserman (2024a)). For example, as we will discuss in the context of the Total Sobol Index, one may estimate the index by refitting multiple models or, alternatively, by estimating the input conditional distribution. The choice of estimation method may depend on the application—for instance, in scenarios with abundant unlabeled data, one might assume complex relationships between features and the response, but simpler relationships among features themselves, as in model-X methods (Candès et al. (2018)). Moreover, since many estimation procedures involve models, alternative approaches can be considered to account for multiple models, such as the Rashomon set Donnelly et al. (2023).

Because the estimation step introduces approximation error, the final estimator may not perfectly reflect the theoretical index. Consequently, to support reliable scientific discovery, it is crucial to provide **statistical guarantees** for the selected features. Depending on the context, these guarantees might take the form of asymptotic (Williamson et al. (2023); Reyero Lobo, Neuvial and Thirion (2025)) or finite-sample (Liu et al. (2021)) type-I error control, or False Discovery Rate control when conducting multiple testing (Candès et al. (2018)). This final step is important for ensuring reliable discoveries. Indeed, Tansey et al. (2022) presented real-world examples where statistically grounded selected features correlated poorly with the rankings given by heuristic importance scores.

5. APPLICATION OF THE FRAMEWORK

The framework defined in Section 4 is intended to guide practitioners not only in selecting an index aligned with their goals, but also in choosing an estimation procedure that balances statistical robustness, computational constraints, and inferential guarantees. In this section, we provide an extensive study of **theoretical indices**, discuss their interpretability and whether they satisfy the minimal axiom, and examine different **estimation** strategies adapted to the data at hand along with their **statistical properties**. This principled approach sheds some light on the interpretation of classical indices.

In particular, we argue that the commonly held belief—that Permutation Feature Importance (PFI, Mi et al. (2021)) is suited only for marginal importance (Ewald et al. (2024); Molnar (2025))—is misguided. Instead, we show that PFI aligns with conditional importance, and therefore the minimal axiom. This common misunderstanding arises from a lack of clarity about the null hypothesis underlying PFI. It is also caused by notation, especially since a conditional counterpart exists (Hooker, Mentch and Zhou, 2021). Additional notation-related misunderstandings will be discussed below, where we show that the marginal versions of SAGE are in fact suitable for conditional importance, while the conditional SAGE, which is the standard Shapley values, is not—contrary to previous claims Ewald et al. (2024).

All the proofs can be found in Appendix C.

5.1 Coefficients of a generalized linear model

The first, most intuitive importance index relies on the coefficients of a linear model. It is simple to interpret because the effect of each variable can be directly inferred from the magnitude of its coefficient. Moreover, the same reasoning applies to logistic regression and, more broadly, to generalized linear models (GLMs). These models are often favored for their interpretability (Rudin (2019); Molnar et al. (2022)) and have been advocated over black-box models, especially in sensitive applications such as hate speech detection (Reyero Lobo et al. (2023)). In this section, we argue that this simple variable importance measure, along with some modifications, satisfies the minimal axiom.

We begin by stating the standard generalized linear model (GLM) assumption:

ASSUMPTION 2 (GLM assumption). *Given a link function g, we assume*

$$g\left(\mathbb{E}[y \mid X]\right) = X^1\beta_1 + \ldots + X^p\beta_p.$$

Under this assumption, as introduced in McCullagh (1989), a natural variable importance measure for covariate j is given by the magnitude of the coefficient. Thus, the **theoretical index** is given by:

DEFINITION 5.1 (GLM indices). Given $j \in [p], P \in \mathcal{M}$, $\psi_{\mathrm{GLM}}(j, P)$ is defined as

$$\psi_{\mathrm{GLM}}(j, P) := \beta_j^2.$$

The square is used to ignore whether the effect is positive or negative; what matters is simply that there is an effect. Also, we implicitly assume that the input features have been standardized. This measure, clearly aligned with Axiom 7, can be useful in some settings due to its simplicity and the direct interpretability of the coefficients, which represent the contribution of each covariate to the predictive function. This simple VIM satisfies the minimal axiom:

PROPOSITION 5.2. *Under the GLM assumption (2),* ψ_{GLM} *satisfies the minimal axiom.*

However, even if this minimal axiom is fulfilled for this importance index, in practice there may be **estimation** complications due to high dimensionality and/or collinearity between input variables. These are issues of inference rather than interpretation problems with the theoretical index. For this reason, many penalization strategies have been introduced, such as the Akaike Information Criterion (AIC, Bozdogan (1987)) or the ℓ_1 penalization used in the Lasso (Tibshirani (1996)). We observe that all these strategies still satisfy the minimal axiom, as the theoretical index continues to assign zero importance to unimportant coordinates. Moreover, there are numerous results on bounding the number of false discoveries made by the inference procedures used in these methods (see, for example, Corollary 5.3 from Giraud (2021), which provide bounds in probability).

Many variable selection procedures with their respective **statistical guarantees** can be derived from this variable importance index. For instance, to control the type-I error using Conditional Randomization Tests (CRT, Candès et al. (2018)), the most effective procedures are based on the difference of coefficients from a Lasso model trained on the original data versus conditional independent data (Liu et al. (2021)). One can easily observe that the sum over the conditional samplings tends to the Lasso index, as the coefficients of the conditional independent generated input coordinates will tend to zero. Therefore, the minimal axiom of the given theoretical index is satisfied with finite sample guarantees on the selected coordinates. Similarly, this approach can be applied to the Knockoffs framework with the popular Lasso Coefficient Difference (LCD, Candès et al. (2018)), which provides False Discovery Rate control on the selected set. Finally, additional guarantees on p-values or on the familywise error rate can be obtained from corrected versions of the Lasso or Ridge (Bühlmann (2013); Dezeure et al. (2015)).

5.2 Perturbation indices

In the general case, assuming a linear model is often unrealistic, and we may want to adapt to more complex settings. Moreover, if the model does not generalize well, there may be no reliable scientific inference, as the model acts merely as a surrogate for the data-generating process. Thus, we are unable to provide interesting insights on feature importance with a model that does not represent the underlying distribution. This relates to the poor model generalization pitfall described by Molnar et al. (2022). Therefore, there is a need for model-agnostic theoretical indices. In the sequel, we present model-agnostic VIMs that use the loss to measure the predictability of the model when incorporating the information of the j-th coordinate versus when it is not used. Consequently, a data-splitting step is required, which may decrease the power of the procedure.

The first approach we present involves reusing the same model while perturbing the inputs to observe the effect of each coordinate on the model's output. Specifically, the model's performance with the j-th coordinate perturbed is compared to its performance with the original input. They appear to fall outside the principled proposed approach. Indeed, they are often presented as heuristics for model interpretation rather than as theoretically grounded indices with a well-defined estimand. It is unclear what is estimated, and therefore, the corresponding null hypothesis. To tackle this, we proceed with the proposed principled approach.

We begin by defining the **theoretical index**:

DEFINITION 5.3 (Perturbation indices). Given $j \in [p], P \in \mathcal{M}$, $\psi_{\text{perturb}}(j,P)$ is defined as

$$\psi_{\text{perturb}}(j,P) := \mathbb{E}\left[\ell(y, m(\widetilde{X}^j))\right] - \mathbb{E}\left[\ell(y, m(X))\right],$$

where $\widetilde{X}^j$ is a perturbed version of X that preserves the $-j$ coordinates (i.e., $\widetilde{X}^{j-j} = X^{-j}$), but the j-th coordinate is *perturbed*, and therefore the joint distribution changes, i.e., $\mathcal{L}(X,y) \neq \mathcal{L}(\widetilde{X}^j, y)$.

Therefore, all the coordinates are preserved except for the j-th coordinate, which is sampled in a way that alters the joint distribution, making it different from the original one.

Now, we present several examples of perturbations along with their corresponding **estimation** procedures, showing, in particular, that for any perturbation, the minimal axiom is satisfied because the model acts as a first filter for conditional independence.

The most popular perturbation-based index is the permutation feature importance (PFI, Breiman (2001); Mi et al. (2021)) due to its simple estimation. This approach relies on a perturbation denoted by $X^{(j)}$ such that: the j-th coordinate preserves the marginal distribution ($\mathcal{L}(X^{(j)j}) = \mathcal{L}(X^j)$), and it is independently distributed ($X^{(j)j} \perp\!\!\!\perp X^{-j}$ and $X^{(j)j} \perp\!\!\!\perp y$). However, the theoretical index associated to PFI was studied by Bénard, Da Veiga and Scornet (2022), who showed that it is not always desirable, as it can lead to misleading conclusions in the presence of correlation. This reinforces our proposal to study the theoretical index carefully, since otherwise, we risk misunderstanding what is considered important.

Regarding its estimation, the simplest method consists of permuting the value across observations. However, this approach also suffers from inference issues because it induces extrapolation (Hooker, Mentch and Zhou (2021)). Indeed, due to permutation, the model attempts to predict in low-density regions where it was not trained, resulting in unpredictable behavior.

To address this estimation issue, Hooker, Mentch and Zhou (2021) and Chamma, Engemann and Thirion (2023) proposed permuting the j-th coordinate conditionally on the other ones, leading to the Conditional Permutation Importance (CPI). This method was primarily introduced to tackle the inferential problem, without thoroughly studying the theoretical quantity to which it converges. This theoretical aspect was analyzed in Reyero Lobo, Neuvial and Thirion (2025), where it was shown that the conditional permutation step corresponds to sampling from the conditional distribution; then, CPI corresponds to Conditional Feature Importance (CFI, Strobl et al. (2008)). We denote this conditional perturbation by $\widetilde{X}^{(j)}$. CFI assumes that the j-th coordinate is independent of the output given the other inputs, i.e., $\widetilde{X}^{(j)} \perp\!\!\!\perp y \mid X^{-j}$, and preserves the conditional distribution, $\mathcal{L}(X^j \mid X^{-j}) = \mathcal{L}(\widetilde{X}^{(j)j} \mid X^{-j})$. Moreover, Reyero Lobo, Neuvial and Thirion (2025) showed that for the quadratic loss, CFI coincides up to a explicit universal constant with the Total Sobol Index, presented below.

Other types of perturbations can also be considered, such as relational perturbations (König et al. (2021)), which operate similarly to the CFI but condition on a smaller subset of coordinates. However, each perturbation gives rise to a different theoretical index, and thus it is not meaningful to compare them directly, as they pursue different objectives and consequently yield different rankings of importance.

In any case, these indices satisfy the minimal axiom for any kind of perturbation defined according to Definition 5.3. In addition, under the following assumption from Reyero Lobo, Neuvial and Thirion (2025), the estimated index is also able to correctly identify the null coordinates.

ASSUMPTION 3 (Asymptotic relevance). *Denote by $g_j(x, s)$ the vector obtained from x by replacing the j-th component of x by $s \in \mathbb{R}$. For $\epsilon > 0, x \in \mathcal{X}, s \in \mathbb{R}$ and $X^j \perp\!\!\!\perp y \mid X^{-j}$, there exists an n large enough such that*

$$|\widehat{m}_n(x) - \widehat{m}_n(g_j(x, s))| \leq \epsilon \text{ a.s.}$$

We observe that this assumption explicitly requires the model to be independent of the null coordinates, and it can be obtained under standard assumptions on the model. In particular, there is no extrapolation under the null hypothesis (see Reyero Lobo, Neuvial and Thirion (2025)).

PROPOSITION 5.4 (ψ_{perturb} satisfies the minimal axiom). *Under additive noise assumption (1a) or classification assumption (1b), ψ_{perturb} satisfies the minimal axiom for any strictly convex loss ℓ. Moreover, under asymptotic relevance assumption (3), if $X^j \perp\!\!\!\perp y \mid X^{-j}$, then $\psi_{\text{perturb}} \to 0$ almost surely for any continuous loss ℓ.*

The first part of the proposition indicates, in particular, that PFI is also suitable for detecting conditional independence. This result contrasts with the existing literature, where there is a common misconception: since the perturbation in PFI is marginal—also with respect to the rest of the input—it is often assumed that PFI is only suited for detecting marginal independence, not conditional independence (see Ewald et al. (2024)). In general, Ewald et al. (2024) first showed that if a method is suited for marginal independence, then it cannot be suited for conditional independence. Then, under strong assumptions, they related PFI to marginal independence and therefore claimed that it could not be used for conditional independence. In Appendix C.2.1, we prove that, under the assumptions of Ewald et al. (2024), marginal independence does in fact imply conditional independence.

It is important to note that this key result comes from a form of equivalence between functional and conditional independence (see Proposition 3.1). Consequently, the trained model $\widehat{m}$ is not just any model; it is a loss minimizer that approximates the data-generating process and thus serves as an initial filter for conditional independence.

Regarding the second part of Proposition 5.4, for inference purposes, when using a consistent estimator, the estimated importance of irrelevant coordinates vanishes asymptotically. This follows from Theorem 3.4 in Reyero Lobo, Neuvial and Thirion (2025), which also states that when using CFI—that is, when the perturbation is conditional—there is a double robustness property. Specifically, it suffices for only one of the models involved to be consistently estimated in order to detect the null hypothesis, which provides a valuable inference guarantee.

Finally, regarding the **statistical properties** of such methods, the extrapolation inherent to PFI makes it challenging to derive theoretical guarantees. While some central limit theorems have been established, allowing for the derivation of asymptotic type-I error control, these results are not model-agnostic (Föge et al. (2024)). Some asymptotic results for the CFI on the type-I error have been established by Reyero Lobo, Neuvial and Thirion (2025) based on Markov inequality using the vanishing influence function. This result could be extended to any perturbation index. Moreover, similarly to the case of Lasso coefficients, certain CRT-based methods can be adapted to provide finite-sample type-I error control, for example by using the Holdout Randomization Test from Tansey et al. (2022). Following the ideas of Watson and Wright (2021), other nonparametric tests such as the sign test or the Wilcoxon test can also be considered to obtain finite-sample type-I error control. The null hypothesis for such tests is that the mean of the original loss distribution coincides with that of the perturbed one.

5.3 Generalized Total Sobol Index

In this section, we begin by defining the theoretical index and then present several equivalent reformulations, each leading to a distinct plug-in estimator. These representations contrast with the previous categorizations of VIMs, as the index can alternatively be expressed through variance, model refitting, perturbation, or marginalization procedures. Finally, we discuss the associated statistical inference.

The Total Sobol Index (TSI) was originally introduced in the context of sensitivity analysis by Homma and Saltelli (1996) as the proportion of output variance attributed to a specific input coordinate when the remaining inputs are known. This notion can be extended beyond variance to accommodate any loss function ℓ (Williamson et al. (2023)). The **theoretical index** is given by:

DEFINITION 5.5 (Generalized Total Sobol Index (TSI)). Given $j \in [p], P \in \mathcal{M}$ and a loss function ℓ, the Generalized Total Sobol Index is defined as

$$\psi_{\mathrm{TSI}}(j, P) := \mathbb{E}\left[\ell\left(m_{-j}(X^{-j}), y\right)\right] - \mathbb{E}\left[\ell(m(X), y)\right],$$

where $m_{-j}(X^{-j}) := \mathbb{E}\left[y \mid X^{-j}\right]$ and $m(X) := \mathbb{E}\left[y \mid X\right]$.

This index is a widely used measure of variable importance (see, e.g., Lei et al. (2018); Rinaldo, Wasserman and G'Sell (2019); Hooker, Mentch and Zhou (2021); Bénard, Da Veiga and Scornet (2022); Williamson et al. (2023)). It is termed *generalized* because, when the loss function ℓ is the squared error, the expression recovers the unnormalized TSI definition. Notably, with the quadratic loss, TSI can also be rewritten in several equivalent forms:

(1)
$$\psi_{\mathrm{TSI}} = \mathbb{E}\left[\mathrm{Var}(y \mid X^{-j})\right]$$

(2)
$$= \underbrace{\mathbb{E}\left[(m_{-j}(X^{-j}) - m(X))^2\right]}_{variance}$$

(3)
$$= \underbrace{\mathbb{E}\left[\left(m_{-j}(X^{-j}) - y\right)^2\right] - \mathbb{E}\left[(m(X) - y)^2\right]}_{loss/refitting}$$

(4)
$$= \underbrace{\mathbb{E}\left[\left(\mathbb{E}\left[m(X) \mid X^{-j}\right] - y\right)^2\right] - \mathbb{E}\left[(m(X) - y)^2\right]}_{marginalization}$$

(5)
$$= \underbrace{\frac{1}{2}\left[\mathbb{E}\left[\left(m(\widetilde{X}^{(j)}) - y\right)^2\right] - \mathbb{E}\left[(m(X) - y)^2\right]\right]}_{perturbation}$$

(6)
$$= \sigma^2(R^2_{-j} - R^2).$$

From these equalities, we first note the richness of the Total Sobol Index (TSI): it admits multiple interpretations and computational forms. This highlights the value of the general estimation pipeline—despite differences in how the index is computed, each formulation targets the same underlying quantity. Yet, the current literature often categorizes these methods into distinct, seemingly incompatible families, claiming that they are not comparable (see Ewald et al. (2024); Molnar (2025); Fumagalli et al. (2025)).

We observe that it can be **estimated** as a direct plug-in using any of these equivalent formulations. First, from (1) to (3), we observe that the distinction between *variance*-based and *loss*-based approaches is artificial. Although variance-based methods do not explicitly use the output y in estimation (2), they nonetheless target the same quantity as loss-based methods based on (3), because the dependence in y is hidden in the models m and m_{-j}.

Equation (3) presents the Total Sobol Index using the squared loss based on two different models m and m_{-j}. Equation (4) corresponds to a marginalization-based approach, similar to that of Covert, Lundberg and Lee (2020). The plug-in in this formulation is often labeled a *one-function* variable importance measure (VIM), since the same estimated model m is used on both sides of the comparison, without retraining a separate m_{-j} for each j, which provides an advantage in terms of computation.

However, this perspective overlooks a critical challenge: the conditional expectation $\mathbb{E}[m(X) \mid X^{-j}]$ is generally unknown and must itself be estimated. This requires both estimating the conditional distribution and the conditional expectation. This estimation framework was formalized in the Sobol-CPI method proposed by Reyero Lobo, Neuvial and Thirion (2025), who also demonstrated that not having an infinite number of conditional samples introduces bias.

Equation (5) falls under the class of *perturbation methods*, since it uses a fixed model m and perturbs the j-th coordinate by drawing from the conditional distribution—a normalized version of the Conditional Feature Importance (CFI) (Strobl et al. (2008)). From (6), we note that this index can also be interpreted as the difference in nonparametric R^2 (Williamson et al., 2021).

Taken together, these results reveal that the usual categorization into refitting, perturbing, and marginalizing is misleading: in this example, all approaches aim to estimate the same functional. As such, they share the same interpretation and comparing them is meaningful. Their differences lie in inference properties rather than conceptual targets. For example, Williamson et al. (2021) studied the difference between plug-in estimators used in (2) and (3), showing that the former requires a one-step correction to achieve nonparametric efficiency, while the latter does not. The method in (3) corresponds to Leave-One-Covariate-Out (LOCO). Furthermore, the Permute-and-Relearn and Condition-and-Relearn importances of Hooker, Mentch and Zhou (2021), which consist of either marginally or conditionally permuting a feature before relearning the model, also fall within this category, since they aim to estimate

$$\mathbb{E}\left[y \mid X^{-j}, \widetilde{X}^j\right] = \mathbb{E}\left[y \mid X^{-j}\right] = m_{-j}(X^{-j}).$$

If a plug-in method is applied to (4), the resulting method is known as Sobol-CPI (Reyero Lobo, Neuvial and Thirion (2025)), which was shown to be nonparametrically efficient. It also corresponds to the conditional SAGE value function (Covert, Lundberg and Lee (2020)). Using the estimator in (5) leads to Sobol-CPI(1) (Reyero Lobo, Neuvial and Thirion (2025)), also known as $0.5{\times}$CFI (Strobl et al., 2008). Importantly, Reyero Lobo, Neuvial and Thirion (2025) proved that this estimator is *double robust*, meaning that it can reliably identify null covariates as long as either the predictive model or the conditional sampler is well-specified—thereby reducing the risk of false positives and making it a strong candidate for scientific discovery.

Another insightful interpretation of the Total Sobol Index arises when using the cross-entropy loss. In this case, it becomes directly connected to information-theoretic quantities such as mutual information (denoted by I) and the Kullback–Leibler (KL) divergence, as discussed by Covert, Lundberg and Lee (2020). Specifically, we have

$$\begin{aligned}(7)\quad \psi_{\mathrm{TSI}}(j,P) &= \mathrm{I}(y; X^j \mid X^{-j}) \\ &= D_{\mathrm{KL}}\left(P_{y,X^j|X^{-j}} \,\middle\|\, P_{y|X^{-j}} P_{X^j|X^{-j}}\right),\end{aligned}$$

which quantifies the conditional mutual information between y and X^j given X^{-j}. Intuitively, this measures the reduction in uncertainty about y obtained by adding X^j to the already known covariates X^{-j}.

Finally, we note that the Total Sobol Index satisfies the minimal axiom.

PROPOSITION 5.6. *Let ℓ be a loss with a unique minimizer which is a function of the conditional distribution of y given X. Under additive noise assumption (1a) or classification assumption (1b), then, $\psi_{\mathrm{TSI}}(j,P)$ satisfies the minimal axiom.*

As discussed in Williamson et al. (2023), most commonly used losses fulfill the condition stated in Proposition 5.6, with their Bayes-optimal predictors being functions of the conditional expectation. Examples include the mean squared error (MSE), deviance, classification accuracy, and the area under the ROC curve (AUC).

This makes the Total Sobol Index a suitable criterion for variable selection, as also supported in the literature (see Bénard, Da Veiga and Scornet (2022)), who stated that it is the best quantity for finding the minimal subset of predictive variables.

Leveraging the relationship with CFI, it is possible to directly inherit all the **statistical guarantees** established in the previous section, such as the asymptotic Type-I error control based on Markov's inequality (Reyero Lobo, Neuvial and Thirion (2025)) and the finite-sample guarantees from the HRT (Tansey et al. (2022)). Furthermore, when analyzing the LOCO version, one can employ the data-splitting variants proposed by Williamson et al. (2023) to achieve asymptotic Type-I error control or, similarly to the CFI, rely on the Markov-based guarantees (Verdinelli and Wasserman (2024a)). Finally, similarly to the perturbation indices, nonparametric tests such as the sign test or the Wilcoxon test can be directly applied by treating the losses using the j-th coordinate as one population and those excluding this information as another. In this way, we obtain finite-sample type-I error guarantees (Lei et al., 2018; Watson and Wright, 2021).

5.4 Total Sobol Index Modifications

Verdinelli and Wasserman (2024a) criticized the TSI due to its tendency to underestimate the importance of features that are correlated with others. This phenomenon is known as *correlation distortion* (see Verdinelli and Wasserman (2024b)). To mitigate this issue, Verdinelli and Wasserman (2024a) proposed a normalized *decorrelated* version of the TSI, defined as

$$\psi_{\mathrm{dTSI}}(j,P) := \frac{\mathbb{E}\left[(m(X) - m_{-j}(X^{-j}))^2\right]}{\mathbb{E}\left[(X^j - \nu_{-j}(X^{-j}))^2\right]},$$

where $\nu_{-j}(X^{-j}) := \mathbb{E}[X^j \mid X^{-j}]$ denotes the conditional mean of X^j given all other covariates.

The motivation behind this normalization comes from the behavior of TSI under a linear model (2). In that case, the TSI can be expressed as

$$\psi_{\mathrm{TSI}}(j,P) = \beta_j^2\, \mathbb{E}\left[(X^j - \nu_{-j}(X^{-j}))^2\right],$$

which shows that the index is scaled by the conditional variance of X^j given X^{-j}. Therefore, when this variance is small due to high correlation, the importance decreases. The proposed normalization corrects this by dividing out the conditional variance, thereby restoring the importance of such variables and better aligning with the axioms discussed in section 2.2.

Verdinelli and Wasserman (2024a) also introduced other decorrelation-based strategies to address correlation distortion. However, they also found those alternatives to suffer from second-order bias and low-density inference issues. In any case, they satisfy the minimal axiom due to its alignment with the functional dependence axiom (see Axiom 5 and Proposition 3.1).

Finally, Du, Roeder and Wasserman (2025) proposed Disentangled Feature Importance (DFI), which is mainly based on the total Sobol index computed in a latent space where input coordinates are rendered independent via a transport map. However, DFI only satisfies the minimal axiom in this transformed space—not in the original input space. As a result, it can inflate the importance of irrelevant coordinates and lead to a high rate of false discoveries, as seen in their experiments.

5.5 Shapley Additive Global importancE (SAGE)

Shapley values are widely used for local variable importance (Lundberg and Lee, 2017; Spadaccini, Fokkema and van de Wiel, 2025), and were adapted to global variable importance by Covert, Lundberg and Lee (2020), who explicitly defined variable importance in terms of the predictive power a feature provides to the model. However, in this section, we argue that this performance criterion is in fact not fulfilled by standard Shapley values, as they do not satisfy the *minimal* axiom. We begin by explicitly introducing SAGE:

DEFINITION 5.7 (ψ_{SAGE}). Shapley Additive Global importancE (SAGE) is given by

$$\psi_{\text{SAGE}}(j,P) := \sum_{S\subset -\{j\}} w_S(v(S\cup\{j\}) - v(S)), \tag{8}$$

where $v(S) := \mathbb{E}\left[\ell(y,\mathbb{E}\left[y\right])\right] - \mathbb{E}\left[\ell(y,m_S(X^S))\right]$, with $m_S(X^S) := \mathbb{E}\left[y \mid X^S\right]$ and $w_S := \frac{1}{p}\binom{p-1}{|S|}^{-1}$.

Each $v(S)$ denotes the SAGE value function associated with the subset $S\subset[p]$, i.e., the change in performance relative to the average prediction when incorporating the information from X^S. In particular, we denote by ψ_{SAGEvf} the value corresponding to $S=\{j\}$.

DEFINITION 5.8 (ψ_{SAGEvf}). SAGE value function associated with the j-th feature is given by

$$\psi_{\text{SAGEvf}} := v(\{j\}) = \mathbb{E}\left[\ell(y,\mathbb{E}\left[y\right])\right] - \mathbb{E}\left[\ell(y,m_j(X^j))\right].$$

Notably, it coincides with the index estimated by the Leave One Covariate In (LOCI) method. However, these two approaches were not previously connected—similarly to the case of LOCO and Sobol-CPI—because SAGE is based on marginalization, whereas LOCI relies on refitting. Despite this inference difference, both aim to estimate the same underlying quantity. As a result, they will often lead to the same covariate rankings and selections.

Nonetheless, we emphasize that neither the SAGE value function (and thus LOCI; see section C.6) nor SAGE itself satisfies the minimal axiom.

PROPOSITION 5.9. *Neither ψ_{SAGE} nor ψ_{SAGEvf} satisfies the minimal axiom.*

In particular, Shapley values may assign importance to features that are not actually used by the model, simply because they are correlated with truly important features (see Figure 1). Consequently, the removal of such features does not degrade model performance, since they do not directly influence the predictions of the model. Moreover, the explainability of the method is hindered by the fact most features are assigned non-null importance: the explanations lack parsimony, and the set of features called important becomes unmanageable.

5.5.1 Surplus SAGE: Ewald et al. (2024) also proposed to study another quantity, namely the *surplus*, defined as the difference between the SAGE value function of all coordinates except the j-th and that of all coordinates.

DEFINITION 5.10 (ψ_{scSAGEj}). Surplus SAGE value function associated with the j-th feature is given by

$$\psi_{scSAGEj} := v(-j\cup j) - v(-j) = v([p]) - v(-j).$$

We note that this VIM does not introduce a new index, it exactly coincides with the Total Sobol Index (expanding the definitions, we recover (4)). Thus, we do not restate the principled approach, as it has already been applied in Section 5.3. In particular, this VIM satisfies the minimal axiom.

PROPOSITION 5.11. *Under additive noise assumption (1a) or classification assumption (1b), a strictly convex loss ℓ, ψ_{scSAGEj} satisfies the minimal axiom.*

5.5.2 Marginal extensions: in any case, the inference of ψ_{SAGE} is complex due to two main challenges: 1) the number of subsets to consider grows exponentially with the number of features, and 2) for each subset $S\subset[p]$, it is necessary to estimate the restricted model m_S.

To address the first issue, various approaches have been proposed, ranging from Monte Carlo Covert, Lundberg and Lee (2020) to more sophisticated methods such as importance sampling guided by prior information, for example from the structure of random forests Bénard et al. (2022).

Regarding the second issue, Covert, Lundberg and Lee (2020) proposed using marginalization instead of refitting. That is, rather than training a new model $\widehat{m}_S$, they approximate m_S by the conditional expectation $\mathbb{E}[m(X) \mid X^S]$. For the original SAGE estimator, this requires conditional sampling, which is computationally expensive. To mitigate this, they further proposed using marginal sampling as a proxy for conditional sampling. Accordingly, they distinguish between two variants: the original *conditional SAGE*, and the more computationally efficient *marginal SAGE* (with similar distinctions for their value function counterparts).

However, these variants target different theoretical quantities and thus have different interpretations. In particular, while conditional SAGE and conditional SAGEvf do not satisfy the minimal axiom (as illustrated by Figure 1), marginal SAGE and marginal SAGEvf do.

PROPOSITION 5.12. *Let ℓ be a loss with a unique minimizer which is a function of the conditional distribution of y given X. Under additive noise assumption (1a) or classification assumption (1b) ψ_{mSAGE} and ψ_{mSAGEvf} satisfy the minimal axiom.*

Similarly to the marginal interpretation of PFI discussed in Section 5.2, Ewald et al. (2024) related the marginal SAGE counterpart to marginal independence. However, Proposition 5.12 shows that this interpretation is not accurate. In fact, similarly to the perturbation indices, the global model acts as a first filter for conditional independence due to its functional dependence on the input features (see Proposition 3.1).

Intuitively, since marginal SAGE involves comparing subsets by altering only coordinate j, if feature j is not important, it will not influence the model's output. As a result, all comparisons involving j yield zero importance.

This confusion between marginal SAGE and marginal independence partly arises from the terminology "marginal" and "conditional" SAGE. Conversely, the conditional SAGE value function, which coincides with LOCI, can only be employed for marginal testing. This highlights that the distinction is not merely theoretical but also reflects conceptual differences, as the two approaches assess feature importance either conditionally or marginally, even though both involve marginalization. Consequently, directly comparing mSAGE and cSAGE is not meaningful.

Finally, note that, as shown in the second part of Proposition 5.4, under the asymptotic relevance assumption (3), perturbation indices assign vanishing importance to the null coordinates. The same result can therefore be obtained under this assumption for both $\widehat{\psi}_{\text{mSAGE}}$ and $\widehat{\psi}_{\text{mSAGEvf}}$.

6. SUMMARY OF THEORETICAL INDICES AND VIM METHODS

In this section, we present a summary overview of the definitions of the main theoretical indices, whether they satisfy the minimal axiom, and the main Variable Importance Measures (VIMs), detailing what they estimate and how they are estimated.

We denote the conditional and marginal Shapley value functions for $S \subset [p]$ as:

$$v(S) := \mathbb{E}\left[\ell(y, \mathbb{E}[y])\right] - \mathbb{E}\left[\ell\left(y, \mathbb{E}[m(X) \mid X^S]\right)\right],$$
$$v^m(S) := \mathbb{E}\left[\ell(y, \mathbb{E}[y])\right] - \mathbb{E}\left[\ell\left(y, \mathbb{E}[m(X^{(-S)}) \mid X^S]\right)\right].$$

The former ($v(S)$) reflects the intrinsic conditional dependence, while the latter ($v^m(S)$) involves marginal dependence. Therefore, $\mathbb{E}[m(X^{(-S)}) \mid X^S]$ means that the $-S$ coordinates are *averaged* without taking into account the relationship with X^S, which are fixed. We also recall the Shapley weighting:

$$w_S := \frac{1}{p}\binom{p-1}{|S|}^{-1}.$$

The main theoretical indices, along with their definitions and whether or not they satisfy the minimal axiom (MA, Axiom 8), are summarized in table 1.

Next, we summarize the main methods and how they are estimated. We use the abbreviations: **P** for perturbation, **M** for marginalization, and **R** for refitting. To maintain readability in the table, we report the theoretical quantity rather than the explicit estimation formulas. In practice, these correspond to plug-in estimators: expectations are approximated using empirical means over a test set, and the function m is replaced by a trained machine learning model.

For the construction of $X^{(j)}$, the j-th column is permuted across the dataset. The conditional version, $\widetilde{X}^{(j)}$, is obtained via a conditional sampler. In CPI-based methods, this conditional permutation is used, although any conditional sampler (as in CFI) is theoretically valid.

To distinguish between marginalization and refitting when estimating the restricted model m_S, we denote marginalization-based approaches as $\mathbb{E}[m(X) \mid X^S]$ (conditional sampling, then we use v), and refitting-based approaches simply as m_S. For marginal sampling, we denote $\mathbb{E}[m(X^{(-S)}) \mid X^S]$ (then, we use v^m).

The difference between LOCO (Lei et al. (2018)) and LOCO-W (Williamson et al. (2023)) lies in the use of a extra data splitting in the latter. This data splitting

Index	Definition	MA
ψ_{TSI}	$\mathbb{E}\left[\ell(m_{-j}(X^{-j}), y)\right] - \mathbb{E}\left[\ell(m(X), y)\right]$	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}	$\mathbb{E}\left[\ell(y, \mathbb{E}\left[y\right])\right] - \mathbb{E}\left[\ell(y, m_j(X^j))\right]$	No
ψ_{mSAGEvf}	$\mathbb{E}\left[\ell(y, \mathbb{E}\left[y\right])\right] - \mathbb{E}\left[\ell(y, \mathbb{E}\left[m(X^{(-j)})\right])\right]$	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}	$\mathbb{E}\left[\ell(m(X^{(j)}), y)\right] - \mathbb{E}\left[\ell(m(X), y)\right]$	Yes

TABLE 1

Summary of the theoretical indices: *Theoretical indices, their target quantities, and whether they satisfy the minimal axiom.*

Method	Theoretical quantity	Index	Estimation
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)	$\frac{n_{\mathrm{cal}}}{n_{\mathrm{cal}}+1}\left(\mathbb{E}\left[\ell(\frac{1}{n_{\mathrm{cal}}}\sum_{k=1}^{n_{\mathrm{cal}}} m(\widetilde{X}_k^{(j)}), y)\right] - \mathbb{E}\left[\ell(m(X), y)\right]\right)$	ψ_{TSI}	P/M

TABLE 2

Summary of the VIMs: *The theoretical quantity targeted by the variable importance measures, the corresponding index, and the estimation procedure considered (M for marginalization, R for refitting, and P for perturbation).*

avoids using the same training set for both model refittings, which ensures that the variance does not vanish and enables asymptotic normality, thereby allowing for valid type-I error control. However, Reyero Lobo, Neuvial and Thirion (2025) discuss the inference consequences of such data splitting. Principally, this causes the variance to explode, which leads to far fewer discoveries. This is also seen in the experiments (Section 7).

Sobol-CPI(n-cal) can be interpreted as a perturbation approach. Indeed, for n-cal $= 1$, it reduces to a normalized CPI/CFI. However, for larger values of n-cal, then marginalizing, it accounts for the estimation bias introduced by not having access to the true conditional expectation, which is not addressed by other marginalization-based methods.

In general, refitting-based approaches incur higher computational costs, as the models used to represent the relationship with the output tend to be more complex than those used for conditional sampling. Additionally, Shapley-based approaches are often slower, as they involve an exponential number of feature combinations. These factors should be taken into account when choosing an estimation procedure.

7. EXPERIMENTS

In this section, we compare all the presented global variable importance measures on both simulated and real datasets to illustrate two main points: (1) whether the theoretical index satisfies the minimal axiom—and therefore whether the estimate tends to zero when a covariate has no predictive power—and (2) the necessity of the general pipeline. Indeed, the estimated importance (and in particular the resulting ranking, which reflects the predictive power of each feature) remains consistent across estimation approaches, whether based on refitting, perturbation, or marginalization.

7.1 Methods

We compare `Sobol-CPI(1)` (a normalization of the `CPI` from Chamma, Engemann and Thirion (2023)) and `Sobol-CPI(100)` from Reyero Lobo, Neuvial and Thirion (2025), `PFI` from Breiman (2001), `CFI` from Strobl et al. (2008), marginal / conditional SAGE (`m/c SAGE`) and their respective value function variants (`m/c SAGEvf`) from Covert, Lundberg and Lee (2020), `scSAGEj` from Ewald et al. (2024), `LOCO` from Lei et al. (2018), `LOCO-W` from Williamson et al. (2023), and `LOCI` (Leave One Covariate In). To enable comparison across all methods, the importance scores have been normalized. Most methods were computed using the Python package **fippy** (https://github.com/gcskoenig/fippy). Code to reproduce these experiments is available at https://github.com/AngelReyero/Principled-VIM-comparison.

The model used is a Gradient Boosting, but in section D, we present the same experiments using a Random Forest to show that, with a less accurate model, the results are similar—albeit with less precise inference. All experiments were repeated at least 50 times using different random seeds.

7.2 Simulated data

For our simulated data, we did not replicate the exact setup from Ewald et al. (2024), in which the unimportant coordinates X_1 and X_2 are nearly identical. Indeed, this induces collinearity, which prevents accurate estimation of the linear model and results in a model that includes the term $0.36X_1 - 0.36X_2$.

As a consequence, Ewald et al. (2024) argue that PFI incorrectly assigns importance to these uninformative covariates. However, this is actually an estimation issue, since the linear model is not consistent in cases of perfect correlation; in contrast, the theoretical PFI does not attribute any importance to such features. Furthermore, this setup violates the identifiability assumption, which requires that no coordinate be a deterministic function of the others.

We generate the input X as a Gaussian vector with zero mean and a Toeplitz covariance matrix defined by $\Sigma_{i,j} = 0.6^{|i-j|}$, with dimension $p = 10$. The output is defined as $y = X_0 + 2X_1 - X_4^2 + X_7X_8$. We use $n = 5000$ samples for the main experiments.

In Figure 2, we plot the estimated importances across all methods. For clarity, we present only one important and one unimportant feature, while the full set of estimated importances—from which similar conclusions can be drawn—is provided in Appendix D.1.

On the left of Figure 2, for an important covariate, we observe that even though the quantities have been estimated using different inference methods, those corresponding to the same theoretical index yield similar importance values. On the right, there is the importance estimated to a not important coordinate. We observe that all the VIMs that satisfy the minimal axiom, represented by plain boxplots, provide a null importance. For instance, this illustrates that PFI is suited for conditional testing, contrary to what was said before.

In Section D.1.1, we examine how the estimation varies as n ranges from 100 to 5000. The conclusions remain similar: numerical comparisons should be made between estimates of the same index to assess different inference properties, while conceptual comparisons should rely on their theoretical counterparts.

7.3 Real data

Following Ewald et al. (2024), we use the bike dataset from Fanaee-T and Gama (2014). In Figure 3, we plot the estimated importance of the variable *year* across all the discussed VIMs. We argue that this covariate is not important, i.e., it does not contribute predictive power to the model. Nevertheless, we observe that VIMs that do not satisfy the minimal axiom assign importance to this irrelevant covariate. Specifically, ψ_{SAGE} (estimated by cSAGE) and ψ_{LOCI} (estimated by cSAGEvf via marginalization and LOCI via refitting) both assign a nonzero importance.

The claim of Ewald et al. (2024) that PFI can assign high FI values to features even if they are not associated with the target but with other features that are associated with the target does not appear to be supported by the results of this experiment. In fact, we observe in Figure 3 that PFI, which satisfies the minimal axiom, assigns zero importance to the variable *year*. This indicactes that the model acts as a filter for conditional independence. As a result, PFI assigns different importance scores (and therefore different rankings) to the relevant covariates compared to methods like LOCO; however, both approaches identify the same set of important covariates.

In Figure 4, we present boxplots of all features using only the VIMs that aim to estimate the total Sobol index. We omit `LOCO-W` due to its high variability, which

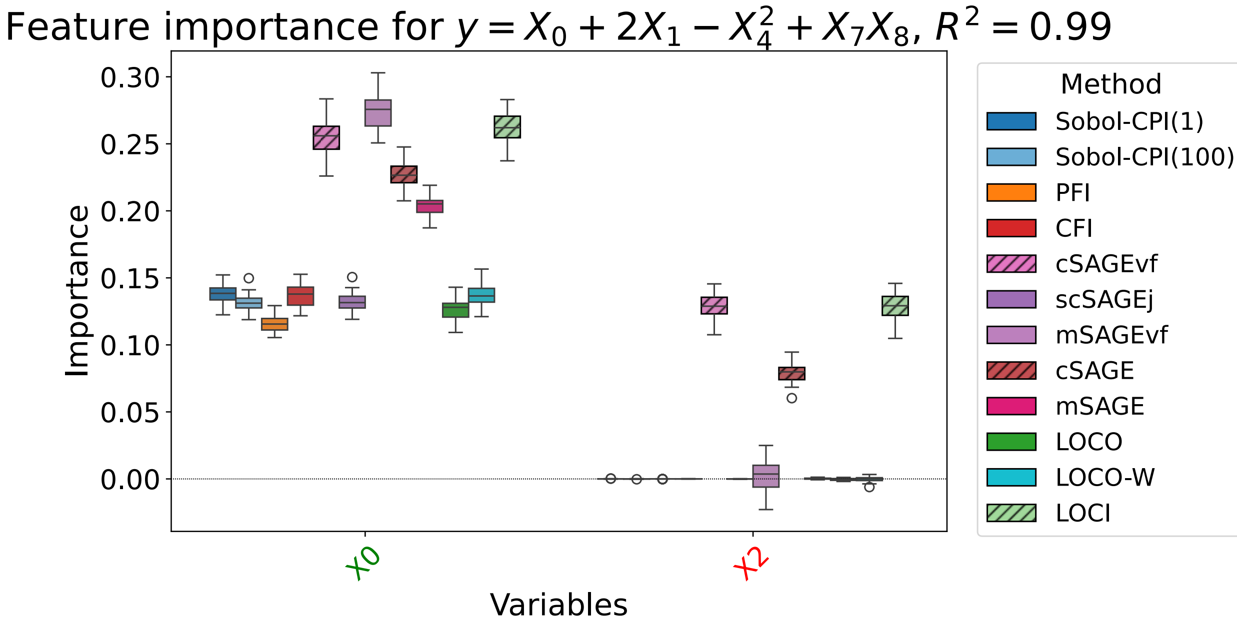

FIG 2. ***Boxplot of estimated VIMs for an important and an unimportant variable:*** *Methods not satisfying theoretically the minimal axiom have boxes filled with diagonal hatch lines. The left panel shows an important covariate, while the right panel shows an unimportant one. Only conditional SAGE (vf) and LOCI fail to satisfy the minimal axiom, assigning non-zero importance to the unimportant variable.* `Sobol-CPI(1)`, `Sobol-CPI(100)`, `CFI`, `scSAGEj`, `LOCO`, *and* `LOCO-W` *aim to estimate* ψ_{TSI}.

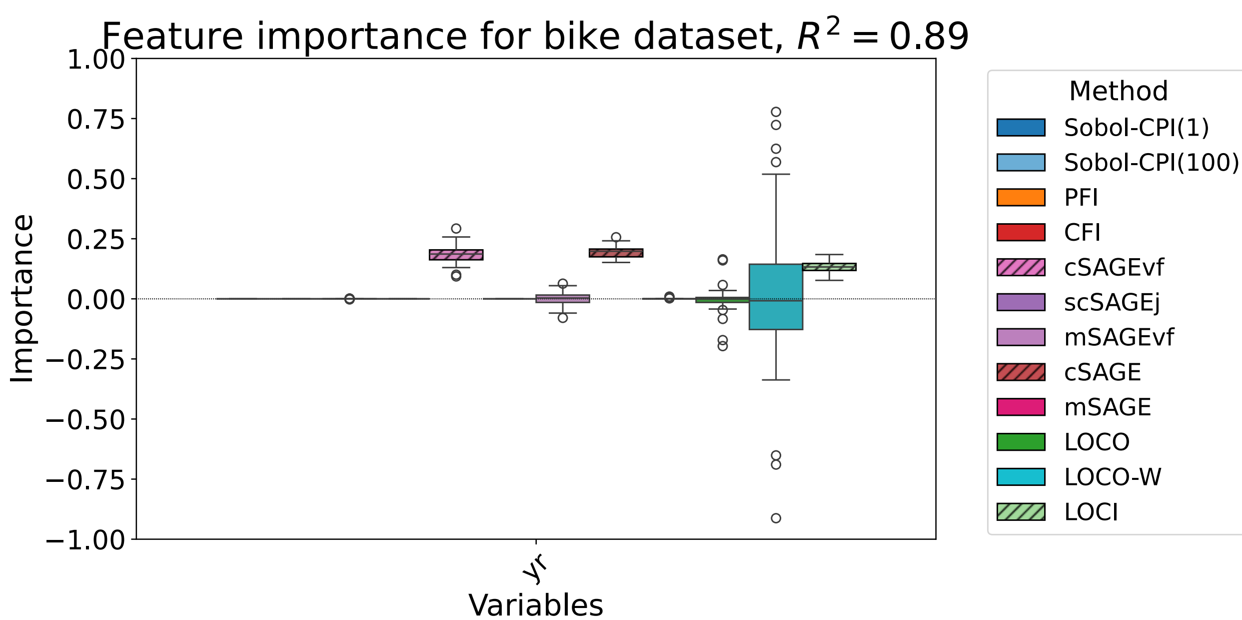

FIG 3. ***Boxplots of the VIMs for the year feature:*** *Methods not satisfying theoretically the minimal axiom have boxes filled with diagonal hatch lines. All methods that estimate an importance index satisfying the minimal axiom assign no importance to this variable.*

compromises readability. The complete figure—including additional methods and results based on a Random Forest—is provided in Appendix D.2.

We observe that all displayed methods produce similar importance values and rankings, even on real data, regardless of whether they rely on marginalization, refitting, or perturbation. This similarity arises because they target the same underlying quantity. However, their estimation properties differ. For instance, refitting-based methods exhibit greater variability (Paillard et al. (2025)), which can lead to inaccurate variable importance estimates that are highly sensitive to the data and the optimization procedure. Consequently, the resulting rankings tend to be less reliable in practice compared to those obtained via marginalization or perturbation. Furthermore, perturbation methods possess a double robustness property (Reyero Lobo, Neuvial and Thirion (2025)), which in practice translates into greater robustness to model misspecification under the null hypothesis, leading to faster convergence rates and reduced variance, as observed for variables such as *month* and *holiday*.

8. DISCUSSION

In this article, we have proposed a new categorization of variable importance measures (VIMs). In doing so, we aim to provide practitioners with more insight into how to proceed when working with real datasets. We argue that comparing VIMs only based on their inference procedures it is not meaningful, and can be misleading. Instead, it is crucial to clearly define the theoretical index corresponding to the intended goal, and then select an appropriate estimation method. In particular, we emphasize that comparing methods aimed at estimating different indices is not meaningful, as each method yields a different ranking aligned with its own objectives.

This principled framework acknowledges that global variable importance aims to capture information about the true underlying distribution, with the model acting as a surrogate that simplifies this information. Furthermore, establishing a connection between variable importance and variable selection is essential for making insightful discoveries with statistical guarantees. Since the index is a theoretical quantity that must be estimated from data, it is crucial to provide statistical guarantees for the selected covariates. This connection is formalized through the minimal axiom, highlighting that both fields pursue the same objective.

The proposed axiom not only formalizes the notion of importance as predictive power but also generalizes it by relating it to the conditional independence testing framework under the established assumptions. We note that some notions of importance do not fall within this definition; for example, a feature may influence the output only through its variance without providing any predictive signal. However, we argue that when performing scientific inference using an ML model to explain the X–y relationship, no stronger notion can be achieved, since such alternative forms of importance are not captured by the predictive model.

While the presented work offers a rigorous framework for the development of VIMs, there remains a need for further research on this subject. On the one hand, new theoretical indices should be developed to capture alternative notions of importance, with explicit articulation of the axioms they are designed to satisfy. On the other hand, more work is needed on the estimation side, particularly for challenging indices such as the decorrelated LOCO, whose inference bias complicates accurate estimation.

APPENDIX A: NOTATION GLOSSARY

The notation used in this paper is gathered in Table 3. Some notation can be combined; for instance,

	Description
$X \in \mathbb{R}^p$	Input
$j \in [p]$	Feature of interest
$X^j \in \mathbb{R}$	j-th input covariate
$X^{-j} \in \mathbb{R}^{p-1}$	X with the j-th covariate excluded
$X^S \in \mathbb{R}^{p-\lvert S\rvert}$	S-th input covariates for $S \subset [p]$
$y \in \mathbb{R}$	Output
$P \in \mathcal{M}$	Distribution of (X, y)
$X^{(j)}$	Marginal perturbation
$\widetilde{X}^{(j)}$	Conditional perturbation
$m(X)$ (resp. $m_{-j}(X^{-j})$)	$\mathbb{E}\left[y \mid X\right]$ (resp. $\mathbb{E}\left[y \mid X^{-j}\right]$)
$m_S(X)$	$\mathbb{E}\left[y \mid X^S\right] = \mathbb{E}\left[m(X) \mid X^S\right]$
$\widehat{m}$ (resp. $\widehat{m}_{-j}$)	Estimation of m (resp. of m_{-j})
ℓ	Loss function
$\mathcal{F}$	Generic space of functions
$\psi(j, P)$	Importance index of j under P
n_{cal}	Size of calibration set

TABLE 3
Notation used in the paper.

$\widetilde{X}^{(j)l}$ denotes the l-th coordinate of $\widetilde{X}^{(j)}$. Also, we recall that marginal perturbation means $X^{(j)-j} = X^{-j}$, $\mathcal{L}(X^{(j)j}) = \mathcal{L}(X^j)$ and $X^{(j)j} \perp\!\!\!\perp X^{-j}, Y$, and conditional is $\widetilde{X}^{(j)-j} = X^{-j}$, $\widetilde{X}^{(j)j} \perp\!\!\!\perp y \mid X^{-j}$ and $\mathcal{L}(\widetilde{X}) = \mathcal{L}(X)$.

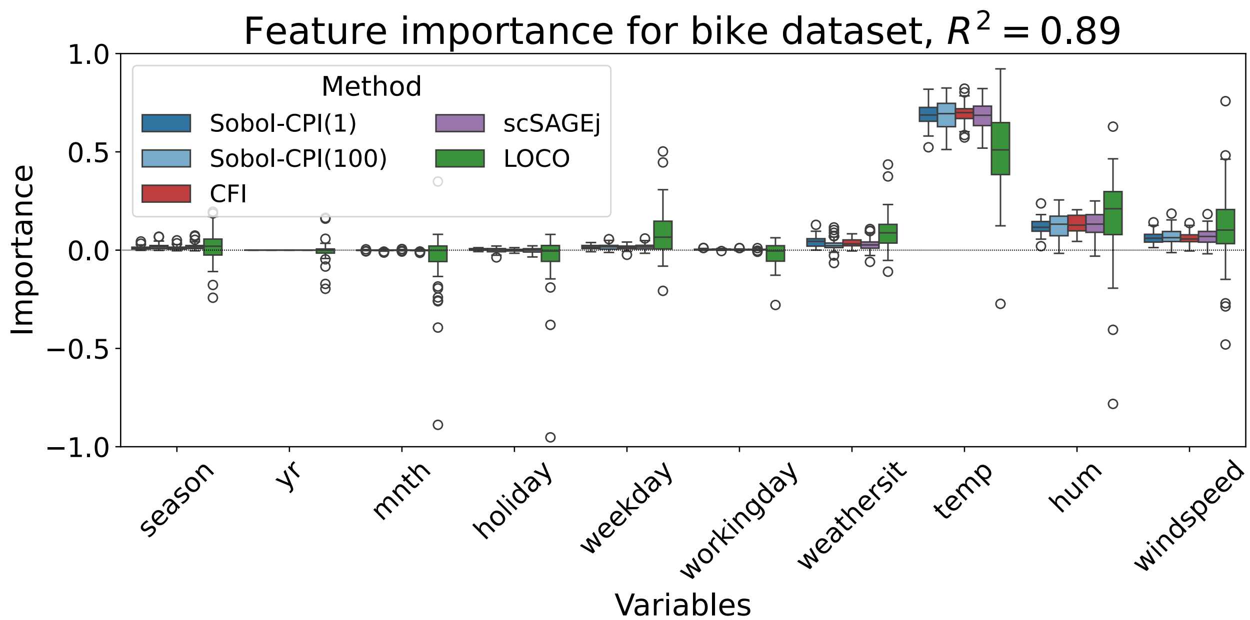

FIG 4. ***Boxplots of the VIMs estimating ψ_{TSI} for all features:*** *All methods aim to estimate the same theoretical quantity. While their estimates are generally close, the refitting-based approaches exhibit poorer inference properties.*

APPENDIX B: CONDITIONAL AND FUNCTIONAL INDEPENDENCE

In this section, we prove that, under certain standard assumptions, it is possible to establish the equivalence between conditional independence and functional independence, i.e., between $y \perp\!\!\!\perp X^j \mid X^{-j}$ and $m(X) \in \mathcal{F}_{-j}$.

First, under the standard additive noise assumption (1a), Reyero Lobo, Neuvial and Thirion (2025) proved this equivalence in Lemma I.1 of the appendix, which is based on proving that $m(X)$ is $\sigma(X^{-j})$-measurable.

For classification, we will establish the result under the general classification assumption (1b).

First, if $y \perp\!\!\!\perp X^j \mid X^{-j}$, we have that

$$
\begin{aligned}
&\text{(using Assumption 1b)} && \sigma(m(X)) = \mathbb{P}(y=1 \mid X) \\
&\text{(using that } y \perp\!\!\!\perp X^j \mid X^{-j}) && = \mathbb{P}(y=1 \mid X^{-j}) \\
& && = f_{-j}(X^{-j}),
\end{aligned}
$$

for some $f_{-j} \in \mathcal{F}_{-j}$, as $\mathbb{P}(y=1 \mid X^{-j})$ belongs to the σ-algebra generated by X^{-j}.

Then, if $m \in \mathcal{F}_{-j}$, then we have that

$$
\begin{aligned}
&\text{(using the tower rule)} \\
&\qquad \mathbb{P}(y=1 \mid X^{-j}) = \mathbb{E}\left[\mathbb{P}(y=1 \mid X) \mid X^{-j}\right] \\
&\text{(using Assumption 1b)} \qquad = \mathbb{E}\left[\sigma(m(X)) \mid X^{-j}\right] \\
&\text{(using that } m \in \mathcal{F}_{-j}) \qquad = \mathbb{E}\left[\sigma(m_{-j}(X^{-j})) \mid X^{-j}\right] \\
&(\sigma(m_{-j}(X^{-j})) \text{ is } \sigma(X^{-j})\text{-measurable}) \\
&\qquad = \sigma(m_{-j}(X^{-j})) \\
&\text{(using definition of } m_{-j}) \qquad = \sigma(m(X)) \\
&\qquad = \mathbb{P}(y=1 \mid X).
\end{aligned}
$$

Therefore, as y binary and $\mathbb{P}(y=1 \mid X^{-j}) = \mathbb{P}(y=1 \mid X)$, then $y \perp\!\!\!\perp X^{-j} \mid X^{-j}$.

APPENDIX C: MINIMAL AXIOM PROOFS

In this section, we study whether the theoretical indices presented in the main text satisfy the minimal axiom.

C.1 Generalized Linear Models $\psi_{\mathbf{GLM}}$

PROOF OF PROPOSITION 5.2. On the one hand, assume that $X^j \perp\!\!\!\perp y \mid X^{-j}$. Then, we have that $g(X^1\beta_1 + \ldots + X^p\beta_p) =: \mathbb{E}[y \mid X] = \mathbb{E}\left[y \mid X^{-j}\right]$, then it is $\sigma(X^{-j})$-measurable. However, as X^j is not $\sigma(X^{-j})$-measurable by identifiability assumption, then $\beta_j = 0$.

On the other hand, assume that $\beta_j = 0$. Then, in particular using the GLM likelihood assumption (McCullagh (1989)):

$$p_y(y; \theta, \phi) = \exp\left(\frac{y\theta - \mathrm{b}(\theta)}{\mathrm{a}(\phi)} + \mathrm{c}(y;\phi)\right),$$

we have that the canonical parameter does not depend on X^j because $\beta_j = 0$, then $p_y(y \mid X) = p_y(y \mid X^{-j})$. $\square$

C.2 Perturbation $\psi_{\mathbf{perturb}}$

PROOF OF PROPOSITION 5.4. First, we prove it in particular for the quadratic loss. For any perturbation $\widetilde{X}^j$, using Assumption 1a we have that

$$\psi_{\mathrm{perturb}}(j, P) = \mathbb{E}\left[(y - m(\widetilde{X}^j))^2\right] - \mathbb{E}\left[(y - m(X))^2\right]$$

$$
\begin{aligned}
&= \mathbb{E}\left[(y - m(X^{-j}, \widetilde{X}^{jj}))^2\right] - \mathbb{E}\left[(y - m(X))^2\right] \\
&= \mathbb{E}\left[(m(X) + \epsilon - m(X^{-j}, \widetilde{X}^{jj}))^2\right] \\
&\quad - \mathbb{E}\left[(m(X) + \epsilon - m(X))^2\right] \\
&= \mathbb{E}\left[(m(X) - m(X^{-j}, \widetilde{X}^{jj}))^2\right].
\end{aligned}
$$

Thus, using the strict convexity of the quadratic loss, we have that $m(X) - m(X^{-j}, \widetilde{X}^{jj}) = 0$ almost surely if and only if $m \in \mathcal{F}_{-j}$. Therefore, we can conclude by applying Proposition 3.1.

More generally, for any strictly convex loss ℓ, we have that $\mathbb{E}\left[\ell(y, m(\widetilde{X}^j))\right] - \mathbb{E}\left[\ell(y, m(X))\right] = 0$ if and only if $m(X) - m(\widetilde{X}^j) = 0$ almost surely. Then, we conclude using Proposition 3.1.

For the second part of the proposition, it suffices to apply Theorem 3.4 from Reyero Lobo, Neuvial and Thirion (2025). Indeed, this result follows from one of the double robustness properties of CFI: for any perturbation, using the asymptotic relevance of $\widehat{m}$, the importance vanishes.

□

C.2.1 PFI is not suitable for marginal testing. In the literature, there is a common misconception that PFI is suitable for marginal testing (see Ewald et al. (2024)), due to the fact that the permutation is performed marginally. However, in this section, we show that this is not the case. First, using Proposition 5.4, we show that if PFI is nonzero, then X^j is dependent on the output given the rest of the input. Then, we demonstrate that the assumptions made in Ewald et al. (2024) to draw conclusions about marginal independence implies conditional independence.

C.2.1.1 PFI different than 0: In Ewald et al. (2024), they claim that if $\psi_{\mathrm{PFI}}(j, P) \neq 0$ and $X^j \perp\!\!\!\perp X^{-j} \mid y$, then $X^j \perp\!\!\!\perp y$. In this section we argue that indeed, if $\psi_{\mathrm{PFI}}(j, P) \neq 0$ then we have that $X^j \perp\!\!\!\perp y \mid X^{-j}$. More generally, this applies to any perturbation index as seen in Proposition 5.4. The problem arises from the fact that the model used in the index is not an arbitrary model, but rather an estimate of the data-generating process. Under standard assumptions, this model acts as a filter for conditional independence. Consequently, if the difference in loss is nonzero, it implies that the model relied on the covariate, and therefore, the covariate is important—i.e., it is conditionally dependent.

C.2.1.2 PFI equal to 0: In Ewald et al. (2024) they claimed that if $\psi_{\mathrm{PFI}}(j, P) = 0, X^j \perp\!\!\!\perp X^{-j}$ and $X^j \perp\!\!\!\perp X^{-j} \mid y$, then $X_j \perp\!\!\!\perp y$. However, this also implies that $X^j \perp\!\!\!\perp y \mid X^{-j}$. Indeed, we have that

$$
\begin{aligned}
p(X^j, y \mid X^{-j}) &= \frac{p(X^j, y, X^{-j})}{p(X^{-j})} \\
&= \frac{p(y)p(X^j \mid y)p(X^{-j} \mid y)}{p(X^{-j})} \\
&= \frac{p(y)p(X^j)p(X^{-j} \mid y)}{p(X^{-j})} \\
&= \frac{p(X^j)p(X^{-j}, y)}{p(X^{-j})} \\
&= \frac{p(X^j)p(y \mid X^{-j})p(X^{-j})}{p(X^{-j})} \\
&= p(X^j)p(y \mid X^{-j}) \\
&= p(X^j \mid X^{-j})p(y \mid X^{-j}).
\end{aligned}
$$

Therefore, we conclude that $X^j \perp\!\!\!\perp y \mid X^{-j}$.

In general, as shown in Proposition 5.4, there is no need to impose such restrictive assumptions to draw conclusions about conditional independence. We present this development because Ewald et al. (2024) first showed that if a method is suited for detecting marginal independence, then it is not suited for detecting conditional independence. Second, they showed that under this restrictive setting, PFI is appropriate for marginal independence. We therefore demonstrate that, in this particular case, it also implies conditional independence, thus aligning with our more general result.

C.3 General Total Sobol Index ψ_{TSI}

PROOF OF PROPOSITION 5.6. Covert, Lundberg and Lee (2020) established this result in the context of the cross-entropy loss. Specifically, they showed that

$$
\mathrm{I}(y; X^j \mid X^{-j}) = 0 \quad \Longleftrightarrow \quad X^j \perp\!\!\!\perp y \mid X^{-j},
$$

which corresponds precisely to the minimal axiom using (7).

For the quadratic loss, we can directly use the CFI formulation of the TSI and apply Proposition 5.4.

Under the assumptions of the proposition, this equivalence follows directly. Since the loss ℓ is strictly convex, it admits a unique minimizer, which by definition is the Bayes predictor, a function of $m \in \mathcal{F}$. Hence, we have that

$$
\begin{aligned}
&\mathbb{E}[\ell(y, m(X))] = \mathbb{E}[\ell(y, m_{-j}(X^{-j}))] \\
&\Longleftrightarrow \quad m(X) = m_{-j}(X^{-j}) \text{ a.s.}
\end{aligned}
$$

This condition implies that X^j provides no additional predictive power beyond X^{-j}, i.e., m is functionally independent of X^j given X^{-j}. By the equivalence between functional and conditional independence under these assumptions, we conclude that $\psi_{\mathrm{TSI}}(j, P)$ satisfies the minimal axiom.

□

C.4 Shapley Additive Global importancE (SAGE, $\psi_{\mathbf{SAGE}}$)

PROOF OF PROPOSITION 5.9. We begin by showing that ψ_{SAGEvf} does not satisfy the minimal axiom, as there exists a spurious correlation between the j-th coordinate and an important coordinate. Similarly, since ψ_{SAGE} combines multiple terms—including those reflecting this spurious importance—it also fails to satisfy the minimal axiom. Finally, we prove that ψ_{scSAGEj} coincides with ψ_{TSI} and, in particular, as established in the previous section, satisfies the minimal axiom.

We recall the notation $m_S(X^S) = \mathbb{E}[y \mid X^S] = \mathbb{E}[m(X) \mid X^S]$ for $S \subseteq [p]$. In particular, for $S = \{j\}$, we denote it simply as m_j.

We start with the SAGE value function. Indeed, it is given by

$$\psi_{\text{SAGEvf}}(j, P) := \mathbb{E}\left[\ell(\mathbb{E}\left[y\right], y)\right] - \mathbb{E}\left[\ell(m_j(x^j), y)\right],$$

with $m_j(x^j) := \mathbb{E}\left[m(X^j, X^{-j}) \mid X^j = x^j\right] = \mathbb{E}[y \mid X^j = x^j]$. Therefore, it is easy to construct the following general counterexample: for $y = m_{-j}(X^{-j}) + \epsilon$, with $\epsilon \perp\!\!\!\perp X$, we observe that $y \perp\!\!\!\perp X^j \mid X^{-j}$, but if $X^j \not\perp\!\!\!\perp X^{-j}$, it will not be 0 as $\mathbb{E}\left[y \mid X^j\right] \neq \mathbb{E}\left[y\right]$.

For the SAGE, it is given by

$$\psi_{\text{SAGE}}(j, P) = \frac{1}{d} \sum_{S \subset -\{j\}} \binom{d-1}{|S|}^{-1} \left(v(S \cup \{j\}) - v(S)\right),$$

with $v(S) = \mathbb{E}\left[\ell(y, \mathbb{E}\left[y\right])\right] - \mathbb{E}\left[\ell(y, m_S(x^S)\right]$. In particular, using the same counterexample as for the SAGE value function, we observe that for $S = \emptyset$ the difference is strictly positive as $v(\emptyset \cup \{j\}) - v(\emptyset) = v(\{j\}) > 0$. Therefore, $\psi_{\text{cSAGE}}(j, P) > 0$, since it is a combination of differences that are either positive or zero, with at least one strictly positive. However, since $X^j \perp\!\!\!\perp y \mid X^{-j}$, the minimal axiom is not satisfied.

For the ψ_{scSAGEj}, we note that it is exactly the total Sobol Index:

$$\begin{aligned}
\psi_{\text{scSAGEj}} &:= v(-\{j\} \cup \{j\}) - v(-\{j\}) \\
&= \mathbb{E}\left[\ell(y, \mathbb{E}\left[y\right])\right] - \mathbb{E}\left[\ell(y, m_{-\{j\} \cup \{j\}}(x^{-\{j\} \cup \{j\}})\right] \\
&\quad - \mathbb{E}\left[\ell(y, \mathbb{E}\left[y\right])\right] + \mathbb{E}\left[\ell(y, m_{-\{j\}}(x^{-\{j\}})\right] \\
&= \mathbb{E}\left[\ell(y, m_{-\{j\}}(x^{-\{j\}})\right] - \mathbb{E}\left[\ell(y, m(x)\right] \\
&= \psi_{\text{TSI}}(j, P).
\end{aligned}$$

In particular, it satisfies the minimal axiom (see Proposition 5.6). ☐

C.5 Marginal SAGE $\psi_{\mathbf{mSAGE}}$

PROOF OF PROPOSITION 5.12. For the notation, we denote

$$m_S^m(X^S) := \mathbb{E}[m(X^{(-S)}) \mid X^S].$$

This means, in particular, that the coordinates in $X^{(-S)S} := X^S$ are fixed, and the expectation is taken over the remaining coordinates, which are independent from X^S due to marginal permutation.

We observe that, for this proof, there is no need to preserve any specific distribution over the variables on which the expectation is taken (i.e., $X^{(-S)-S}$); it suffices to have independence. In this way, the coordinates over which we take the expectation are not influenced by the others, and thus the conditional filtering of the trained function remains valid—unlike in the original SAGE.

To make this point more explicit, note that by independence,

$$\begin{aligned}
m_S^m(x^S) &:= \mathbb{E}[m(X^{(-S)}) \mid X^S = x^S] \\
&= \mathbb{E}[m(X^{(-S)-S}, x^S) \mid X^S = x^S] \\
&= \mathbb{E}[m(X^{(-S)-S}, x^S)].
\end{aligned}$$

Therefore, the information provided by the coordinates in S does not influence the distribution of the remaining coordinates.

If the sampling is done marginally, we have in particular that $m_S^m(x^S) = \mathbb{E}[m(X^{-S}, x^S)]$.

We start with the marginal SAGE value function. It is defined as

$$\psi_{\text{mSAGEvf}}(j, P) := \mathbb{E}\left[\ell(\mathbb{E}\left[y\right], y)\right] - \mathbb{E}\left[\ell(m_j^m(x^j), y)\right].$$

Note that $m_j^m(x^j) := \mathbb{E}\left[m(x^j, X^{-j})\right]$. Under the null hypothesis, $m \in \mathcal{F}_{-j}$, and therefore

$$m_j^m(x^j) = \mathbb{E}[m(x^j, X^{-j})] = \mathbb{E}\left[m_{-j}(X^{-j})\right] = \mathbb{E}\left[y\right].$$

Thus, its value is 0 if and only if x^j is conditional independent. Similarly, for the marginal SAGE, it can be written as

$$\psi_{\text{mSAGE}} = \frac{1}{d} \sum_{S \subset -\{j\}} \binom{d-1}{|S|}^{-1} \left(v^m(S \cup \{j\}) - v^m(S)\right),$$

with $v^m(S) = \mathbb{E}\left[\ell(y, \mathbb{E}\left[y\right])\right] - \mathbb{E}\left[\ell(y, m_S^m(x^S)\right]$. Thus, we just need to note that $v^m(S \cup \{j\}) = v^m(S)$ for all $S \subset -\{j\}$. To see this, similarly as before, using the equivalence between conditional dependence and functional dependence, we observe that

$$\begin{aligned}
m_{S \cup \{j\}}^m(x_{S \cup \{j\}}) &= \mathbb{E}\left[m(x^{S \cup \{j\}}, X^{-(S \cup \{j\})})\right] \\
&= \mathbb{E}\left[m_{-j}(x^S, X^{-(S \cup \{j\})})\right] \\
&= \mathbb{E}\left[m(x^S, X^{-S \cup \{j\}})\right] \\
&= m_S^m(x_S).
\end{aligned}$$

Then, we have that

$$\begin{aligned} v^m(S \cup \{j\}) &= \mathbb{E}\left[\ell(y, \mathbb{E}\left[y\right])\right] - \mathbb{E}\left[\ell(y, m^m_{S\cup\{j\}}(x^{S\cup\{j\}})\right] \\ &= \mathbb{E}\left[\ell(y, \mathbb{E}\left[y\right])\right] - \mathbb{E}\left[\ell(y, m^m_S(x^S)\right] \\ &= v^m(S). \end{aligned}$$

□

C.6 Leave One Covariate In (LOCI, ψ_{LOCI})

PROPOSITION C.1. *ψ_{LOCI} does not satisfy the minimal axiom.*

PROOF OF PROPOSITION C.1. Similarly to the (conditional) SAGE value function for the subset $\{j\}$, we have that it is given by

$$\mathbb{E}[\ell(y, \mathbb{E}[y])] - \mathbb{E}[\ell(y, m_j(X^j))],$$

except that the index j appears as a refitting index for the estimation rather than a marginalization index. Nevertheless, the index remains the same, and the same counterexample applies to show that this difference is non-zero, even under conditional independence.

□

APPENDIX D: ADDITIONAL EXPERIMENTS

In this section, we present the remaining results on variable importance and additional convergence behaviors. This provides a more complete view of the experiments.

D.1 Simulated data

From Figure 5, where we used a Random Forest, we observe the same overall pattern as in Figure 2, where the main model is a Gradient Boosting machine. We also note that the Gradient Boosting model is more accurate (achieving a better R^2), but its estimated importance values exhibit higher variability. Therefore, even though both models are expected to converge to the same quantity, some inference methods may be more desirable than others.

For instance, Sobol-CPI (and similarly, CFI) benefits from a double robustness property (see Reyero Lobo, Neuvial and Thirion (2025)), which enhances its ability to detect null covariates and leads to better performance on X^2 compared to PFI. Additionally, we observe that LOCI not only fails to satisfy the minimal axiom—by assigning importance to uninformative covariates—but also assigns negative importance to important covariates, exhibiting behavior that is clearly undesirable. This could be sign of a strong model overfitting.

For completeness, in Figures 6 and 7 we present the estimated importance for all the features, using a Gradient Boosting and a Random Forest respectively. From these figures, we observe that the VIMs satisfying the minimal axiom do not assign any importance to the red (uninformative) features. We also note the equivalence between `Sobol-CPI(1)`, `Sobol-CPI(100)`, `CFI`, `scSAGEj`, `LOCO`, and `LOCO-W`, even though they rely on different estimation approaches and models.

From Figure 7, we first observe that the estimation quality is slightly lower due to the reduced performance of the model. Notably, LOCI exhibits highly undesirable behavior: it not only assigns importance to uninformative covariates, violating the minimal axiom, but also assigns negative importance to important covariates. This is likely a consequence of the variability in the optimization process. Overall, Sobol-CPI demonstrates more reliable inference results (see Paillard et al. (2025)).

D.1.1 Convergence In Figure 8, we illustrate that the behavior observed in the boxplots from Figure 2 is not specific to a fixed sample size n, but rather remains stable across different values of n. On the left panel, which corresponds to an important covariate, we observe that several distinct trends emerge early in the sampling process and remain consistently separated as n increases. On the right panel, we observe that both SAGE and LOCI assign importance to features that are, in fact, not important. This suggests that these methods may fail to satisfy the minimal axiom, as they can attribute relevance to covariates that have no true impact on the outcome.

In Figure 9, we focus exclusively on VIMs that target the total Sobol index. As such, this figure does not aim to compare different theoretical importance indices, but rather to assess the inference properties of estimators targeting the same quantity—making it a purely methodological comparison. On the left, we observe that for the important covariate, the CPI-based methods exhibit lower variance compared to refitting-based approaches (`LOCO(-W)`), consistent with findings from Paillard et al. (2025). On the right, we report the mean bias across null covariates. Again, CPI-based methods show more favorable behavior due to their double robustness property, as discussed in Reyero Lobo, Neuvial and Thirion (2025).

D.2 Real data

In Figure 10, we reproduce the same figure as in Figure 4, but using a Random Forest instead of Gradient Boosting. We observe that the results and rankings remain essentially unchanged, assigning the highest importance to temperature and humidity, and consistently identifying the same covariates as unimportant.

Finally, for completeness, we present all the VIMs across all methods for both Gradient Boosting and Random Forest in Figure 11 and Figure 12, respectively.

We first observe that although `LOCO-W` targets the same quantity as the other TSI methods, it exhibits substantial variability due to the data-splitting and refitting steps required to ensure valid type-I error control. While

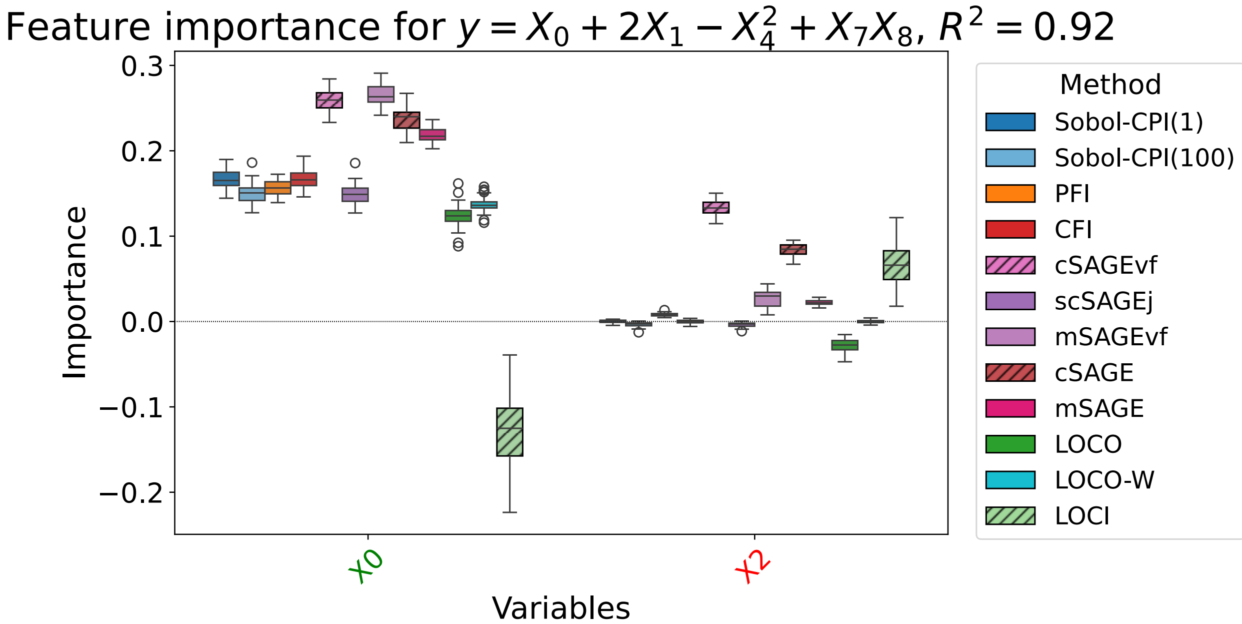

FIG 5. ***Boxplot of estimated VIMs for an important and an unimportant variable using a Random Forest:*** *Methods not satisfying theoretically the minimal axiom have boxes filled with diagonal hatch lines. The left panel shows an important covariate, while the right panel shows an unimportant one. Only conditional SAGE (vf) and LOCI fail to satisfy the minimal axiom, assigning non-zero importance to the unimportant variable.* `Sobol-CPI(1)`, `Sobol-CPI(100)`, `CFI`, `scSAGEj`, `LOCO`, *and* `LOCO-W` *aim to estimate* ψ_{TSI}.

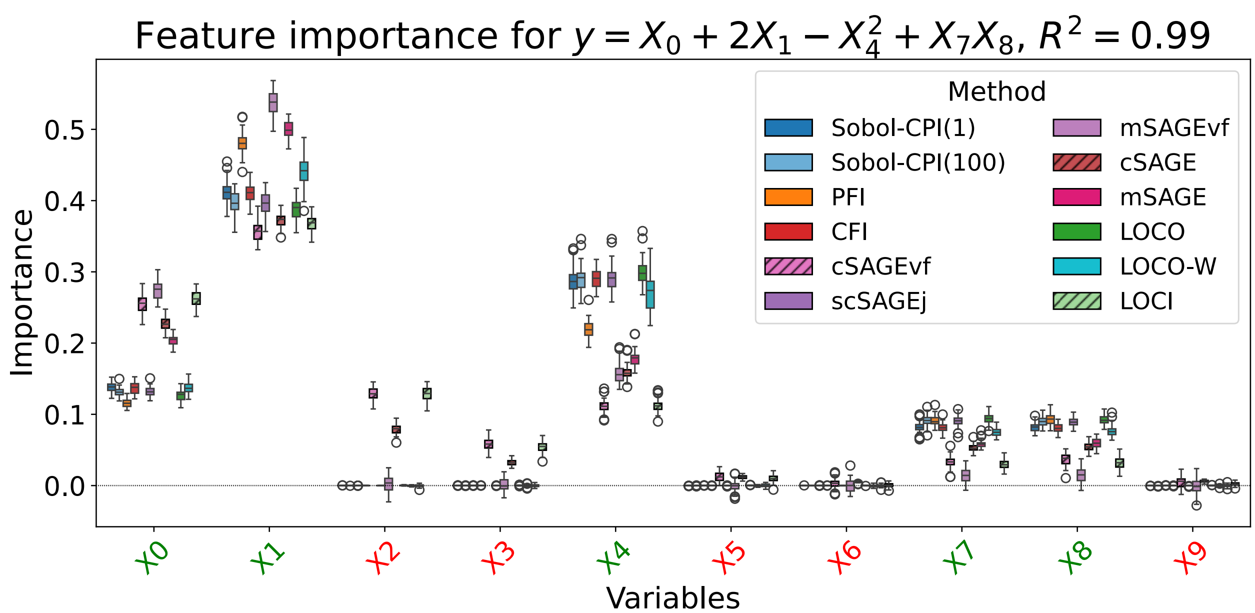

FIG 6. ***Boxplot of estimated VIMs for all features using Gradient Boosting:*** *Methods not satisfying theoretically the minimal axiom have boxes filled with diagonal hatch lines. Important covariates are shown in green, and unimportant ones in red. Only conditional SAGE (vf) and LOCI fail to satisfy the minimal axiom by assigning non-zero importance to uninformative variables.* `Sobol-CPI(1)`, `Sobol-CPI(100)`, `CFI`, `scSAGEj`, `LOCO`, *and* `LOCO-W` *aim to estimate* ψ_{TSI}.

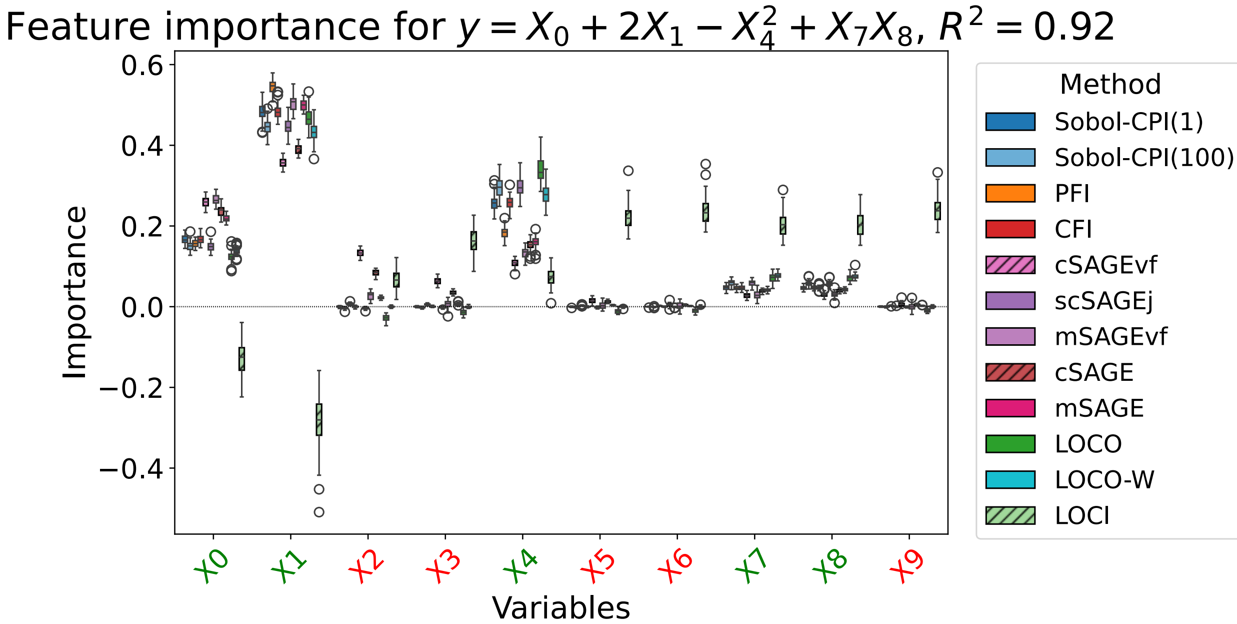

FIG 7. ***Boxplot of estimated VIMs for all features using Random Forest:*** *Methods not satisfying theoretically the minimal axiom have boxes filled with diagonal hatch lines. Important covariates are shown in green, and unimportant ones in red. Only conditional SAGE (vf) and LOCI fail to satisfy the minimal axiom by assigning non-zero importance to uninformative variables.* `Sobol-CPI(1)`, `Sobol-CPI(100)`, `CFI`, `scSAGEj`, `LOCO`, *and* `LOCO-W` *aim to estimate* ψ_{TSI}.

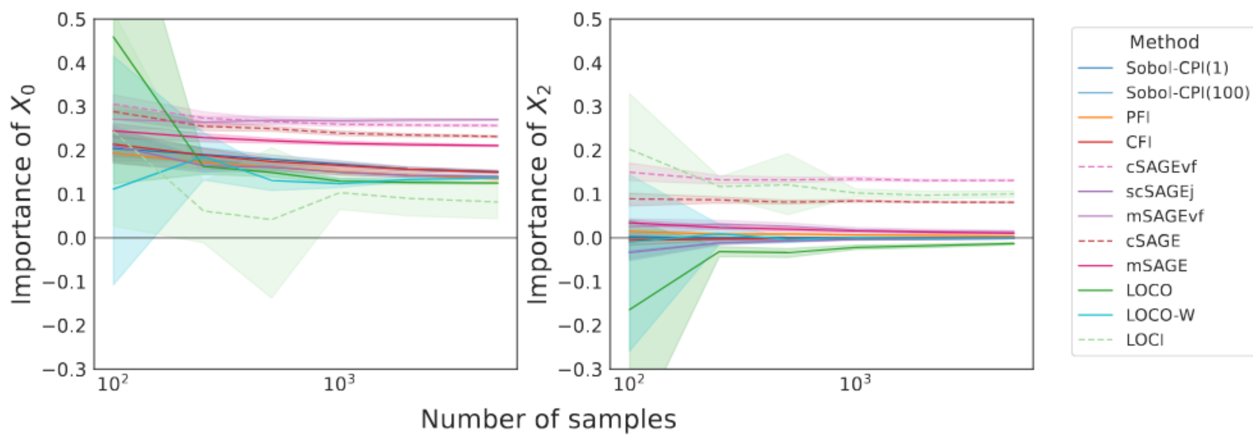

FIG 8. ***Convergence of VIMs for an important and an unimportant covariate:*** *Methods not satisfying theoretically the minimal axiom have discontinuous lines. The left panel shows the convergence behavior of various VIMs for an important covariate, highlighting different trends as each method targets distinct indices with different objectives. The right panel illustrates the behavior for an unimportant covariate, where we can observe which VIMs satisfy the minimal axiom by assigning negligible or zero importance.*

this approach provides formal statistical guarantees, we argue that its high variability limits its practical usefulness.

Additionally, we note that VIM methods satisfying the minimal axiom do not consistently assign importance to the same set of variables.

REFERENCES

BÉNARD, C., BIAU, G., DA VEIGA, S. and SCORNET, E. (2022). SHAFF: Fast and consistent SHApley eFfect estimates via random Forests. In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics* (G. CAMPS-VALLS, F. J. R. RUIZ and I. VALERA, eds.). *Proceedings of Machine Learning Research* **151** 5563–5582. PMLR.

BOZDOGAN, H. (1987). Model selection and Akaike's Information Criterion (AIC): The general theory and its analytical extensions. *Psychometrika* **52** 345–370. https://doi.org/10.1007/BF02294361

BREIMAN, L. (2001). Random Forests. *Machine Learning* **45** 5–32. https://doi.org/10.1023/A:1010933404324

BÜHLMANN, P. (2013). Statistical significance in high-dimensional linear models. *Bernoulli* **19** 1212 – 1242. https://doi.org/10.3150/12-BEJSP11

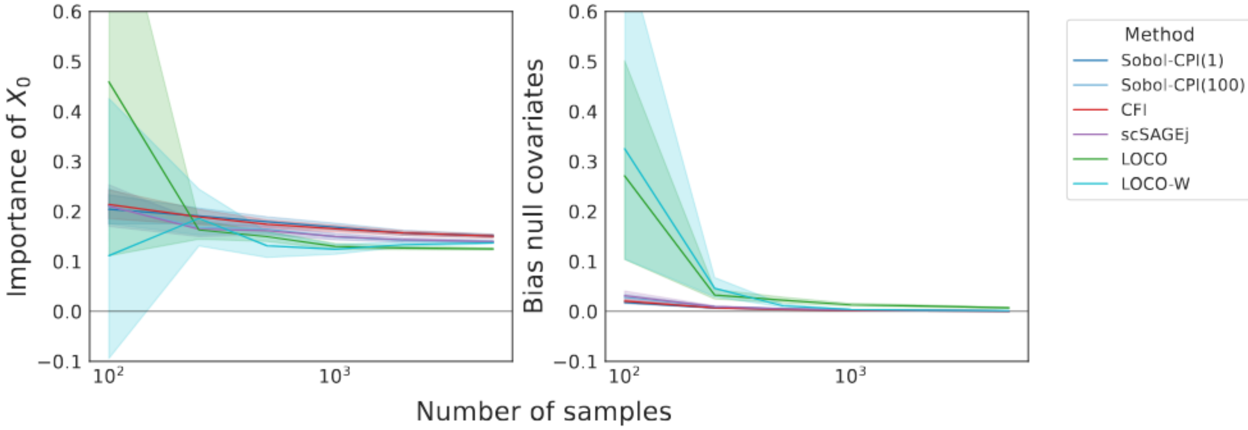

FIG 9. ***Convergence of ψ_{TSI} estimates:*** *The left panel shows the convergence behavior of the ψ_{TSI} estimates for an important covariate. The right panel displays the mean bias across unimportant covariates.*

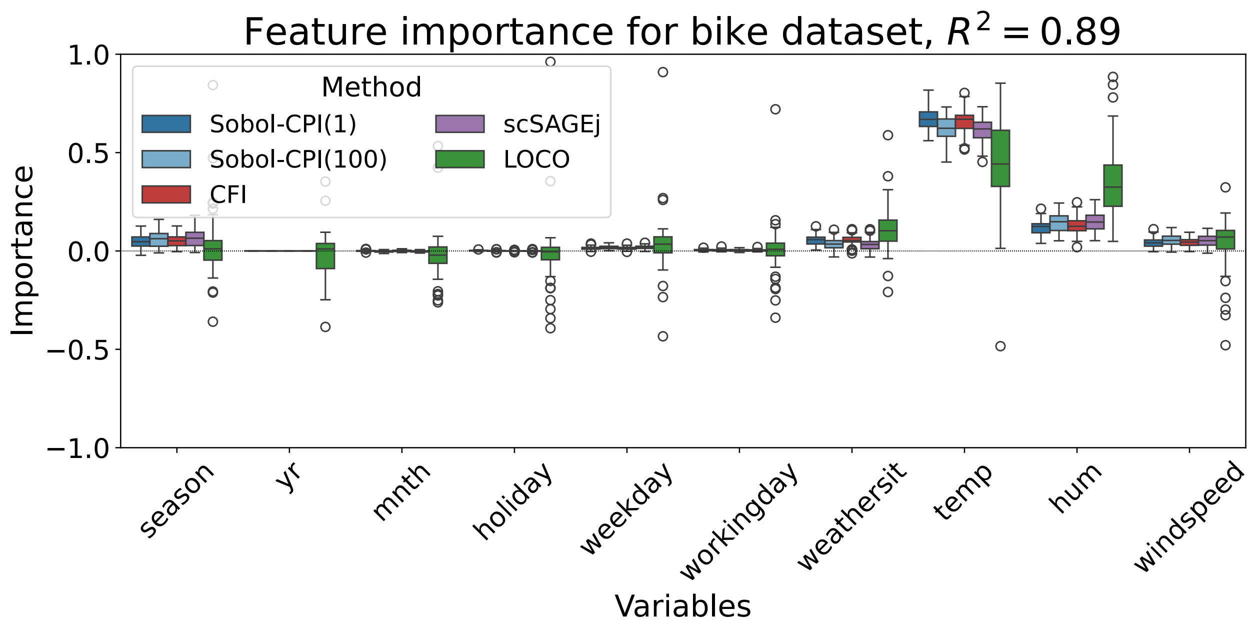

FIG 10. ***Boxplots of the VIMs estimating ψ_{TSI} for all features using a Random Forest:*** *All methods aim to estimate the same theoretical quantity. While their estimates are generally close, the refitting-based approaches (*`LOCO`*) exhibit poorer inference properties.*

BÉNARD, C., DA VEIGA, S. and SCORNET, E. (2022). Mean decrease accuracy for random forests: inconsistency, and a practical solution via the Sobol-MDA. *Biometrika* **109** 881-900.

CANDÈS, E., FAN, Y., JANSON, L. and LV, J. (2018). Panning for Gold: 'Model-X' Knockoffs for High Dimensional Controlled Variable Selection. *Journal of the Royal Statistical Society Series B: Statistical Methodology* **80** 551-577. https://doi.org/10.1111/rssb.12265

CHAMMA, A., ENGEMANN, D. A. and THIRION, B. (2023). Statistically Valid Variable Importance Assessment through Conditional Permutations.

CHAMMA, A., ENGEMANN, D. and THIRION, B. (2024). Variable Importance in High-dimensional Settings Requires Grouping. In *Proceedings of the Thirty-Eighth Conference on Artificial Intelligence (AAAI 2024)*. Underline Science Inc. https://doi.org/10.48448/pj0s-yg09

COVERT, I., LUNDBERG, S. M. and LEE, S.-I. (2020). Understanding global feature contributions with additive importance measures. *Advances in Neural Information Processing Systems* **33** 17212–17223.

DEZEURE, R., BÜHLMANN, P., MEIER, L. and MEINSHAUSEN, N. (2015). High-Dimensional Inference: Confidence Intervals, p-Values and R-Software hdi. *Statistical Science* **30** 533–558. https://doi.org/10.1214/15-STS527

DONNELLY, J., KATTA, S., RUDIN, C. and BROWNE, E. P. (2023). The rashomon importance distribution: getting RID of unstable, single model-based variable importance. In *Proceedings of the 37th International Conference on Neural Information Processing Systems. NIPS '23*. Curran Associates Inc., Red Hook, NY, USA.

DU, J.-H., ROEDER, K. and WASSERMAN, L. (2025). Disentangled Feature Importance. Arxiv preprint https://arxiv.org/abs/2507.00260.

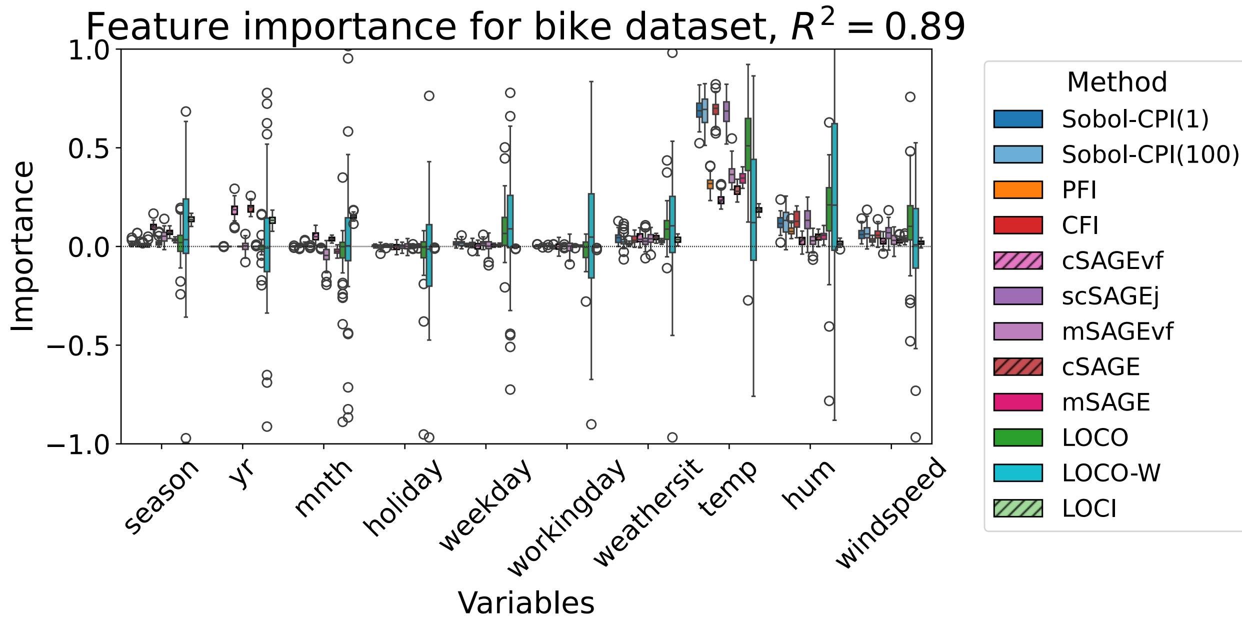

FIG 11. ***Boxplots of all VIMs for all features using Gradient Boosting:*** *Methods not satisfying theoretically the minimal axiom have boxes filled with diagonal hatch lines. Refitting-based methods (`LOCO(-W)`) exhibit high variability, which limits their applicability in real-world scenarios. Methods estimating the total Sobol index produce similar results, and VIMs that satisfy the minimal axiom lead to consistent feature selection.*

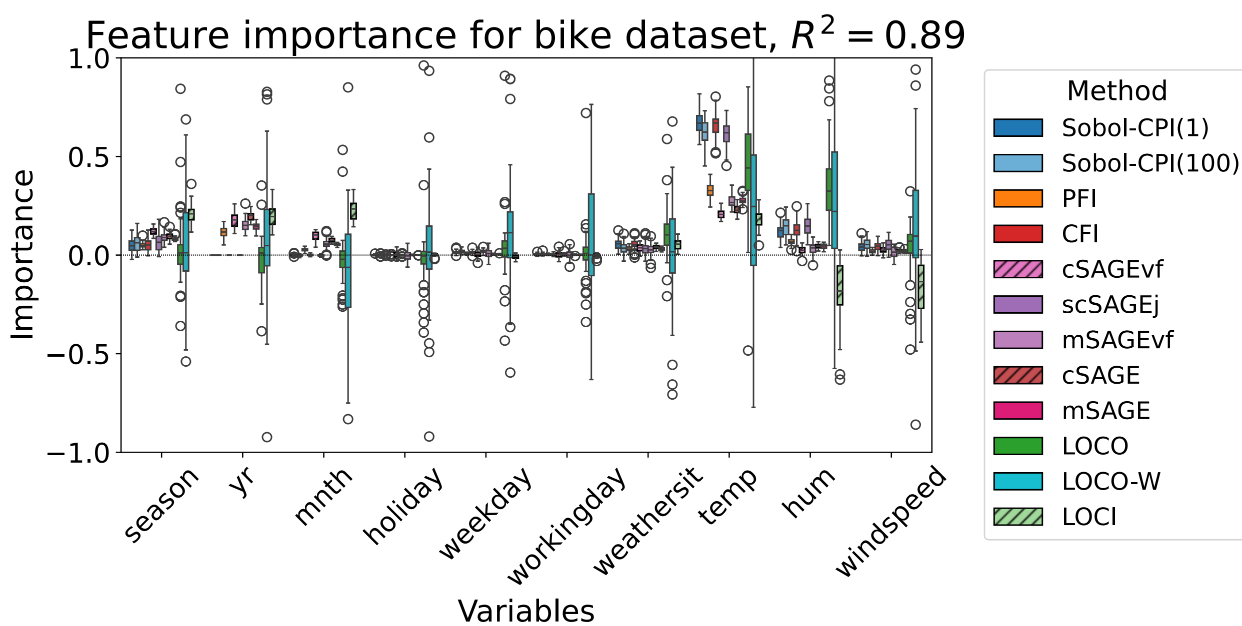

FIG 12. ***Boxplots of all VIMs for all features using Random Forest:*** *Methods not satisfying theoretically the minimal axiom have boxes filled with diagonal hatch lines. Refitting-based methods (`LOCO(-W)`) exhibit high variability, which limits their applicability in real-world scenarios. Methods estimating the total Sobol index produce similar results, and VIMs that satisfy the minimal axiom lead to consistent feature selection.*

EWALD, F. K., BOTHMANN, L., WRIGHT, M. N., BISCHL, B., CASALICCHIO, G. and KÖNIG, G. (2024). *A Guide to Feature Importance Methods for Scientific Inference* In *Explainable Artificial Intelligence* 440–464. Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-63797-1_22

FANAEE-T, H. and GAMA, J. (2014). Event labeling combining ensemble detectors and background knowledge. *Progress in Artificial Intelligence* **2** 113–127. https://doi.org/10.1007/s13748-013-0040-3

FUMAGALLI, F., MUSCHALIK, M., HÜLLERMEIER, E., HAMMER, B. and HERBINGER, J. (2025). Unifying Feature-Based Explanations with Functional ANOVA and Cooperative Game Theory.

FÖGE, N., SCHMID, L., DITZHAUS, M. and PAULY, M. (2024). A Central Limit Theorem for the permutation importance measure.

GIRAUD, C. (2021). *Introduction to High-Dimensional Statistics*, 2nd ed. Chapman and Hall/CRC. https://doi.org/10.1201/9781003158745

HOMMA, T. and SALTELLI, A. (1996). Importance measures in global sensitivity analysis of nonlinear models. *Reliability Engineering & System Safety* **52** 1-17.

HOOKER, G., MENTCH, L. and ZHOU, S. (2021). Unrestricted permutation forces extrapolation: variable importance requires at least one more model, or there is no free variable importance. *Statistics and Computing* **31** 1–16. https://doi.org/10.1007/s11222-021-10057-z

KÖNIG, G., MOLNAR, C., BISCHL, B. and GROSSE-WENTRUP, M. (2021). Relative Feature Importance. In *2020 25th International Conference on Pattern Recognition (ICPR)* 9318–9325. IEEE. https://doi.org/10.1109/ICPR48806.2021.9412143

KUMAR, I. E., VENKATASUBRAMANIAN, S., SCHEIDEGGER, C. and FRIEDLER, S. A. (2020). Problems with Shapley-value-based explanations as feature importance measures. In *Proceedings of the 37th International Conference on Machine Learning*. *ICML'20*. JMLR.org.

LEI, J., G'SELL, M., RINALDO, A., TIBSHIRANI, R. J. and WASSERMAN, L. (2018). Distribution-Free Predictive Inference for Regression. *Journal of the American Statistical Association* **113** 1094–1111. https://doi.org/10.1080/01621459.2017.1307116

LIU, M., KATSEVICH, E., JANSON, L. and RAMDAS, A. (2021). Fast and Powerful Conditional Randomization Testing via Distillation.

LUNDBERG, S. M. and LEE, S.-I. (2017). A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*. *NIPS'17* 4768–4777. Curran Associates Inc., Red Hook, NY, USA.

MCCULLAGH, P. (1989). *Generalized Linear Models*, 2nd ed. Routledge, New York. eBook published 22 January 2019. Subjects: Mathematics & Statistics. https://doi.org/10.1201/9780203753736

MEINSHAUSEN, N. (2013). Discussion of Grouping Strategies and Thresholding for High Dimension Linear Models. *Journal of Statistical Planning and Inference* **143** 1439-1440. https://doi.org/10.1016/j.jspi.2013.03.005

MI, X., ZOU, B., ZOU, F. and HU, J. (2021). Permutation-based identification of important biomarkers for complex diseases via machine learning models. *Nature Communications* **12** 3008. Published: 21 May 2021. https://doi.org/10.1038/s41467-021-22756-2

MOLNAR, C. (2025). *Interpretable Machine Learning*, 3 ed.

MOLNAR, C., KÖNIG, G., HERBINGER, J., FREIESLEBEN, T., DANDL, S., SCHOLBECK, C. A., CASALICCHIO, G., GROSSE-WENTRUP, M. and BISCHL, B. (2022). *General Pitfalls of Model-Agnostic Interpretation Methods for Machine Learning Models* In *xxAI - Beyond Explainable AI: International Workshop, Held in Conjunction with ICML 2020, July 18, 2020, Vienna, Austria, Revised and Extended Papers* 39–68. Springer International Publishing, Cham. https://doi.org/10.1007/978-3-031-04083-2_4

PAILLARD, J., REYERO LOBO, A., KOLODYAZHNIY, V., THIRION, B. and ENGEMANN, D. A. (2025). Measuring Variable Importance in Heterogeneous Treatment Effects with Confidence. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*. PMLR arXiv:2408.13002.

REYERO LOBO, A., NEUVIAL, P. and THIRION, B. (2025). On the double robustness of Conditional Feature Importance. Arxiv preprint https://arxiv.org/abs/2501.17520.

REYERO LOBO, P., DAGA, E., ALANI, H. and FERNANDEZ, M. (2023). Knowledge-Grounded Target Group Language Recognition in Hate Speech. *Studies on the Semantic Web* **56** 1–18. https://doi.org/10.3233/SSW230002

RINALDO, A., WASSERMAN, L. and G'SELL, M. (2019). Bootstrapping and Sample Splitting for High-Dimensional, Assumption-Lean Inference. *The Annals of Statistics* **47** 3438–3469. https://doi.org/10.1214/18-AOS1765

RUDIN, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence* **1** 206–215. https://doi.org/10.1038/s42256-019-0048-x

SHAPLEY, L. S. (1953). A Value for n-Person Games. In *Contributions to the Theory of Games II* (H. W. Kuhn and A. W. Tucker, eds.) 307–317. Princeton University Press, Princeton. https://doi.org/10.1515/9781400881970-018

SPADACCINI, G., FOKKEMA, M. and VAN DE WIEL, M. A. (2025). Hypothesis-free discovery from epidemiological data by automatic detection and local inference for tree-based nonlinearities and interactions.

STROBL, C., BOULESTEIX, A.-L., KNEIB, T., AUGUSTIN, T. and ZEILEIS, A. (2008). Conditional variable importance for random forests. *BMC Bioinformatics* **9** 307. https://doi.org/10.1186/1471-2105-9-307

TANSEY, W., VEITCH, V., ZHANG, H., RABADAN, R. and AND, D. M. B. (2022). The Holdout Randomization Test for Feature Selection in Black Box Models. *Journal of Computational and Graphical Statistics* **31** 151–162. https://doi.org/10.1080/10618600.2021.1923520

TIBSHIRANI, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)* **58** 267–288.

VERDINELLI, S. and WASSERMAN, L. (2024a). Feature Importance: A Closer Look at Shapley Values and LOCO. *Statistical Science* **39** 623–636. https://doi.org/10.1214/24-STS911

VERDINELLI, I. and WASSERMAN, L. (2024b). Decorrelated Variable Importance. *Journal of Machine Learning Research* **25** 1–27.

WATSON, D. S. and WRIGHT, M. N. (2021). Testing conditional independence in supervised learning algorithms. *Machine Learning* **110** 2107–2129. https://doi.org/10.1007/s10994-021-06030-6

WILLIAMSON, B. D., GILBERT, P. B., CARONE, M. and SIMON, N. (2021). Nonparametric variable importance assessment using machine learning techniques. *Biometrics* **77** 9-22.

WILLIAMSON, B. D., GILBERT, P. B., SIMON, N. R. and CARONE, M. (2023). A general framework for inference on algorithm-agnostic variable importance. *Journal of the American Statistical Association* **118** 1645–1658.